Universitá degli Studi di Salerno

PhD thesis in
COMPUTER SCIENCE

Automatic Discovery of Drug Mode of Action and Drug Repositioning from Gene Expression Data

Candidate: Francesco Iorio

Supervisors:
Prof. Giancarlo Raiconi, Dr. Diego di Bernardo
Coordinator:
Prof. Margherita Napoli

IX Cycle - 2007/2010

To Giorgia:
the greatest result
I have ever achieved

Acknowledgements

During my PhD training I have spent more than 3 years in the Systems, Synthetic and Computational Biology Laboratory (diB-LAB) led by Diego di Bernardo at the TeleThon Institute of Genetics and Medicine (TIGEM). I joined the group after my graduation in computer science, in a joint PhD program between the University of Salerno and the TIGEM.
When I arrived I felt like a stranger in a strange land: until that moment I used to deal mainly with numbers, codes and programming languages and from that moment I started to tackle molecular biology problems.

However I was really lucky because in that place I found nice people, with different backgrounds, talking to each other with ease, with humility and friendship. They taught me that true science is communication, knowledge sharing and creativity. They were the diB-LAB "first-generation" members: Mukesh Bansal, Giusy della Gatta, Alberto Ambesi-Impiombato, and Giulia Cuccato. Thanks!

I started my training program with three other PhD students. We grew together, learning from each other and we had really nice parties and a lot of fun: Thanks to Vincenzo Belcastro for helping me at different stages of my project, for giving me the opportunity to collaborate with him, and... for getting drunk together in different parts of the world; thanks to Velia Siciliano for kindly introducing me to the wonderful world of biotechnology experimental tools and for being one of the most "positive" people I have ever known; thanks to Lucia Marucci (mathematics, art, music, madness, genius and candor mixed together in an explosive recipe) for having been a real friend in some difficult moments.

I received precious help and suggestions from him, and we had an incredible number of insightful discussions: thanks to Mario Lauria, senior scientist in our lab and my deskmate. Thanks also for giving to me the opportunity to work with him on NIRest and other algorithms.

During these years our lab was enriched by novel and really smart guys: Thanks to Filippo Menolascina for being the best lab-mate one could imagine and for listening to me patiently; Thanks to Gennaro Gambardella for being my personal Java trainer and "food-shopping-assistant"; Thanks to Mariaurelia Ricci and Alda Graziano for their kindness and friendship.

Thanks to Nicoletta Moretti (aka Alfia), Stefania Criscuolo, Immacolata Garzilli and Chiara Fracassi: we did not share the lab for a long time because some of them joined our group in the last months of my permanence there and others were constantly involved in wet-lab experiments, but that time was sufficient to me to understand what nice people they are.

Thanks to Gennaro Oliva, of the ICAR institute, for his kind assistance and his (massive) competence, which were of great help to me during the implementation and management of the MANTRA web-tool.

Irene Cantone did incredible work, and I found IRMA to be one of the most exciting projects I encountered since I started to work in this field. However, I principally wish to thank this "mad" girl for her genuine and unruly friendship.

Thanks to Vincenza Maselli: truly one of the most kind and good-natured people I have ever met.

Thanks to the TIGEM bioinformatic-core for its support and kind help while I was dealing with statistical tests and microarray data. Rossella Rispoli, Gopuraja Dharmalingam, Annamaria Carissimo, and Margherita

Mutarelli: thanks!
A special thank you to Luisa Cutillo for her ideas about ranked lists, which inspired my work, for being the first person that patiently worked with me when I joined the TIGEM, and for her really fun jokes.

A really great special thanks to my friend Santosh Anand!

I wish to thank Graciana Diez-Roux for very critically reading and revising the manuscript of my most important paper.

A great thank you to Nicola Brunetti-Pierri that contributed ideas and the design of the research about Fasudil: a significant effort to my results. Thanks to Pratibha Mithbaokar and Rosa Ferriero, of the Brunetti-Pierri lab, doing the experiments that confirmed one of my nicest results.

Being a member of a TIGEM group has been one of the most stimulating experiences of my life. I wish to thank Maria Pia Cosma, who gave me the opportunity to collaborate with people in his lab in really great projects, and the other group leaders that involved me in their work: Alberto Auricchio, Giancarlo Parenti, and Alberto Luini. A great thank you to Seetharaman Parashuraman for the same reasons.

Working at TIGEM has been great and comfortable also because of the nice people working in the administration and human resources: Silvana Ruotolo, Federico Barone, Brunella Summaria, Barbara zimbardi, Mariolina Pepe et al. Thanks!

... And How to forget our IT core technicians? Giampiero Lago, Marco Savarese, Giancarlo Sambrini and Mario Traditi: Thanks for your great work, your kind assistance and all the nice chats and funny jokes.

What makes TIGEM a special place is each single person working there: thanks to Signor Agostino, Antonio and Dina. Your smiles were the best way to start each working day!

I want to thank my housemate Carmine Spampanato for having kindly put up with me for almost three long years.

Last but (obviously) not least a huge thank to Diego di Bernardo (my supervisor, at the TIGEM): literally the best mentor that anyone trying to do science could have!
He has an incredible ability in helping young scientists to discover (and to do) what they do best. He saw some little talent in me and cheered me up even in the most discouraging moments.
THANKS!

At the University of Salerno I have been working in the Neural and Robotic Network (NeuRoNe) Laboratory led by Prof. Roberto Tagliaferri and Prof. Giancarlo Raiconi.
I met these two Professors when I was an undergraduate and I wish to thank them for insightfully introducing me to the world of Machine Learning, Data Mining, Complex Systems and Neural Networks.
Moreover I gratefully thank them for their support, encouragement and the

inspiring discussions we had during my periodical reports.

I wish to thank Loredana Murino (PhD student at the NeuRoNe lab) for her great work on the GO:Fuzzy-Enrichment analysis.

Thanks to Francesco “Ciccio” Napolitano: a real friend and one of the most intelligent and stimulating people I ever met.

Thanks to the people of the lab for their friendship and the time spent together: Andrea Raiconi, Carmine Cerrone, Donatella Granata, Ekaterina Nosova, Ivano Scoppetta (aka Vittorio Santoro), Francesco Carrabs and Ida Bifulco.

A huge thank you to Antonella Isacchi, Roberta Bosotti, and Emanuela Scacheri of the Nerviano Medical Science conceiving the blind test of my method, for producing novel Microarray-Data for me and performing the experiments validating some of my most original and interesting results.
Additionally, thanks to them for their great contribution to my paper, for their help in interpreting MANTRA results, their great expertise in oncology and the extremely stimulating chats we had via Skype, Phone and Email.
A great team: it has been really a pleasure to work with them.
THANKS Nerviane!

A special thanks to Dr. Julio Saez-Rodriguez for hosting me in his laboratory at the European Bionformatics Institute in the last months of my PhD, for giving me the opportunity of joining a great group and the chance

of pursuing other scientifically exciting results together in the coming years.

Thanks to my new lab-mates (both the temporary and the permanent ones): Beatriz Penalver, Ioannis Melas, Camille Terfve, Jordi Serra i Musach, Aidan MacNamara, Jerry Wu and David Henriques.

Thanks to Gabriella Rustici for her kind help when myself and my family were searching for accommodation in Cambridge and for her friendship.

Now it is the turn of the most important people in my life...

First of all, I wish to thank my parents, who always supported me in every possible way. I believe that they should be cited as co-authors of this and all the other successes in my life. Mum and Dad: I love you.

Secondly, I want to thanks my in-laws: they were literally my second parents in these last few years and helped the little new family that myself and my wife were composing with infinite love and patience. Nonna Brenda and Nonno Pasquale: a huge THANK YOU and a hug!

A huge hug and a thank you to my "brothers and sisters":
Ylenia, Giuseppe (the greatest mathematician I have ever known) and Raffaella, Davide and Annarita, for their love and all the funny moments together.

There are no words to describe my love for you nor do numbers exist to quantify it. You have been my force and together the source of all my happiness. To say "thank you" would be improper, as is improper (and impossible) to list all the reasons why I should do it. Annalisa, I love you and I am so happy you are my wife.

Finally, to you: I hope that your eyes will always be full of this curiosity and vivacity, and that you will look at me always as you do now; I hope to deserve your love for ever and that your smiles will be always so real and happy.
With Love, Dad :)

Abstract

The identification of the molecular pathway that is targeted by a compound, combined with the dissection of the following reactions in the cellular environment, i.e. the drug mode of action, is a key challenge in biomedicine.
Elucidation of drug mode of action has been attempted, in the past, with different approaches. Methods based only on transcriptional responses are those requiring the least amount of information and can be quickly applied to new compounds. On the other hand, they have met with limited success and, at the present, a general, robust and efficient gene-expression based method to study drugs in mammalian systems is still missing.
We developed an efficient analysis framework to investigate the mode of action of drugs by using gene expression data only. Particularly, by using a large compendium of gene expression profiles following treatments with more than 1,000 compounds on different human cell lines, we were able to extract a synthetic consensual transcriptional response for each of the tested compounds. This was obtained by developing an original rank merging procedure. Then, we designed a novel similarity measure among the transcriptional responses to each drug, endingending up with a "drug similarity network", where each drug is a node and edges represent significant similarities between drugs.
By means of a novel hierarchical clustering algorithm, we then provided this network with a modular topology, contanining groups of highly interconnected nodes (i.e. network communities) whose exemplars form second-level modules (i.e. network rich-clubs), and so on. We showed that these topological modules are enriched for a given mode of action and that the hierarchy of the resulting final network reflects the different levels of similarities among the composing compound mode of actions.
Most importantly, by integrating a novel drug X into this network (which

can be done very quickly) the unknown mode of action can be inferred by studying the topology of the subnetwork surrounding X. Moreover, novel potential therapeutic applications can be assigned to safe and approved drugs, that are already present in the network, by studying their neighborhood (i.e. drug repositioning), hence in a very cheap, easy and fast way, without the need of additional experiments.

By using this approach, we were able to correctly classify novel anti-cancer compounds; to predict and experimentally validate an unexpected similarity in the mode of action of CDK2 inhibitors and TopoIsomerase inhibitors and to predict that Fasudil, a known and FDA-approved cardiotonic agent, could be repositioned as novel enhancer of cellular autophagy.

Due to the extremely safe profile of this drug and its potential ability to traverse the blood-brain barrier, this could have strong implications in the treatment of several human neurodegenerative disorders, such as Huntington and Parkinson diseases.

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

List of Algorithms

PARVA
LIBELLUM
SUSTINE
PATIENTIA

1

Introduction

A bottleneck in drug discovery is the identification of the molecular targets of a compound and of its off-target effects (Figure 1.1).

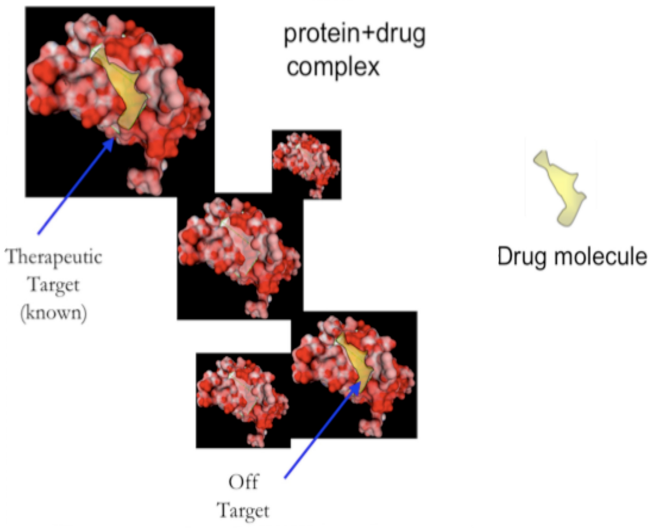

Figure 1.1: Discovery of drug mode of action -

The recognition of a set of interacting genes, proteins and metabolites (i.e. a biological pathway) whose activity is modulated by the drug treatment, combined with the dissection of the resulting reactions in the cellular environment (Figure 1.2), is nowadays a key challenge in biomedicine.

Addressing these problems means to investigate drug Mode of Action (MoA).
On the other hand, the detection of the complex regulatory relationships occurring

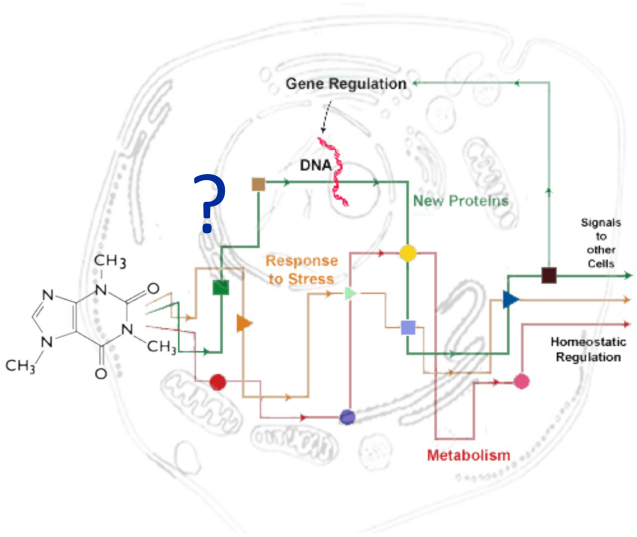

Figure 1.2: Reactions to drug-substrate interaction -

among genes (i.e. gene regulatory network) is of major importance in order to understand the working mechanisms of the cell in patho-physiological conditions and it also allows the discovery of novel drug targets.

Thanks to the development of new experimental methods and to the constant decrease of the data-storage costs, current technology allows the analysis of massive quantities of data. One of the scientific fields that are taking great advantages of this capabilities is molecular biology in which a key role is played by DNA-microarray technology, in the context of functional genomics.

Each single cell contains a copy of the entire genome of the organism to which it belongs.

The genome is composed by DNA molecules and it contains the set of informations needed for the transmission of the hereditary factors and the protein synthesis. Once a gene is "activated" a corresponding molecular intermediate, the messenger RNA (mRNA), is generated through a process called "transcription" and released in the cytoplasm (the thick liquid residing among the nucleus, the cellular membrane and the organelles). Here, the mRNA is translated into proteins through the assembly of amino acids by a ribosomes.

The amount of mRNA equivalent to the DNA sequence of a given gene, in a give instant, quantifies the "level of expression" of that gene.

A DNA-microarray is a tool able to measure gene expression levels at a genome-wide scale simultaneously (i.e. it is able to produce a genome-wide Gene Expression Profile (GEP)). By using DNA-microrrays it is possible to monitor the expression of all the genes in the cells of a given tissue in pathological conditions or to measure how cells respond to a pharmacological treatment at a transcriptional level.

Even if elucidation of drug MoAs has been attempted, in the past, with different approaches, the drug discovery pipeline typically has been guided by knowledge of the biological mechanisms underlying the disease to treat. Based on this knowledge, "druggable" molecular targets have been hypothesized and libraries of chemical structures have been systematically analyzed in order to find drug candidates on the basis of their chemical "affinity" with the desired target.

More recently, alternative approaches such as high-throughput screening of drug libraries have been developed to allow identification of molecules acting on specific cellular targets experimentally. However, they are generally based on assays or binding studies that focus narrowly on the molecular target, not taking into account the complexity of the cell response.

Strategies based on the analysis of drug-induced changes in gene expression profiles have the potential to elucidate the cellular response to specific drugs. Moreover, methods based only on transcriptional responses are those requiring the least amount of experiments and can be quickly applied to new compounds. On the other hand, they have met with limited success and, at the present, a general, robust and efficient gene-expression based method to study drugs in mammalian systems is still missing.

Other important problems are linked to the concept of "drug repositioning": the large number of drug candidate failures has been enormously costly for the pharmaceutical industry, but has also created the opportunity of re-purposing these molecules for therapeutic applications into new disease areas. Companies which can systematically "reposition" unsuccessful drug candidates could create significant value by reinforcing their pipeline (the set of drugs under development or in testing) and meeting the needs of innovative medicines.

In this PhD thesis we present the methodology and the results of our research, which has been focused on the development of a novel and efficient analysis framework to investigate the MoA of new drugs by using gene expression data only and for suggesting novel therapeutic uses of well-known and already approved drugs.

Our leading assumption was a simple but strong postulate: if two drugs elicit similar effects on the transcriptional activity of the cell then they could share a MoA (and possibly a therapeutic application) even if they act on distinct intracellular direct targets (point 1 in Figure 1.3).

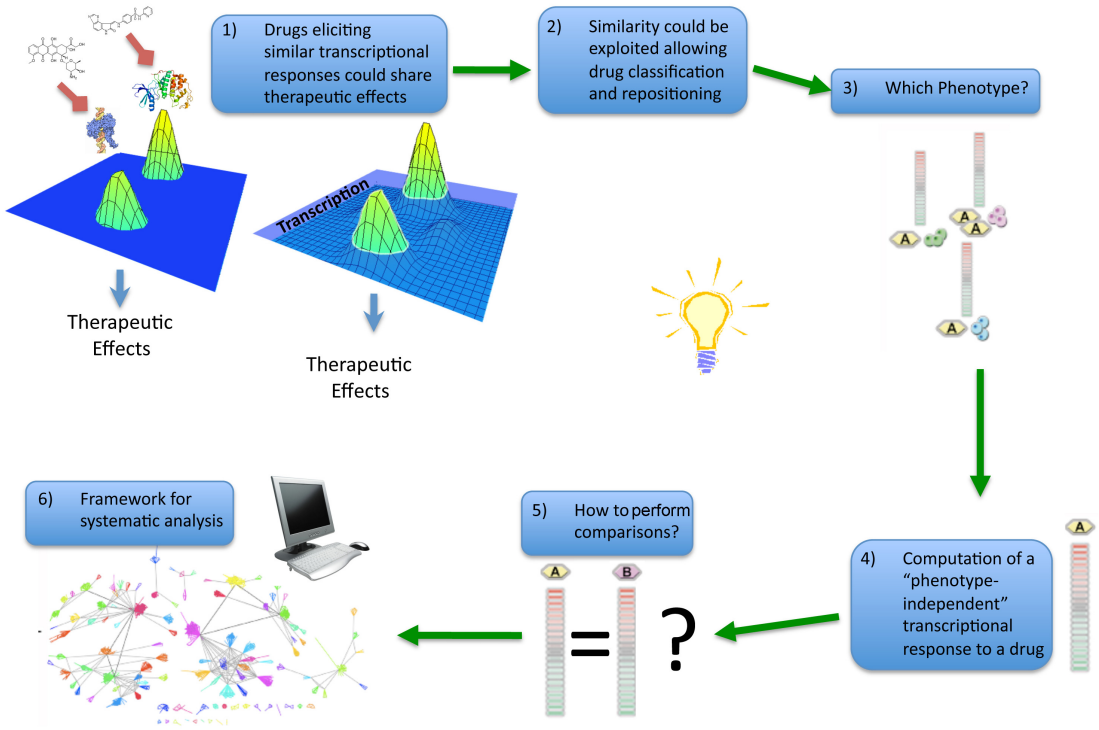

Figure 1.3: Project: Leading ideas and problems -

As a consequence, similarities in the transcriptional responses to drugs could be exploited allowing drug classification and repositioning (i.e. re-purposing for novel uses) (point 2 in Figure 1.3). The first problem we had to tackle was due to a phenomenon known for microarray studies: cells grown at the same time and in the same experimental setting tend to respond similarly at a transcriptional level even if they are differently stimulated. In other word similarity of gene expression profiles can be recorded for unrelated stimuli in the same experimental setting (also called batch effect)(81). On the other hand, cells in different pathological conditions obey to the rules of the transcriptional program in the corresponding disease phenotype (the condition characteristics) hence they tend to respond differently to the same drug treatment. Consequently, poor results can be achieved with classic micro-array analysis approach, which tends to dis-

criminate gene expression profiles on the basis of the experimental settings (kind of cells, observation time, microarray platform) in which they have been produced rather than on the basis of the stimuli they are responding to (for example a drug treatment) (point 3 in Figure 1.3).
We addressed this problem by using a large compendium of gene expression data following treatments with more than 1,000 compounds on different human cell lines, being able to compute a synthetic consensual transcriptional response for each of the tested compounds. This response is a proxy of a "phenotype independent" transcriptional response and we considered it a sufficiently general summary of the drug MoA (point 4 in Figure 1.3). This was obtained by using a novel and original data merging procedure. In order to pair-wise compare the drugs in our reference dataset (point 5 in Figure 1.3) we conceived a novel similarity measure among these responses, which was based on a non-parametric statistic, ending up to a "drug similarity network" (point 6 in Figure 1.3).
Finally we used this network as a classification template and as a predictor of drug candidates for drug repositioning, a task which has been growing in importance in the last few years as an increasing number of drug development and pharmaceutical companies see their drug pipelines drying up. To assess our results, novel experimental data were produced on purpose in order to validate computational predictions.

1.1 Outline

This thesis in computer science describes a computational approach to a practical problem of drug discovery, which has been tackled based on principles of molecular biology and making use of available biomedical data.
Moreover, a number of experiments has been conducted with different experimental tools in order to produce de-novo data and to verify computational results. Some biological concepts and principles together with the background informations needed to understand the experiment outcomes are provided in Chapter 2.
In the same chapter we covered some concepts from complex network theory and traditional drug discovery. In Chapter 3 we briefly discuss gene expression and systems biology approaches to drug discovery.
In 4 we exhaustively describe the design of our analysis framework while in Chapter 5

we present Mode of Action by Network Analysis (MANTRA), the online implementation of our method, and we describe the drug classification algorithm.
In Chapter 6 we present the results that were obtained while testing our method in classifying novel drugs and their experimental validation.
Chapter 7 contains the description of a drug repositioning proposal predicted by our method. Future directions and a final discussion are presented in Chapter 8.

Part of the work described in this thesis has been published in:

- Iorio F, Isacchi A, di Bernardo D, Brunetti-Pierri N.
  Identification of small molecules enhancing autophagic function from drug network analysis.
  Autophagy. 2010 Nov 16; 6(8): 1204-5.

- Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A, di Bernardo D.
  Discovery of drug mode of action and drug repositioning from transcriptional responses.
  Proc Natl Acad Sci U S A. 2010 Aug 17; 107(33): 14621-6.

- Iorio F, Murino L, di Bernardo D, Raiconi G, Tagliaferri R.
  Gene ontology fuzzy-enrichment analysis to investigate drug mode-of-action
  Neural Networks (IJCNN), The 2010 International Joint Conference on. 2010 July: 1-7.

- Cantone I, Marucci L, Iorio F, Ricci MA, Belcastro V, Bansal M, Santini S, di Bernardo M, di Bernardo D, Cosma MP.
  A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches.
  Cell. 2009 Apr 3; 137(1): 172-81.

- Lauria M, Iorio F, di Bernardo D.
  NIRest: a tool for gene network and mode of action inference.
  Ann N Y Acad Sci. 2009 Mar; 1158: 257-64.

- Iorio F, Tagliaferri R, di Bernardo D.
  Identifying network of drug mode of action by gene expression profiling.
  J Comput Biol. 2009 Feb; 16(2): 241-51.

- Iorio F, Tagliaferri R, di Bernardo D.
  Building Maps of Drugs Mode-of-Action from Gene Expression Data
  Computational Intelligence Methods for Bioinformatics and Biostatistics. **LNCS**. 2009, Volume 5488/2009, 56-65.

Supplementary data sheets referred in the text are contained in the **Supplementary Data Disc (SDD)** attached to this thesis.

2

Background

2.1 Introduction

This chapter contains basic informations, needed to fully understand the biology underlying the main presented results the experimental data that we analyzed and produced "de-novo", and a short description of the experimental platforms that we used to verify our results (Section 2.2).

In Section 2.3 some definitions and methods of network theory that we used while designing our computational approach are listed.

In the final section computational drug discovery is briefly discussed and existing applications of network analysis in this field are introduced.

The content of this chapter (figures and some portions of text) are from the following web resources:

`www.nigms.nih.gov`,

`www.ebi.ac.uk`,

`www.wordiq.com`,

`www.bio.davidson.edu`,

`www.molegro.com`.

2.2 Molecular biology: basic principles and techniques

2.2.1 Overview of the Cell

A cell is the simplest and most elementary functional basic unit of life and it can be considered as the building block of all living beings. Some organisms, such as most bacteria, consist of a single cell (unicellular organisms) while others, such as humans, are composed by about 100 trillion of cells.

In an "eukaryotic" cell (see Figure 2.1) the nucleus is a membrane enclosed organelle that can occupy up to 10 percent of the cellular space. It contains the equivalent of the cell's "program", its genetic material, the Deoxyribonucleic acid (DNA). DNA contains the instructions used in the development and functioning of all known living organisms with the exception of some viruses. The main role of DNA molecules is the long-term storage of information. DNA is often compared to a set of blueprints, like a recipe or a code, since it contains the instructions needed to construct other components of cells, such as proteins and Ribonucleic acid (RNA) molecules. The DNA segments that carry this genetic information are called genes, but other DNA sequences have structural purposes, or are involved in regulating the use of this genetic information.

The nucleus is surrounded by two pliable membranes, together known as the nuclear envelope. Normally, the nuclear envelope is pockmarked with octagonal pits and hemmed in by raised sides. These nuclear pores allow chemical messages to exit and enter the nucleus.

Between the cell membrane (a selectively-permeable phospholipidic layer) and the nuclear envelope resides a thick and clear liquid called the cytoplasm. The cell's outer membrane is made up of a mix of proteins and lipids (fats). Lipids give membranes their flexibility. Proteins transmit chemical messages into the cell, and they also monitor and maintain the cell's chemical climate.

On the outside of cell membranes, attached to some of the proteins and lipids, are chains of sugar molecules that help each cell type do its job.

Close to the nucleus resides a groups of interconnected sacs snuggling close by. This network of sacs, the Endoplasmatic Reticulum (ER), often makes up more than 10 percent of a cell's total volume.

Made up of more than 70 proteins and 4 strands of RNA (a chemical relative, of DNA

that we will describe later), ribosomes have a critical job: assembling all the cell's proteins. To make a protein, ribosomes weld together chemical building blocks one by one (as explained in the following sections).
Another important component of the cellular endomembrane system (the set of different membranes that are suspended in the cytoplasm) is the Golgi apparatus. Composed of stacks of membrane-bound structures known as cisternae, the Golgi apparatus processes and packages macromolecules, such as proteins and lipids, after their synthesis and before they make their way to their destination; it is particularly important in the processing of proteins for secretion.
The waste disposal system of the cell is composed by the lysosomal machinery. Lysosomes are cellular organelles which contain acid hydrolase enzymes to break up waste materials and cellular debris.

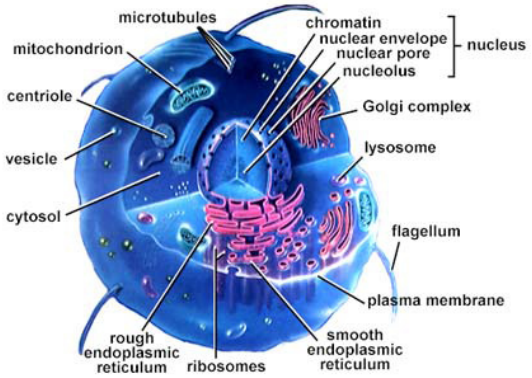

Figure 2.1: The cell - [Image from: http://www.ebi.ac.uk]

The subtle movements of the cell as well as the many chemical reactions that take place inside organelles require vast amounts of cellular energy. The main energy source of the cell is a small molecule called Adenosine-5'-triphosphate (ATP). ATP is often referred as the "molecular unit of currency" of intracellular energy transfer because it transports chemical energy within cells for metabolism. It is produced by membrane-enclosed organelles called mitochondria. ATP is used by enzymes and structural proteins in many cellular processes, including biosynthetic reactions, motility, and cell division. Among these processes one of the most important is phosphorilation.
Phosphorylation is the transfer of a phosphate (PO_4) group from a high-energy donor

molecule (such as ATP) to a protein (in this case, a substrate). Phosphorylation is conducted by specific enzymes called phosphotransferases or kinases, it activates or deactivates many protein enzymes, causing or preventing the mechanisms of diseases such as cancer and diabetes. Protein phosphorylation in particular plays a significant role in a wide range of cellular processes and usually it results in a functional change of the target protein by changing enzyme activity, cellular location, or association with other proteins.

The series of events that takes place in a cell leading to its division and duplication (replication) is called the cell cycle, or cell-division cycle. Cell cycle is tightly regulated by the activity of a group of protein kinases, i.e. Cyclin-Dependent kinases (CDKs). A CDK is activated by association with a cyclin, forming a cyclin-dependent kinase complex. Cyclins are proteins whose concentrations varies in a cyclical fashion during the cell cycle. The oscillations of the cyclins, namely fluctuations in cyclin gene expression and destruction by proteolysis, induce oscillations in CDK activity to drive the cell cycle.

A normal component of the development and health of multicellular organisms is cellular apoptosis, or programmed cell death. Cells die in response to a variety of stimuli and during apoptosis they do so in a controlled, regulated fashion. This makes apoptosis distinct from another form of cell death called necrosis in which uncontrolled cell death leads to lysis of cells, inflammatory responses and, potentially, to serious health problems. Apoptosis, by contrast, is a process in which cells play an active role in their own death (which is why apoptosis is often referred to as cell suicide).

2.2.2 DNA structure and function

DNA is the main information carrier molecule in a cell. A single stranded DNA molecule, also called a polynucleotide, is a chain of small molecules, called nucleotides (see Figure 2.2). There are four different nucleotides grouped into two types, purines: adenine and guanine and pyrimidines: cytosine and thymine. They are usually referred to as bases and denoted by their initial letters, A, C, G and T. Different nucleotides can be linked together in any order to form a polynucleotide. The end of the polynucleotides are marked either 5' and 3' . By convention DNA is usually written with 5' left and 3' right, with the coding strand at top. Two such strands are termed complementary, if one can be obtained from the other by mutually exchanging A with T and C with G,

and changing the direction of the molecule to the opposite. Although such interactions are individually weak, when two longer complementary polynucleotide chains meet, they tend to stick together.

Two complementary polynucleotide chains form a stable structure, which resembles a helix and is known as a the DNA double helix. About 10 base pairs (bp) in this structure takes a full turn, which is about 3.4 nm long. Complementarity of two strands in the DNA is exploited for copying (multiplying) DNA molecules in a process known as the DNA replication, in which one double stranded DNA is replicated into two identical ones. (The DNA double helix unwinds and forks during the process, and a new complimentary strand is synthesized by specific molecular machinery on each branch of the fork. After the process is finished there are two DNA molecules identical to the original one.) In a cell this happens during the cell division and a copy identical to the original goes to each of the new cells.

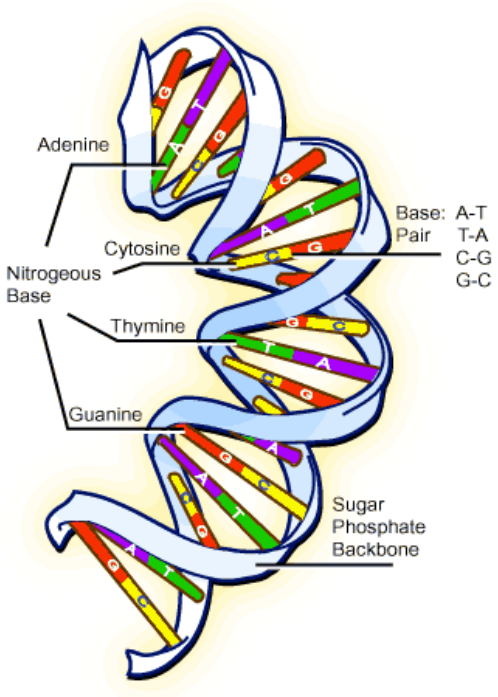

Figure 2.2: DNA - [Image from: http://www.scq.ubc.ca]

During replication an enzyme called Topoisomerase (Topo) prevents DNA tangling and damaging. As a replication fork moves along double-stranded DNA, it creates what

has been called the "winding problem". Every 10 bp replicated at the fork corresponds to one complete turn about the axis of the parental double helix. Therefore, for a replication fork to move, the entire chromosome ahead of the fork would normally have to rotate rapidly. This would require large amounts of energy for long chromosomes, and an alternative strategy is used instead: a swivel is formed in the DNA helix by DNA Topo.

DNA winds around proteins called histones. These proteins play an important role in gene regulation in eukaryotic cells and they are are highly water soluble. The six histone classes are H1, H2A, H2B, H3, H4, and Archaeal. All but the H1 and Archaeal classes create nucleosome core particles by wrapping DNA around their protein spools; the H1 then binds nucleosomes and entry and exit sites of the DNA. Histones and DNA assembled in this way are called chromatin. Packed in this way DNA are 50,000 times shorter than unpacked ones. Histones also perform a function in gene regulation; their "methylation" (modification of certain amino acids by the addition of one, two, or three methyl groups) causes tighter bindings to down-regulate or even inhibit gene transcription, while "acetylation" (addition of acetyl groups) loosens bindings to help encourage transcription and translation.

Histone proteins are packaged into structures called chromosomes. Chromosomes are not visible in the cells nucleus, not even under a microscope, when the cell is not dividing. However, the DNA that makes up chromosomes becomes more tightly packed during cell division and is then visible under a microscope. Most of what researchers know about chromosomes was learned by observing chromosomes during cell division. Each chromosome has a constriction point called the centromere, which divides the chromosome into two sections, or "arms". The short arm of the chromosome is labeled the "p arm". The long arm of the chromosome is labeled the "q arm". The location of the centromere on each chromosome gives the chromosome its characteristic shape, and can be used to help describe the location of specific genes.

Together with DNA and proteins, RNA is one of the major macromolecules that are essential for all known forms of life. The sequence of nucleotides composing a molecule of RNA allows it to encode genetic information. For example, some viruses use RNA instead of DNA as their genetic material, and all organisms use messenger messenger

RNA (mRNA) to carry the genetic information that directs the synthesis of proteins. This process can be divided into two parts (as summarized in Figure 2.3):

- *Transcription*: Before the synthesis of a protein begins, the corresponding RNA molecule is produced by RNA transcription. One strand of the DNA double helix is used as a template by the RNA polymerase to synthesize an mRNA. This mRNA migrates from the nucleus to the cytoplasm. During this step, mRNA goes through different types of maturation including one called splicing when the non-coding sequences are eliminated. The coding mRNA sequence can be described as a unit of three nucleotides called a codon;

- *Translation*: The ribosome binds to the mRNA at the start codon (AUG) that is recognized only by the initiator Transfer RNA (tRNA) (a small RNA molecule that transfers a specific active amino acid to the ribosome). The ribosome proceeds to the elongation phase of protein synthesis. During this stage, complexes, composed of an amino acid linked to tRNA, sequentially bind to the appropriate codon in mRNA by forming complementary base pairs with the tRNA anticodon. The ribosome moves from codon to codon along the mRNA. Amino acids are added one by one, translated into polypeptidic sequences dictated by DNA and represented by mRNA (Figure 2.4). At the end, a release factor binds to the stop codon, terminating translation and releasing the complete polypeptide from the ribosome.

The rule that deals with the detailed residue-by-residue transfer of sequential information establishing that information cannot be transferred back from protein to either protein or nucleic acid is known as the "central dogma of molecular biology". In other words, once information gets into protein, it can't flow back to nucleic acid.
Like shoelaces, the polypeptidic sequences released by the ribosomes loop about each other in a variety of ways (i.e., they fold). But, as with a shoelace, only one of these many ways allows the protein to function properly. Yet lack of function is not always the worst scenario. For just as a hopelessly knotted shoelace could be worse than one that wont stay tied, too much of a misfolded protein could be worse than too little of a normally folded one. This is because a misfolded protein can actually poison the cells around it.

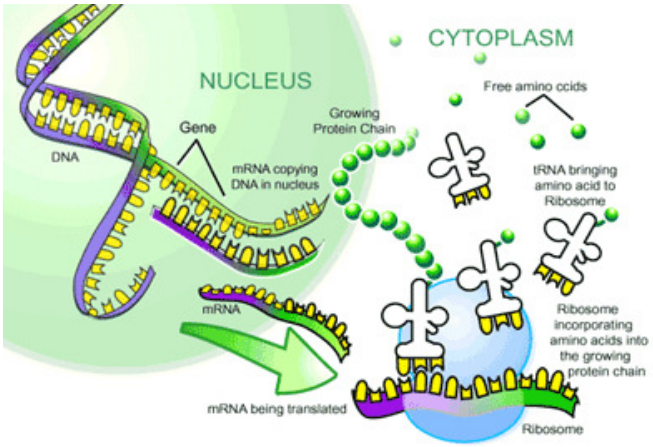

Figure 2.3: The central dogma of molecular biology - [Image from: http://www.scq.ubc.ca]

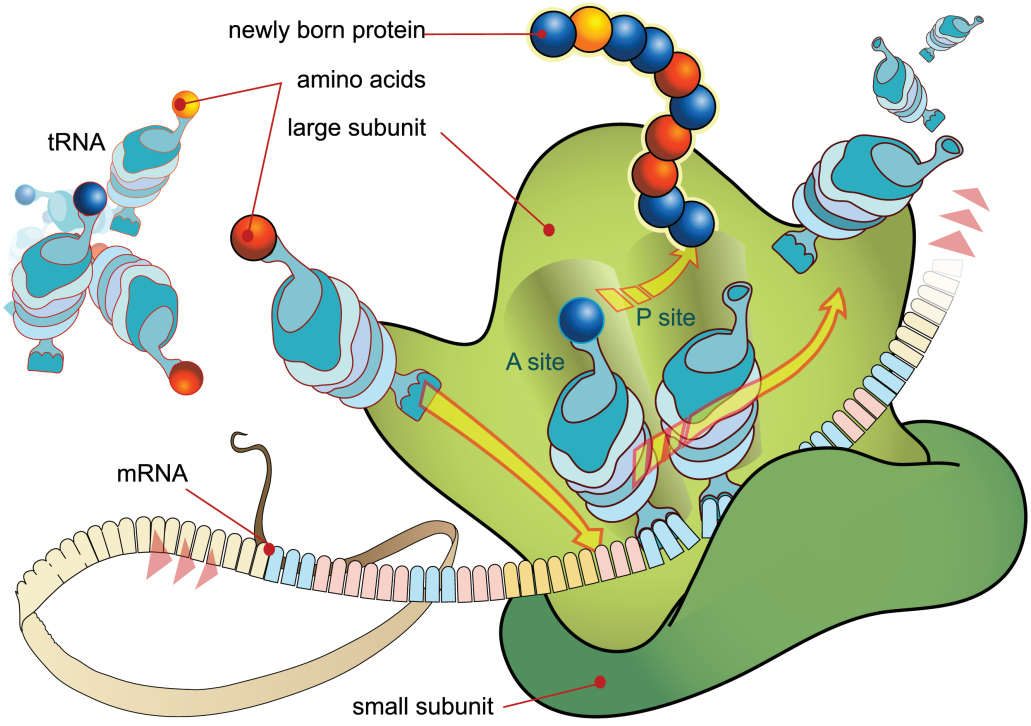

Figure 2.4: Protein synthesis - [Image from: http://www.wikipedia.org]

2.2.3 Gene Expression and Regulation

Living cells are the product of gene expression programs involving "regulated" transcription of thousands of genes. The central dogma, briefly introduced in the previous section, defines a paradigm in molecular biology: genes are perpetuated as sequences of nucleic acid and translated in functional units, i.e. the proteins. In these process mRNA provides a molecular intermediate that carries the copy of a DNA sequence that represents a protein. It is a single-stranded RNA identical in sequence with one of the strands of the duplex DNA. In protein-coding genes, translation will convert the nucleotide sequence of mRNA into the sequence of amino acids comprising a protein. This transformation of information (from gene to gene product) called gene expression. Gene expression is a complex process regulated at several stages by other biological phenomenon.

For example, a large group of proteins play an important role by this point of view. These proteins are known as transcription factors and they can regulate the expression of a gene in a positive or a negative sense. In positive regulation, an "excitatory" protein binds to the promoter (usually a region of the DNA up-streaming the gene sequence), and increases (or activates) the level of mRNA transcribed for that gene (as summarized in Figure 2.5). Some other transcription factors exert a negative regulation by decreasing the mRNA transcription rate of a gene.

Several other aspect of the gene expression process may be modulated. Apart from DNA transcription regulation, the expression of a gene may be controlled during RNA processing and transport (in eukaryotes), RNA translation, and the post-translational modification of proteins. The degradation of gene products can also be regulated in the cell. Recently, RNA has been discovered to play a direct role in regulation of gene expression and it is known that small RNA molecules can act, through RNA interference mechanism, as "silencers" of gene expression (see (28, 40)).

The different cellular components (mRNA, proteins and DNA) compose complex hierarchical networks of interactions that regulate and supervise all the cellular processes, and among these its "transcriptional program". Particularly, the levels of interactions in which gene expression activity tightly regulate itself are referred as "transcriptional networks" or "gene expression networks".

In these networks the interacting entities are genes whose product act as transcriptional

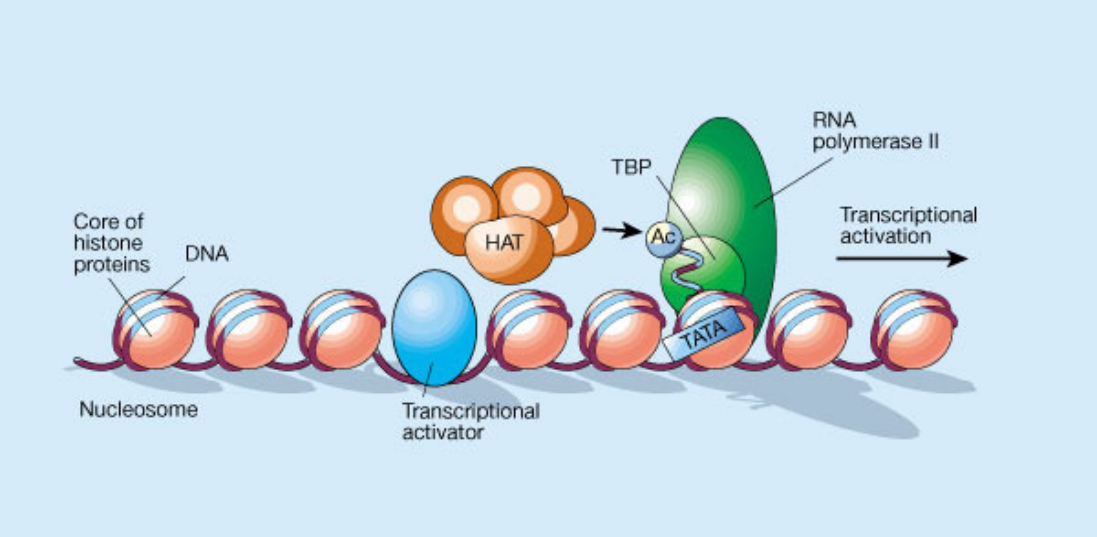

Figure 2.5: A view of gene regulation - [Image from: http://www.nature.com]

factor (or regulates in other ways) the transcriptional activity of other genes (i.e. target genes).

2.2.4 How to measure gene expression level

Measuring gene expression is an important task in several fields of life sciences and the ability to quantify the level at which a particular gene is expressed within a cell, tissue or organism can give a huge amount of information.
When dealing with a small number of genes a possible option for measuring their expression levels consists in using realtime Polymerase Chain Reaction (PCR). Often referred also as qPCR or qrt-PCR, realtime PCR is used to amplify and simultaneously quantify a targeted DNA molecule. It enables both detection and quantification (as absolute number of copies or relative amount when normalized to DNA input or additional normalizing genes) of one or more specific sequences in a DNA sample.
When used for quantifying the level of expression of a given gene, real-time PCR is combined with reverse transcription and actually complementary DNA (cDNA) is quantified.
The procedure follows the general principles of polymerase chain reaction. The starting point is a portion of the sequence of the DNA (or cDNA) molecule that one wishes

to replicate and "primers": short oligonucleotides (containing about two dozen nucleotides) that are precisely complementary to the sequence at the 3' end of each strand of the DNA to amplify (as depicted in Figure 2.6).

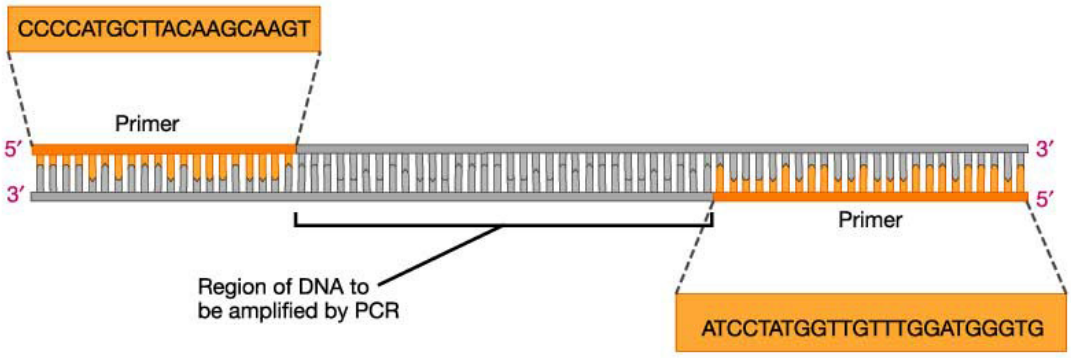

Figure 2.6: Polymerase chain reaction - [Image from: http://campus.queens.edu]

In a series of iterative step (PCR cycles) the DNA samples are heated to separate their strands and mixed with the primers. If the primers find their complementary sequences in the DNA, they bind to them. Synthesis begins (as always from 5' to 3') using the original strand as the template (see Figure 2.8).

This "polymerization" continues until each newly-synthesized strand has proceeded far enough to contain the site recognized by the other primer. Now there are two DNA molecules identical to the original molecule. Their are heated, separated into their strands, and the whole process repeat. In conclusion, each cycle doubles the number of DNA molecules of the previous cycle (see Figure 2.8).

Fluorescent reporter probes detect only the DNA containing the sequence of inter-

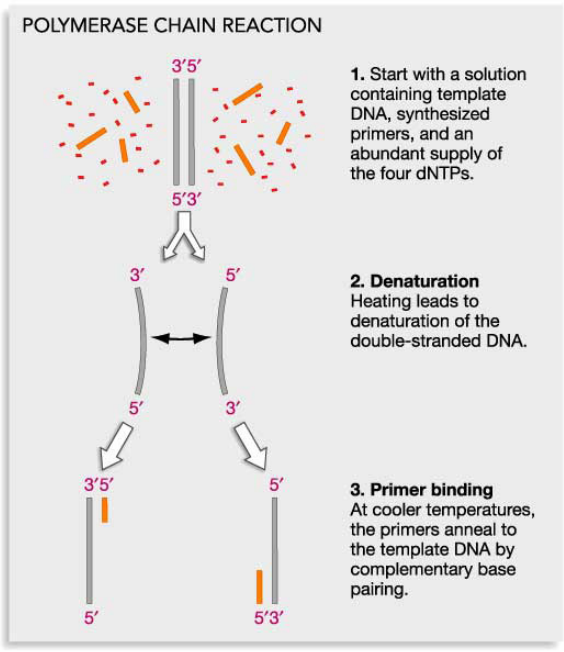

Figure 2.7: Polymerase chain reaction - [Image from: http://campus.queens.edu]

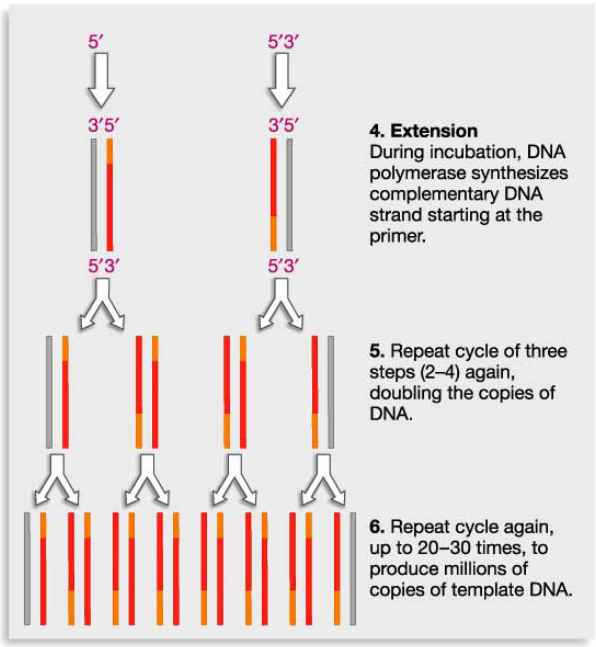

Figure 2.8: Polymerase chain reaction - [Image from: http://campus.queens.edu]

est thus allowing the detection of relative concentrations of DNA present during the reactions, which are determined by plotting fluorescence against cycle number, as summarized in Figure 2.9. A threshold for detection of fluorescence above background is

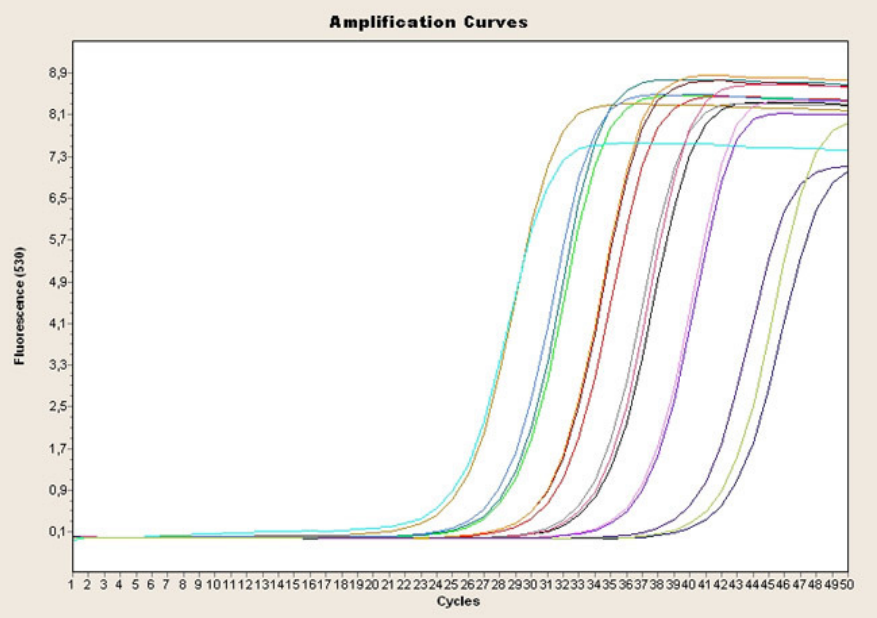

Figure 2.9: Realtime PCR outcomes -

determined. The cycle at which the fluorescence from a sample crosses the threshold is called the cycle threshold, C_t. The quantity of DNA theoretically doubles every cycle during the exponential phase and relative amounts of DNA can be calculated, for example a sample whose C_t is 3 cycles earlier than another's has $2^3 = 8$ times more template. Since all sets of primers don't work equally well, one has to calculate the reaction efficiency first. Thus, by using this as the base and the cycle difference C_t as the exponent, the precise difference in starting template can be calculated.

Amounts of RNA or DNA are then determined by comparing the results to a standard curve produced by realtime PCR of serial dilutions of a known amount of RNA or DNA. As mentioned above, to accurately quantify gene expression, the measured amount of RNA from the gene of interest is divided by the amount of RNA from a housekeeping gene (i.e. a gene that is constitutively expressed) measured in the same sample to normalize for possible variation in the amount and quality of RNA between different samples. This normalization permits accurate comparison of expression of the

gene of interest between different samples, provided that the expression of the reference (housekeeping) gene used in the normalization is very similar across all the samples.
Measuring gene expression with this method for a large number of genes or at a genome-wide scale is infeasible and other tools, such as DNA microarray, are used for this case.
A DNA microarray (or DNA chip) is basically a set of microscopic fragments (i.e. spots) of DNA oligonucleotides on a solid surface of roughly 1 cm^2. These fragments are divided in about 250.000 cells (depending from the platform model) and each of these cells contains millions of copies of a specific sequence of DNA (i.e. probes). These can be a short section of a gene or other DNA element that are used to hybridize (i.e. permanent bind) complementary sequences of cDNA.
cDNA is DNA synthesized from a mature mRNA template in a reaction catalyzed by the enzyme reverse transcriptase and the enzyme DNA polymerase. DNA microarrays can be used to measure changes in expression levels, to detect Single Nucleotide Polymorphisms (SNPs), or to genotype or re-sequence mutant genomes.
Here we focus on microarrays as tools to simultaneously measure the level of expression of thousands of genes (i.e. genome-wide measurement). In a typical microarray experiment like this, the nucleic acid of interest (in this case RNA) is purified and isolated (total as it is nuclear and cytoplasmic). After a quality control of the RNA a labelled product cDNA is generated via reverse transcription. The labeling is typically obtained by tagging fragments with fluorescent dyes. Finally, the labeled samples are then mixed, denatured and added to the microarray surface. Here the cDNA fragments hybridizes with the corresponding complementary DNA spot forming a double helix structure. At this point, the fragments that did not hybridized are washed away and it is possible to count the number of formed double helix thus allowing the quantification of the corresponding RNA levels hence the level of expression of the corresponding coding genes. This counting is achieved after drying the microarray, by using a special machine where a laser excites the dye and a detector measures its emission. The whole process is summarized in Figure 2.10.

In a cDNA microarray experiment fragments are labelled with different dyes (thus different colors) depending on the biological condition they come from (i.e. a cell culture grown in an condition of interest such as, for example, the treatment with a drug or in an anaerobic environment, and in a “normal” condition, respectively) (see Figure

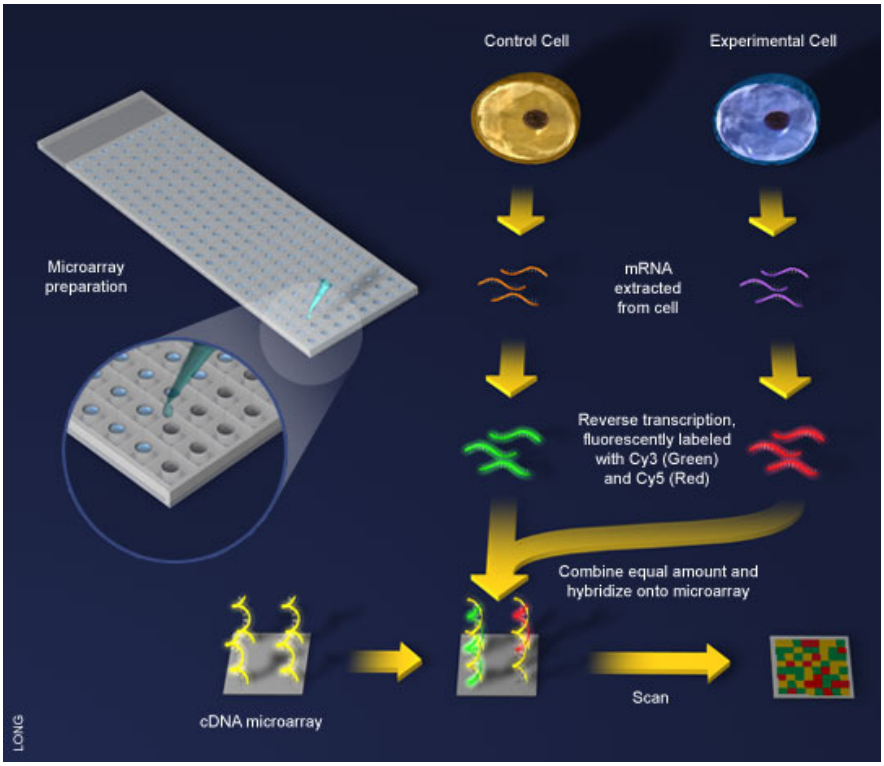

Figure 2.10: cDNA Microarray technology - [Image from: http://www.columbia.edu]

2.10). The final measurement is a "differential" expression value for each probe, quantifying wether the gene corresponding to a given probe is expressed the more in the condition of interest, or in the normal one, or in both of them.

In oligonucleotide microarrays (for example, Affymetrix gene chips), hybridized samples come from just one biological condition during an experiments (see Figure 2.11). Once labeled, the sample of cRNAs can be hybridized to the array and measured values are not differential but absolute. In order to compute differential expression values hybridizations of the control condition samples are needed on different chips.

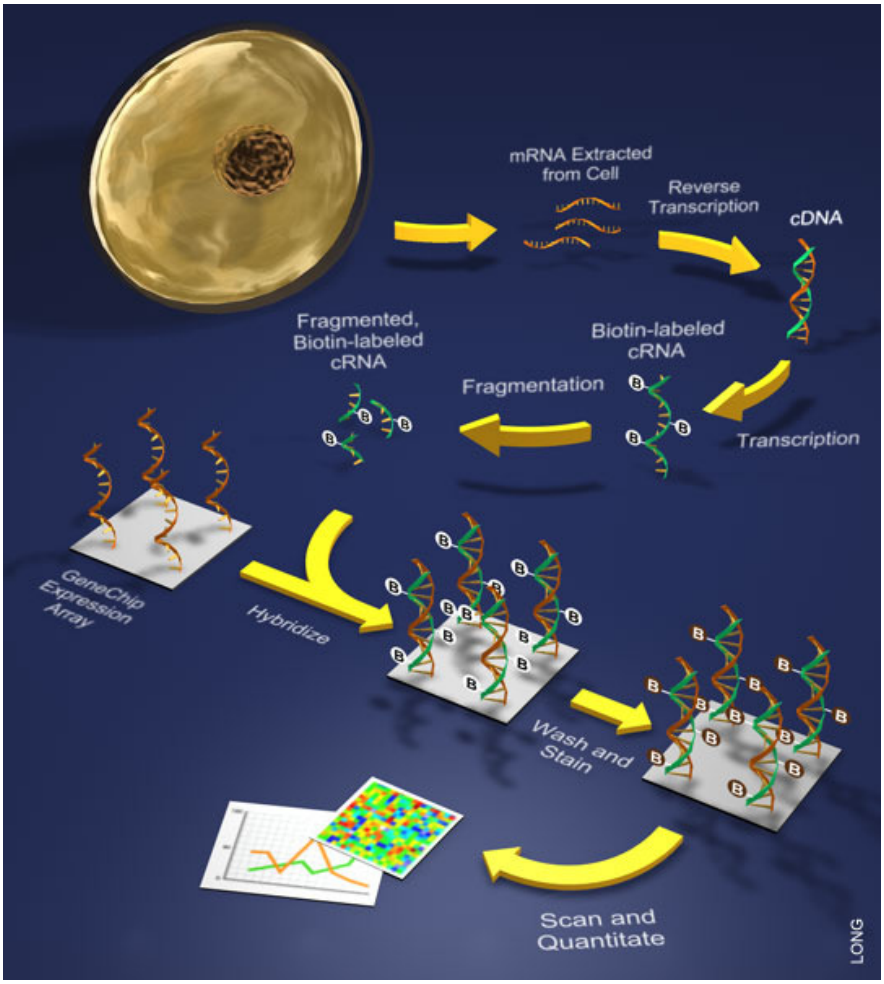

Figure 2.11: Affymetrix GeneChip Scheme - [Image from: http://www.columbia.edu]

2.2.5 Protein detection and localization assays

The analysis of the location of proteins is a powerful tool applicable on a whole organism or at a cellular scale. Investigation of localization is particularly important for study of development in multicellular organisms and as an indicator of protein function in single cells.

Western blotting (or immunoblotting) is a technique used to identify and locate proteins based on their ability to bind to specific antibodies. This kind of analysis can detect a protein of interest from a mixture of a great number of proteins and can provide information about the size of the protein (with comparison to a size marker), and also give information on protein expression (with comparison to a control such as untreated sample or another cell type or tissue).

The first step is "gel electrophoresis". The proteins in the sample are separated according to size on a gel. Usually the gel has several lanes so that several samples can be tested simultaneously. The proteins in the gel are then transferred onto a membrane made of nitrocellulose or PVDF, by pressure or by applying a current. This is the actual blotting process and is necessary in order to expose the proteins to antibody. The membrane is "sticky" and binds proteins non-specifically (i.e. binds all proteins equally well). Protein binding is based upon hydrophobic interactions as well as charged interactions between the membrane and protein.

The membrane is then blocked, in order to prevent non-specific protein interactions between the membrane and the antibody protein. The first antibody (often called the primary antibody) is incubated with the membrane. "Incubation" is typically accomplished by diluting the antibody in a solution containing a modest amount of a salt such as sodium chloride, some protein to prevent non-specific binding of the antibody to surfaces and a small amount of a buffer to keep the solution near neutral pH. The diluted antibody solution and the membrane can be sealed in a plastic bag and gently agitated for an "incubation" of about half an hour. The primary antibody recognizes only the protein of interest, and will not bind any of the other proteins on the membrane. After rinsing the membrane to remove unbound primary antibody a secondary antibody is incubated with the membrane. It binds to the first antibody. This secondary antibody is usually linked to an enzyme that can allow for visual identification of where on the membrane it has bound. The enzyme can be provided with a substrate molecule that

will be converted by the enzyme to a colored reaction product that will be visible on the membrane. Alternately, the reaction product may produce enough fluorescence to expose a sensitive sheet of film when it is placed against the membrane. The whole process is summarized in Figure 2.12.

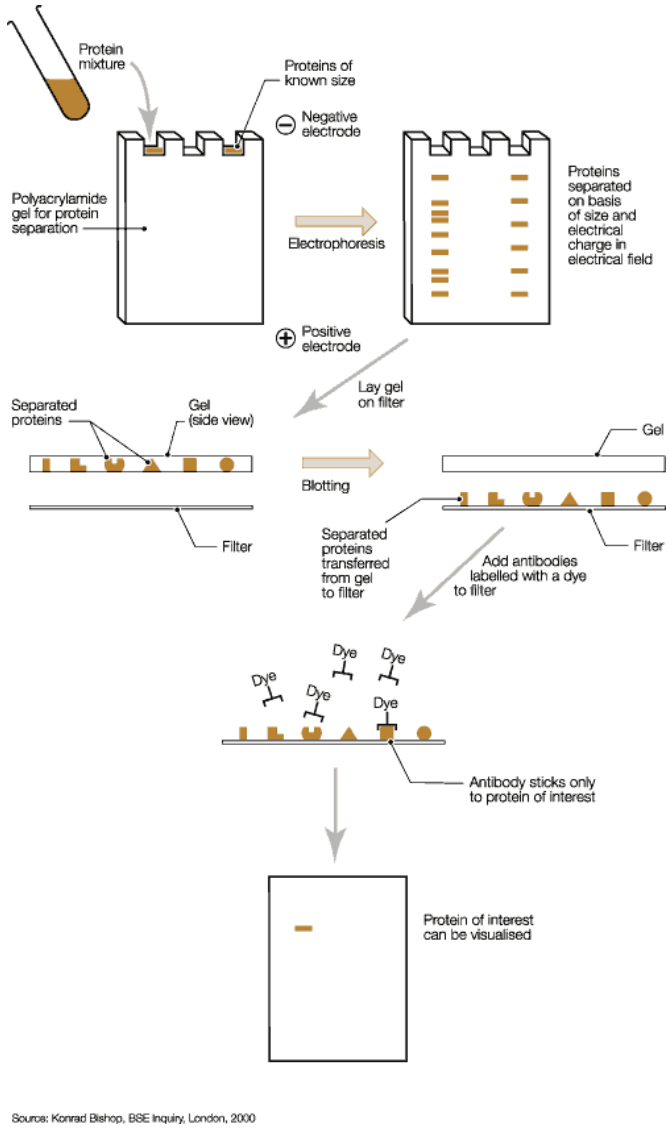

Figure 2.12: Immunoblotting - [Image from: http://www.elec-intro.com]

Same principles lead other immunofluorescence techniques: the specificity of antibodies to their antigen is exploited to target fluorescent dyes to specific biomolecule targets within a cell, and therefore to allows visualization of the distribution of the

target molecule through the sample. Immunofluorescence is a widely used example of immunostaining and is a specific example of immunohistochemistry that makes use of fluorophores to visualize the location of the antibodies.

2.3 Network Theory: basic principles

A network (or equivalently, a graph) is the natural abstraction combined with the corresponding logical-mathematical formalism of a set of objects and their relations. The concept of network is general, "cross disciplinary", and independent from the kind of its composing objects and relations. Sometimes these are strictly dependent from the field in which, in turn, the concept of network is used and they can represents concepts and properties very different among each others: similarity, capacity, interaction, cooperation, transition, etc.

Formally, a network G is defined as the pair (V, E), which is composed by the set of objects $V = \{v_1, \ldots, v_n\}$ called the network "vertex" or "nodes", and by the set $E = \{e_1, \ldots, e_m\}$ called the network "edges" or "links", representing the relations occurring among the network nodes.

The single edge $e \in E$ is a pair of nodes $(x, y) \in V^2$ and it represents the relation occurring between node x and node y. In this case, nodes x and y are said to be joined by the edge e. If in the pair (x, y) the sorting order is relevant then the network G will be considered as "oriented network" (or directed graph, or di-graph), being the edge (x, y) different from the edge (y, x). Otherwise the network will be said "not oriented" (or simple graph).

An edge $e = (x, y)$ of a directed network is composed by a "source" node $s(e) = x$ (or head of e) and a "destination" node $d(e) = y$ (or "tail" of e). On the other hand, in a simple graph, the two nodes joined by the edge e can be considered as source or destination interchangeably.

Nodes of a network are visually denoted as circles or points while edges of a directed network (i.e. directed edges) are denoted as arrows going from the source to the destination node. Segments are used to denote simple edges.

In Figure 2.14 two example of network are shown: a directed (a) and an undirected network (b), respectively.

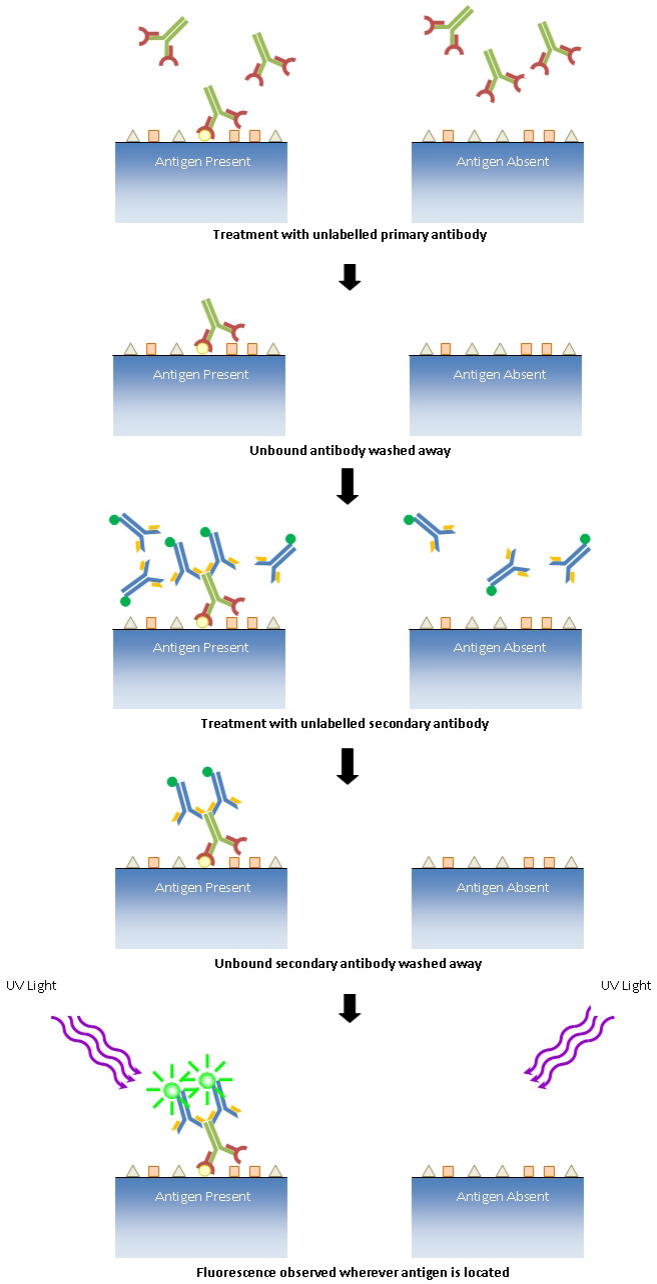

Figure 2.13: Indirect immunofluorescence - [Image from: http://www.di.uq.edu.au/indirectif]

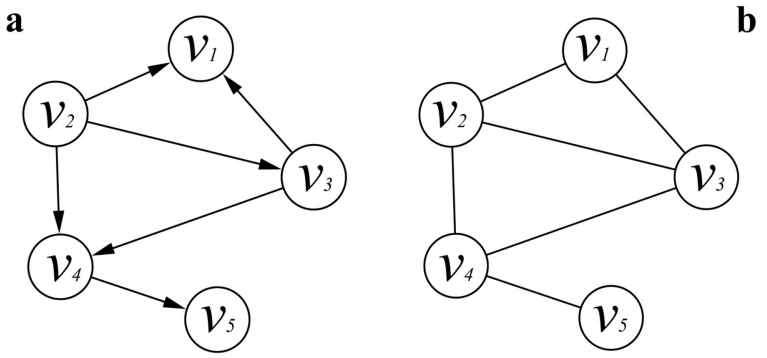

Figure 2.14: Examples of networks - a) directed network, b) undirected network

There is a huge amount of problem in operations research that are modeled through "weighted networks". Generally in these models the weight of an edge represent a "loss" or a "gain" and the solution to most of these modeled problems implies the identification of a sub-network (composed by a sub-set of nodes and edges) for which a function defined on the weights of its edges assumes an optimal value given a certain numbers of constraints.

For example the identification of a path of edges with minimal total weight between two nodes is a classical problem of this kind. Another very famous problem is the identification of the "minimum spanning tree" (i.e. the sub-network obtained by linking together each pair of nodes with path of minimal total weight).

Many other definitions, statistics and measures on networks define their "topological" (i.e. structural) features, their "modularity", the density and localization of their edges. As an example, in a network, a "clique" is a subset of its vertex that are mutually joined by an edge and a "maximal clique" is a clique for which is not possible to add other nodes without violating the condition of mutual conjunction.

On the other hand the "local clustering coefficient" quantifies the edge density on a defined subset of nodes. Specifically, it measures the tendency of the nodes neighboring (i.e. connected to) a given node to form a clique.

Another important property of network is "modularity": the tendency to contain "communities". A network community is a group of nodes that are "densely" interconnected among each other and with fewer connections to nodes outlying group (where "densely" and "fewer" are statistically defined).

Community structures are quite common in real networks. Social networks often include community groups (the origin of the term, in fact) based on common location, interests, occupation, etc. Metabolic networks have communities based on functional groupings. In some network other concept of modularity can be observed for example when nodes of higher degree (hubs) are better connected among themselves than are nodes with smaller degree. A network with this property is said to presents modules termed "rich-clubs". The presence of the rich-club phenomenon (100) may be an indicator of several interesting high-level network properties, such as tolerance to hub failures. Being able to identify these sub-structures within a network can provide insight into how network function and topology affect each other. Many algorithms have been proposed to find communities in network (43, 48, 109).
Other important features of networks regard the statistical distribution the degree and basically are used to discriminate random, hierarchical and complex networks.

2.3.1 Networks as alternative to euclidean embeddings

When the object of study is the analysis of "similarities" among a set of objects then "similarity-networks" are a very effective alternative to "euclidean embeddings". An euclidean embedding is a placement (i.e. a disposition) of a set of objects in a (usually) low-dimensional and visualizable space and it is an important tool in unsupervised learning and in preprocessing data for supervised learning algorithms. Euclidean embeddings are especially valuable for exploratory data analysis and visualization because they provide easily interpretable representations of relationships among objects. Many dimensionality reduction techniques such as Principal Components Analysis (PCA) (71) or MultiDimensional Scaling (MDS) (23) make possible the euclidean embedding of a set of objects, starting from their coordinates in a high dimensional original space or their pair-wise similarity/distance scores. With respect to euclidean embeddings similarity networks are more easily computable and visualizable and can be exploited for a number of tasks in data exploration analysis through topological tools from network theory.
In these networks the edges have weights corresponding to a similarity scores (or distances, inversely proportional to similarity) and, usually, only "significant" edges (whit weights above a statistically significant threshold) are included.
In some cases (i.e. when the network is too dense, it contains to much connections) is

very useful to "compress" it by identifying its modules and communities. Anyway it must be stressed that in this case modularity is deemed as a property to be provided to the network rather than being one of its intrinsic features and it is strictly linked to cluster analysis. Cluster Analysis is one of the most famous tool for data exploration whose goal is grouping objects of similar kind into respective categories without knowing anything but the data itself. One could look to cluster analysis as to a classification tool in which no sets containing already classified samples are available as well as no prior knowledge about the composition that the output clusters should have. Hence nothing could be learnt in any kind of preliminary matter training.

Formally, the problem tackled with cluster analysis is a particular case of a more general class of problems: the partitioning problems. In this class of problems, given a set of objects N and a set of K functions $f = (f_1, f_2, \ldots, f_k)$ from the set N to $\mathbb{R}$ (real numbers), the aim is to find a partition $A = (A_1, A_2, \ldots, A_k)$ of the set N that minimizes or maximizes an objective function $g(f_1(A_1), ..., f_k(A_k))$. In the case of cluster analysis the function defined on the subsets A_i of N is the same for every $i = 1, \ldots, k$ and usually it is the sum of the pairwise similarity between the elements of A_i. The function g is usually a sum and it should be maximized. In other words, in this class of problems, the production of a partition in which data points belonging to the same subset (cluster) are as similar as possible is aimed. So, the first observation we can make is that the ability to quantify a similarity (or distance, its inverse) between two objects is fundamental to clustering algorithms.

The greatest part of clustering algorithms are based on the concept of distance so we have to choose a similarity measure that allows the set of objects to be embedded in a metric space. Usually the guidelines that we can follow are: we use a lot of detailed knowledge to make a metric space embedding then we use classical distance metric (i.e. euclidean, correlation, cosine, etc.) in this space to make clustering or we use a user defined distance metric in a clustering algorithms directly.

Almost all the clustering approaches can be divided in two major class: hierarchical clustering algorithms and partitional clustering algorithms. The methods of the first class build (agglomerative algorithms), or breaks up (divisive algorithms), a hierarchy of clusters. The traditional representation of this hierarchy is a tree (called a dendrogram), with individual elements as leaves and a single cluster containing every element as root. Agglomerative algorithms begin from the leaves of the tree, whereas divisive

algorithms begin from the root. The methods of the second class attempt to directly decompose the data set into a set of disjoint clusters. The most famous algorithm of this second class is the K-means algorithm. In partitional clustering algorithms the value of the input parameters (like i.e. the value of K in K-means or the map dimension in Self Organizing Map approach) plays a key role and in many cases it determines the final number of clusters. In hierarchical clustering algorithms the same role is played by the choice of the dendrogram cutting threshold. A way to justify the choice of these values consists in making same preliminary matter statistical analysis on the set that we want to cluster instead of blindly make clustering on it. Alternatively, an heuristic can be used. Moreover, this first analysis can check the effective clusterizability of a set, in other words, it check the presence of well localized and well separable homogeneous (by the similarity point of view) groups of objects in the set. The tools allowing this kind of analysis are based on the clustering stability concept(11, 134). In these methods many clusterings are computed by previously introducing perturbations into the original set, and the candidate clustering is considered reliable if its composition is approximately reflected across all the computed clusterings. Informally, the stability of a given clustering is a measure that quantifies the change the clustering is affected by, after a perturbation on the data set.
Usually the objective functions that clustering algorithms tries to minimize has multiple local minimums. It means that multiple and, in some cases very different, solutions grant very close optimal values for the objective function.
For all these reasons, with cluster analysis the ability to form meaningful groups of objects (one of the most fundamental modes of intelligence) can be approximatively simulated by automatic procedures. However, enabling computers to accomplish this task is a difficult and ill-posed problem.

2.4 Computational Drug Discovery

Drug industries are part of a very segmented market in which the largest company (Pfizer) only has an 11% market share. In this highly competitive field risks are extremely significant for investors and the term of profits is very long. Usually a novel drug takes from 10 to 20 years to be developed, and most drugs fail to get to the market.

Additionally this industry area is highly regulated by governmental regulatory agencies like the U.S. Food and Drug Administration (FDA) and the European Medicine Agency (EMEA).
The production process in drug has traditionally been divided in four main phases that can be roughly summarized as:

- **Discovery**: after a disease or a pathological condition of interest has been identified, the responsible proteins are isolated and characterized (for example, by identifying causal genetic changes); then a "pharmacophore" (i.e. a set of structural features in a molecule that is recognized at a receptor site and is responsible for that molecule's biological activity) is identified on the basis of these proteins; usually this is done by searching for compounds that interacts with the target protein by using huge libraries of compounds;

- **Development**: the objective of this phase is to synthesize lead compounds, new analogs with improved potency, reduced off-target activities, and physiochemical/metabolic properties suggestive of reasonable in vivo pharmacokinetics; this optimization is accomplished through chemical modification of the pharmacophore (also called "hit structure"), with modifications chosen by employing structure-activity analysis;

- **Clinical trials**: depending on the type of product and the stage of its development, investigators enroll healthy volunteers and/or patients into small pilot studies initially, followed by larger scale studies in patients that often compare the new product with the currently prescribed treatment; as positive safety and efficacy data are gathered, the number of patients is typically increased; clinical trials can vary in size from a single center in one country to multicenter trials in multiple countries;

- **Marketing**: the process of advertising or otherwise promoting the sale of the novel approved drug.

Before computational drug discovery was introduced, drugs were discovered by chance in a trial-and-error manner. Not even the introduction of new technologies, such as High-Throughput Screening (HTS) that can experimentally test hundreds of thousands of compounds a day for activity against the target protein, have resulted in a

more successful identification of promising drug candidates or reduced the process costs. Additionally, the use of HTS is very expensive.

Computational methods can be used to predict or simulate how a particular compound interacts with a given protein target. They can be used to assist in building hypotheses about desirable chemical properties when designing the drug and they can be used to refine and modify drug candidates. Computational Methods can also be used to automate repetitive tasks such as searching large compound databases. Virtual Screening is a general term for computational methods that use computers to screen a database of virtual drug candidates to identify promising candidates (leads). This can be seen as an alternative to perform laboratory experiments or to perform HTS. The major advantages compared to laboratory experiments are: low costs, it is possible to conduct investigation on compounds that have not yet been synthesized, possible preliminary step to select set of compounds for real HTS, investigation are conducted in a huge chemical search space in which the number of possible virtual molecules is higher than the number of real ones.

When the structure of the target is known the most commonly used virtual screening method is molecular docking. Molecular docking programs try to predict how a drug candidate binds to a protein target without performing a laboratory experiment.

Other important applications of computational drug discovery are led by the following requirements and offer proper solutions: chemical structure and other biological data need to be stored for millions of datapoints so computational representation of 2D chemical structure are widely used; thousands of active compounds need to be organized into meaningful groups hence cluster analysis or machine learning methods are effective in grouping similar structures together and relate to activity; as much information as possible must be learnt from the data (data mining) so statistical methods and other unsupervised computational techniques are applied to the structures and related information; finally, microarray technology allows to look for changes in protein expression for different people with a variety of conditions, and to see if the presence of drugs changes that expression, thus making possible the design of drugs to target different phenotypes.

2.5 Network analysis improves understanding of drug use and effects

In several recent works it has been shown that biological network analyses can contribute in understanding the effects of clinically used pharmaceuticals (12). Various approaches can identify previously unknown targets of the drug, pathways affected by the drug and pharmacogenomic factors affecting the usage of the drug. These can in turn be used to explain off-target effects, adverse events or suggest additional indications or contraindications for the usage of a drug. While many drugs have known therapeutic targets, many other drugs that are currently used work through unknown mechanisms. Furthermore, even drugs with a known target often have off-target effects. These are effects, often undesirable, of a drug which can not be explained through its interaction with its primary targets.

Network studies of drugs have allowed identification of some of these secondary targets of drugs (18). Another way drug targets can be linked together into a network involves a chemoinformatic approach in a method for scoring the similarity between the sets of ligands for different receptor (76). Then this score can be used to construct a network of receptors connected together if they bind structurally similar ligands . This analysis showed that many biologically related drug targets clustered together by ligand similarity even though the targets themselves have minimal sequence similarity.

Drug action is not only related to the targets of the drug, but can also be affected by variations in metabolic enzymes, transporter proteins and downstream effects of drug action. The field of pharmacogenomics identifies genetic variations that can change drug effects. Network analyses can contribute to identification of such pharmacogenes, genes which modulate the response to a drug (55).

In addition to identifying unknown targets of drugs and pharmacogenes, network-based approaches can suggest potential alternative uses of drugs (91).

By using a text-mining based approach information about drugs, treatments and diseases can be integrated into a Disease-Drug Correlation Ontology (117). By querying the complex network structure surrounding a disease drugs that might modify its course can be predicted.

These kind of study (referred as Systems pharmacology approaches) are changing the traditional drug discovery pipe-line and the way about drug actions are tougth. They

allow a deep understanding of the mechanism of action of the drug enabling proposal predictions for “drug repositioning”.
Also known as Drug repurposing, Drug re-profiling and Therapeutic Switching, drug repositioning has been growing in importance in the last few years. Using drug repositioning, pharmaceutical companies have achieved a number successes, for example Pfizer’s Viagra in erectile dysfunction and Celgene’s thalidomide in severe erythema nodosum leprosum. Smaller companies are also performing drug repositioning on a systematic basis. These companies use a combination of approaches including in silico biology and in vivo/in vitro experimentation to assess a compound and develop and confirm hypotheses concerning its usage for new indications.
The most significant advantage of drug repositioning over traditional drug development is that since the repositioned drug has usually already passed a significant number of toxicity and other tests, its safety is known and the risk of failure for reasons of adverse toxicology are reduced. Therefore, the walk of a repositioned drug to the marked is easier, cheaper and faster.

3

Gene Expression Based Methods and Systems Biology

3.1 Introduction

Systems biology approaches are naturally suited to capture the complexity of drug activity in cells (12, 61, 96) and the discovery of drug direct target, or more generally of the drug MoA is strictly linked to the problem of the inference of regulatory mechanisms in the cell.
In this chapter we introduce some basic concepts in this field and we report about some results that we obtained in two side-projects we were involved in.
Finally, we describe the gene-signature based methods for the analysis of phenotypes and the study of drug effects and we introduce the starting point of our main project: the Connectivity Map (cMap).

3.1.1 Inference of Gene Regulatory Network

In the context of systems and computational biology a "regulatory network" is a complex set of interactions occurring among different entities within the cell (i.e. DNA, mRNA, proteins, etc.), which tightly regulates its behavior. Depending on which cellular entities are taken into account, different kind of regulatory networks can be distinguished as a hierarchy of layers, as depicted in Figure 3.1.
A "transcriptional network" can be viewed as a graph in which the nodes are the genes and there is an edge between two nodes if the transcriptional activities of the corre-

sponding genes are correlated, in some way. As an example, the gene denoted by 1 in Figure 3.1 could encode for a transcriptional factor that modulates the expression of the genes corresponding to the gene denoted by 3.

In protein networks, the nodes represent proteins and the edges identify physical interactions among them. An important subset of these networks are the phosphoproteomic signaling networks that can be viewed as direct graphs in which nodes are kinase proteins and their substrates.

At the third level of the hierarchy of Figure 3.1 are the "metabolic networks". Metabolites are usually small molecules produced, absorbed and/or transformed by the cells through different chemical reactions. Since these reactions are typically linear and unidirectional, metabolic networks are usually represented as trees (i.e. directed acyclic graphs) in which nodes are metabolites and there is an edge between two nodes if an enzyme catalyzes a chemical reaction transforming the metabolite represented by the first node into the metabolite corresponding to the second one.

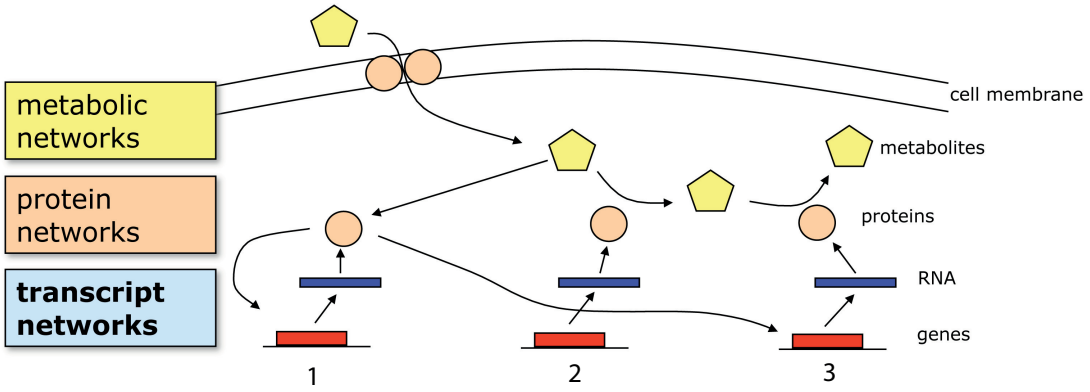

Figure 3.1: Biological network layers - Depending on the involved cellular entities, regulatory networks are organized in different layers of specificity

The development of computational methods to reconstruct these regulatory networks is crucial for the following tasks:

- Identification of functional modules (i.e. subset of genes densely interconnected among each other thus consistently involved in the same biological processes);
- Identification of physical interactions (i.e. new transcription factors or transcription factor targets) elucidating the mechanism of a disease and/or suggesting new

"druggable" targets;

- Prediction of the system response to an external perturbation as well as the identification of the specific target of the perturbation (i.e. identification of drug target or, more generally, of drug MoA).

The problem of "reverse engineering" gene regulatory network from gene expression data can be stated as follows: given a set of gene-expression data obtained from multiple microarray experiments, we would like to infer the network of genes that produced such data, that is, the gene-gene interactions describing the underlying biological process (9). A number of approaches have been proposed, based on correlation analysis (120), on mutual information relevance networks (97), on Bayesian networks (153), on clustering algorithms (37) and on deterministic Ordinary Differential Equations (ODEs) models (34, 45).
All these methods follow the computational pipeline depicted in Figure 3.2: differential gene expression measurements are taken from a system (i.e. a cell culture) grown in different conditions and/or exposed to different perturbations (i.e. drug treatment, stress, etc.) then they are given in input to a learning algorithm, which predicts a model of transcription regulation that could underlie the system under investigation.

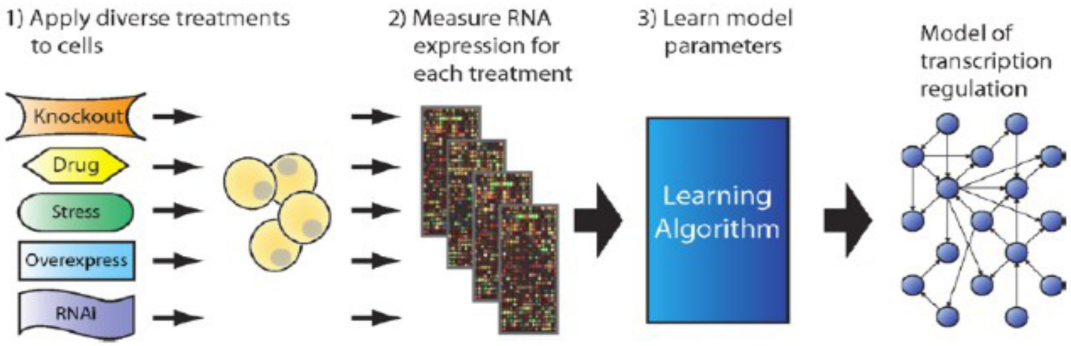

Figure 3.2: Computational pipeline - The general computational pipeline of the algorithms for gene-regulatory network inference.

3.1.2 The Network Inference by multiple Regression (NIR) algorithms

As described in (45), in the NIR approach, the gene network dynamics describing the time evolution of the mRNA concentration transcribed by each gene are modeled by a

set of ordinary differential equations:

$$\frac{dx}{dt} = f(\mathbf{x}, \mathbf{u}) \tag{3.1}$$

where $\mathbf{x}$ represents the mRNA concentrations of the genes in the network and $\mathbf{u}$ is a set of transcriptional perturbations. Basically, NIR is based on the strong assumption that the rate of change in expression for each gene can be modeled as a linear combination of the expression level of its regulator genes plus a proper component of the external perturbation to the system (as summarized in Figure 3.3).

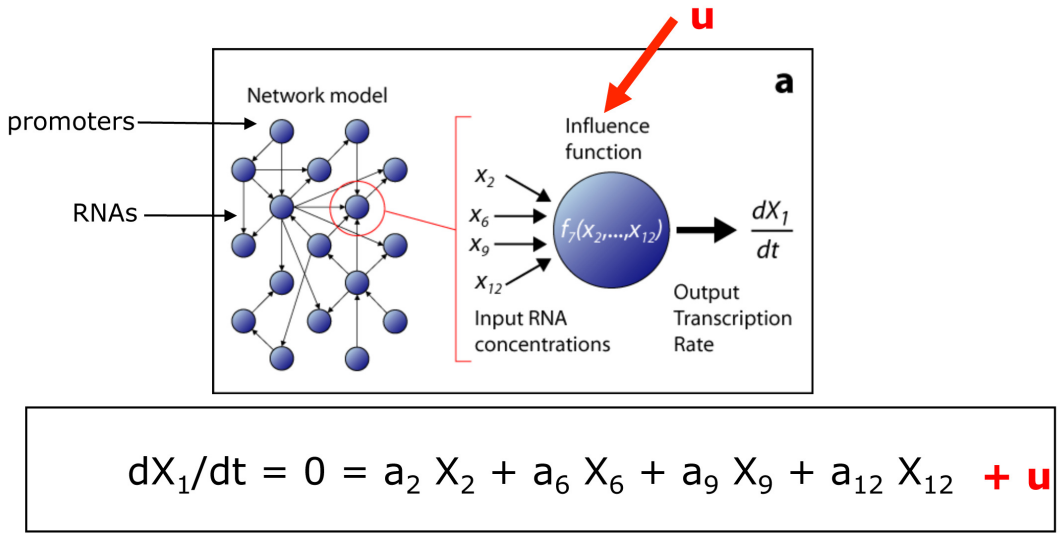

Figure 3.3: The NIR assumption - The rate of change in expression of the generic gene X_1 can be modeled as a linear combination of a subset of other genes plus an external perturbation.

This assumption of linearity is supported by the following considerations. Assuming that the system under investigation is at equilibrium near a stable steady-state point (i.e. the genes are close to an equilibrium point in which their level of expression are sufficiently kept constant), then we can apply a small perturbation to each of its genes. A perturbation is small if it does not drive the system out of the basin of attraction of its stable steady-state point and if the stable manifold in the neighborhood of the steady-state point is approximately linear. With these assumptions the set of nonlinear rate equations can be linearized near their stable steady-state point.

Thus for each gene i, in a network of N genes, we can write the equations

$$\frac{dx_{il}}{dt} = \sum_{j=1}^{m} a_{ij}x_{jl} + u_{il} = \mathbf{a}_i\mathbf{x}_l + u_{il}, \quad i = 1, \ldots, N, \quad l = 1, \ldots, M \tag{3.2}$$

where x_{il} is the mRNA concentration of gene i following the perturbation in experiment l; a_{ij} represents the influence of gene j on gene i; and u_{il} is an external perturbation to the expression of gene i in experiment l.

Identifying the gene interactions network means to derive the matrix $\mathbf{A}$ of the coefficient a_{ij} for each gene i in the model described above. This can be accomplished if we measure the mRNA concentration of all the N genes at steady state (i.e., $dx_i/dt = 0$) in M experiments and then we solve the system of equations:

$$\mathbf{AX} = -\mathbf{U} \tag{3.3}$$

where $\mathbf{X}$ is an $N \times M$ matrix whose columns are the $\mathbf{x}_l$ vector and $\mathbf{U}$ is an $N \times M$ matrix whose columns are the $\mathbf{u}_l$ vectors. This system can be solved only if M is at least equal to N, however the recovered weights $\mathbf{A}$ will be extremely sensitive to noise both in the data and in the perturbations and thus unreliable unless we overdetermine the system (increasing the number of experiments or assuming the number of regulators of each gene, k, is much smaller than M).

In order to estimate the coefficients of the gene interaction network (i.e. the matrix $\mathbf{A}$), NIR essentially solves a linear regression problem for each equation in equation 3.2 assuming an upper bound of k regressors for each predicted gene. That is, it assumes that each gene can be regulated at most by k other genes; the value of k can be computed empirically as the best compromise between computational complexity and completeness of the results. Otherwise it can be computed adaptively by using strategies such as the Akaike's final prediction error (2).

The set of variables comprising the regressor set is chosen according to the Residual Sum of Square (RSS) error minimization criterion.

As explained in the next section NIR is one of the most powerful algorithms for the reverse engineering of gene regulatory network but it requires to *a priori* know which is the gene that has been perturbed in each of the experiment.

3.1.3 The DREAM initiative

We competed, for the challenge number 4 of the Dialogue for Reverse Engineering Assessment and Methods (DREAM) 2 competition (138) by applying the NIR algorithm. The DREAM initiative (together with the annual conference), as claimed by the organizers, aims at "catalyzing the interaction between experiment and theory in the area of cellular network inference and quantitative model building in systems biology" in order to rigorously definine standards and methods in the assessment of reverse-engineering approaches and to create a repository of benchmark data and algorithms.
Thanks to this initiative, experimental data (both real and "in-silico" simulated data) are made publicly available every year and people working in this field can try to infer the true-biological/simulated model (for example, the gene regulatory network) through which those data have been generated.
We won one of the challenges of the DREAM initiative by applying NIR and by obtaining a very high final score.
Moreover, we designed NIR with perturbation Estimates (NIRest) , a tool that builds upon the original NIR and extends its use to cases in which the generating perturbations are not known (82).

3.1.4 The IRMA project: In-vivo Reverse-engineering and Modelling Assessment

The goal of the IRMA project was to provide the systems biology community with an in vivo benchmark, which can be used as "ground truth" to test and compare modeling approaches and reverse-engineering inference strategies. At present, the usefulness and predictive ability of computation approaches in the field of systems and synthetic biology cannot be assessed and compared rigorously. To this aim in (19), we built, in the yeast *Saccharomyces Cerevisiae*, a synthetic network of five genes regulating each other for In-vivo Reverse-engineering and Modelling Assessment (IRMA).
The network was designed to be negligibly affected by endogenous genes, and to respond to galactose, which triggers transcription of its genes. Our network (Figure 3.4), apparently simple, is in fact very articulated in its interconnections generated by the combination of transcriptional activators and repressors.

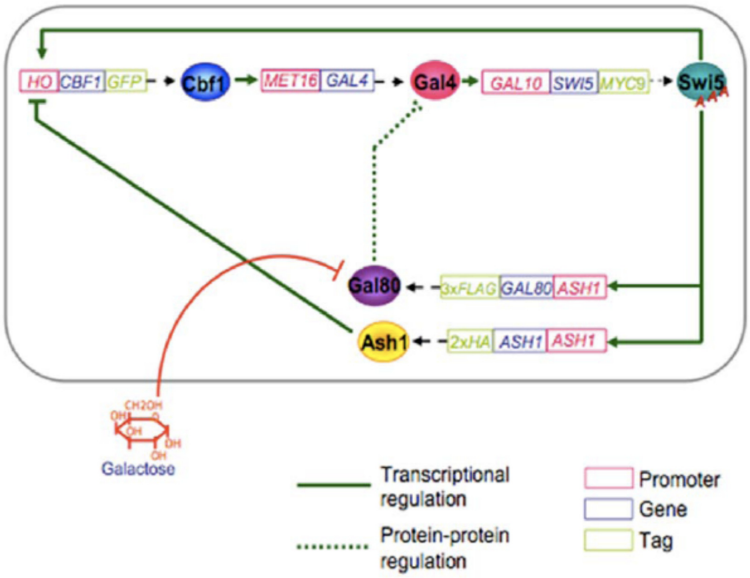

Figure 3.4: IRMA - Schematic diagram of IRMA. New transcriptional units (rectangles) were built by assembling promoters (red) with non-self coding sequences (blue). Genes were tagged at the 3' end with the specified sequences (green). Each cassette encodes for a protein (represented as a circle) regulating the transcription of another gene in the network (solid green lines). The resulting network, is fully active when cells are grown in presence of galactose, while it is inhibited by the Gal80-Gal4 interaction in presence of glucose.

In order to infer the topology of IRMA, we analyzed the transcriptional response of network genes after the following perturbation strategy: we performed multiple perturbation experiments (to each of the network genes, in turn) and collected mRNA measurements at the steady state (as explained before, the situation in which the gene expression levels are sufficiently constant). We then applied NIR to these data. In each of the perturbation experiments we performed only one gene was perturbed (see Figure 3.5). Perturbations were realized by over-expressing each of the five network genes under the control of a promoter that is strongly constitutively expressed in yeast. In this case we considered a fixed number of regressors for each of the 5 genes ($k = 2$) (i.e. we assume that each gene can be regulated by a maximum of 2 other genes). The regressor set was chosen according to the RSS minimization criterion. Since we have only 5 genes in the network we exhaustively searched the best regressors in the space of all the possible couples of genes.

As shown in Figure 3.6, 60% of the network predicted by NIR was composed by connection that are actually present in IRMA and these performances are clearly better than those expected to be obtained by chance (40%).

3.2 Analysis of Phenotypic Changes

A phenotype is any observable feature, or trait, of an organism in normal, diseased or perturbed condition.

"Phenotypic drug discovery", primarily abandoned in the 1980's in favor of targeted approaches to drug development, is nowadays demonstrating its value when used in conjunction with new technologies (89).

Differential gene expression data following drug treatment can be considered as an important aspect of a phenotypic change in response to the drug. The goal in this case is to identify small molecules that modulate the expression of a target gene in a specific manner, thereby either increasing or decreasing the concentration of the corresponding protein product. Transcriptional modulation not only provides a potential means to replace recombinant proteins as drugs, but also provides a novel approach to manipulate key gene targets in many therapeutic areas.

On the other hand drug perturbations of human cells lead to complex responses upon target binding. Therefore to focus narrowly on a single target gene does not take into

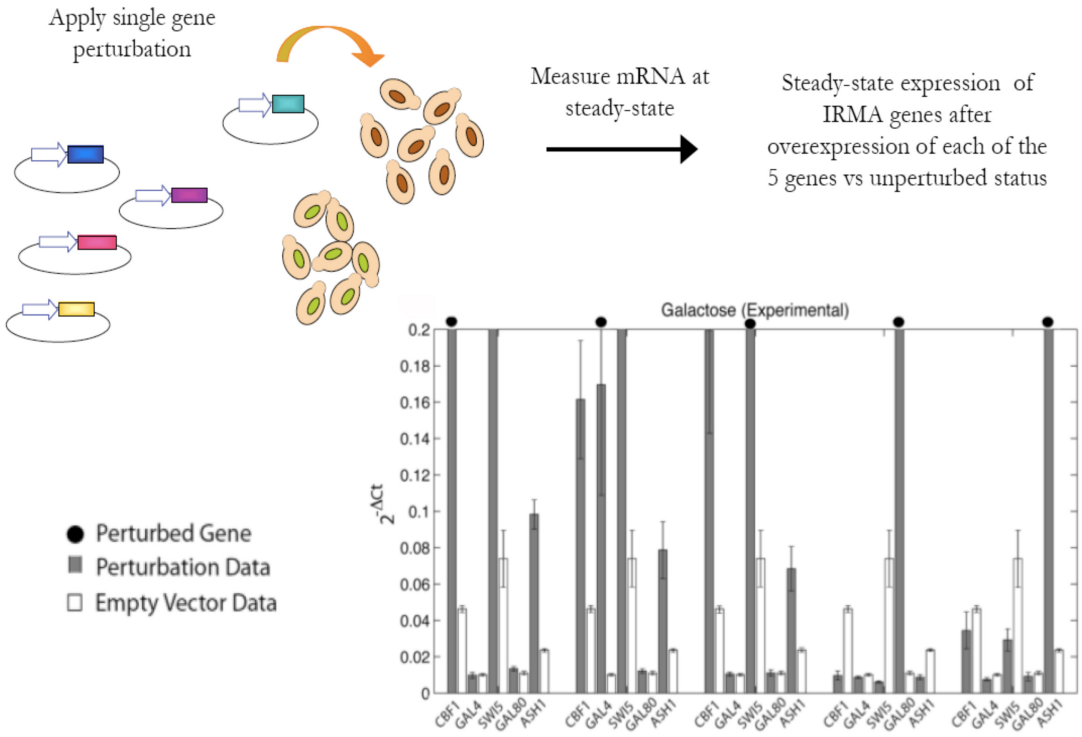

Figure 3.5: Perturbing IRMA - In vivo expression levels (by real-time PCR) of IRMA genes after over-expression of each gene (perturbed gene, indicated by the black dots on the bars) from the constitutive GPD promoter (gray bars) and after transformation of the empty vector (white bars). IRMA cells were transformed with each of the constructs containing one of the five genes or with the empty vector

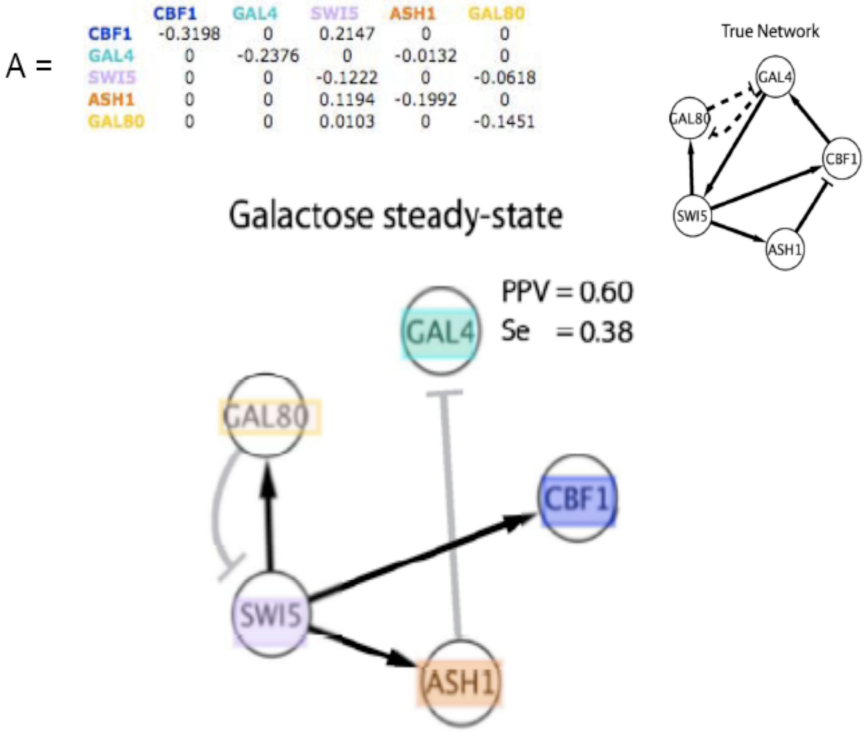

Figure 3.6: Inferring IRMA with NIR - Performances of the NIR algorithm when applied to the inference of IRMA

account the complexity of physiological functions. For this reason, strategies based on the analysis of drug-induced changes in gene expression at a genome-wide scale have the potential to elucidate the cellular response to specific drugs hence to provide significant efforts to the drug discovery process. Basing on these concepts, prediction of drug MoA has been attempted by using gene expression profiles following drug treatment in a valuable number of recent approaches (45, 62, 65, 79, 80, 152)

3.2.1 The Connectivity Map dataset

The Connectivity Map (cMap) data-set is a large public database of genome-wide gene expression data from five different human cancer cell lines, treated with $\approx 1,300$ bioactive small molecules at different concentrations (79, 80).
Data are organized in experiments (batches) composed by two or more microarray hybridizations of the treated cell line and one or more hybridizations of the untreated cell line as negative control, for a total number or 6,100 "instances". An instance is the basic unit of data and metadata in cMap and consists of a treatment and control pair and the list of probe sets ordered by their extent of differential expression between this treatment and control pair. Every instance has a number of attributes including a unique identifier, the batch in which it was produced, the cMap name of the treatment, the source of that treatment, the concentration of that treatment, the cMap cell line used, and the scan (i.e. the chip) numbers for the treatment and its control(s).
The number of treatments and controls per batch can vary as the number of total treatments across batches per single drug.
The change in expression of a cell line after a treatment is computed by considering the differential expression values of a treated hybridization with respect to those of the untreated one (or the set of untreated ones). Hence, each treatment with a drug in a batch yields a genome-wide differential Gene Expression Profile (GEP) (see Figure 3.7).
The aim of the cMap project is to generate a detailed map that links gene patterns associated with disease to corresponding patterns produced by drug candidates and a variety of genetic manipulations.

Together with the data, a pattern-matching tool (detailed in the following sections and summarized in Figure 3.8) allowing users to find connections between a well defined

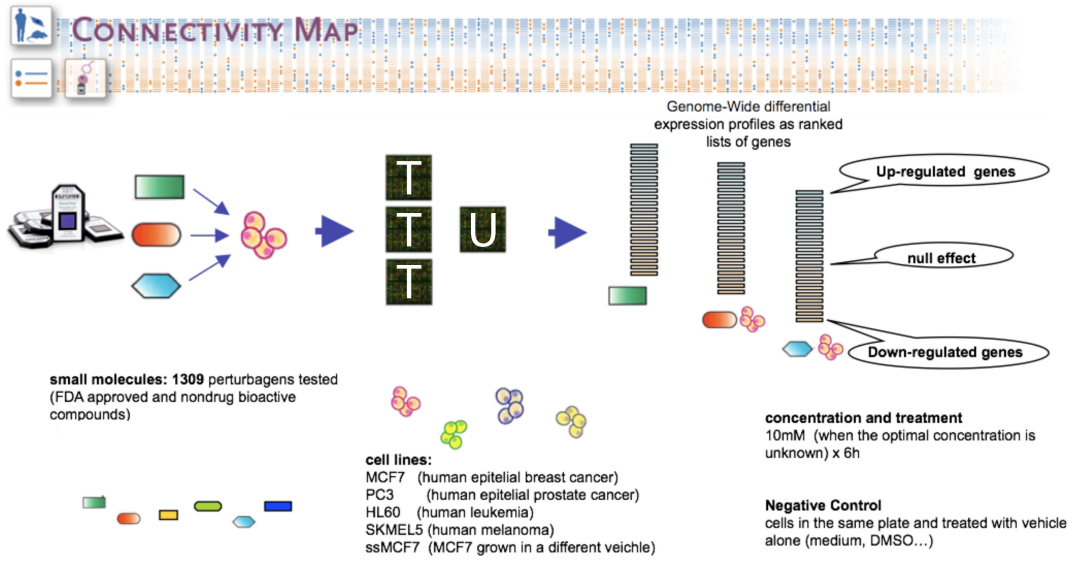

Figure 3.7: The Connectivity Map - Dataset overview

set of genes (a signature) and the GEPs of the cMap has been designed and its implementation is publicly available on the cMap website (`http://www.broadinstitute.org/cmap/`).

3.3 Gene Signature Based Methods

Given a gene expression profile for a set of cells in a specific condition and for a set of control cells, an important problem is to identify "patterns" of differential gene expression that can be used as a "summary" of that biological condition. These identified patterns are useful in a large number of problems (for example, to classify phenotypes (17)).

A "gene signature" is, in this case, a set of genes whose differential expression pattern is specific for the condition under consideration and gene signature based methods for drug study and phenotype characterization are based on the following ideas. Given a set of microarray gene expression profiles from two different conditions (i.e. healthy vs. diseased tissues or treated vs. untreated cells) the difference in gene expression can be used to compose a gene signature. Then the obtained signature can be "connected" to other conditions (drug treatment, diseases, etc.) consistently expressing the composing

genes. This form the basis and the leading concept of the tool described in the following section, upon which the cMap query system is built.

3.3.1 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) (140) is a computational method that determines whether an *a priori* defined set of genes (i.e. a gene signature) shows statistically significant, concordant differences between two biological states (i.e. phenotypes). The basic idea behind GSEA is that by using predetermined sets of genes, perhaps based on a function, gene expression profiles could be better interpreted.
GSEA considers experiments with genome-wide expression profiles from samples belonging to two conditions (for example, untreated vs. treated cells). Genes are ranked according to their differential expression between these two conditions yielding the ranked list of genes L. Given an *a priori* defined set of genes S (i.e., genes encoding products in a metabolic pathway, located in the same cytogenetic band, or sharing the same Gene Ontology (GO) category, or selectively expressed in a disease, or consistently up-regulated in response to a drug treatment), the goal of GSEA is to determine whether the members of S are randomly distributed throughout L or primarily found at the top or bottom. The key assumption is that if the genes in S are related to the phenotypic distinction then they will tend to show the latter distribution. These propensity is quantified by a measure called the Enrichment Score (ES).
ES is calculated by walking down the list L, increasing a running-sum statistic when a gene in S is encountered and decreasing it when a gene not in S is encountered. ES is the maximum deviation from zero encountered in this walk; it corresponds to a weighted Kolmogorov-Smirnov-like statistic (59).

3.3.2 The Connectivity Map query system

One of the cMap's unique features is that it allows researchers to screen a huge set of compounds against an *a priori* defined gene signature. The query system makes use of an extension of the GSEA.
The computation starts with a "query signature" and measures the extent of its similarity to each of the reference GEPs in the cMap data set. This query signature is any list of genes known to be involved in a biological state of interest (i.e. genes correlated with a subtype of disease or regulated by a biological process of interest etc.).

The signature is composed by two subset of genes: those that are typically up-regulated in the described biological state and those that are down-regulated.
The reference gene-expression profiles in the cMap data set are represented in a non-parametric fashion. Each profile is compared to its corresponding intra-batch untreated hybridization (the negative control). Then the genes on the array are rank-ordered according to their differential expression relative to the control; each treatment instance thus gives rise to a rank-ordered list of $\approx 22,000$ genes.
The query signature is finally compared to each rank-ordered list to determine whether up-regulated query genes tend to appear at the top of the list and down-regulated query genes at the bottom or vice versa, yielding a Connectivity Score (CS) ranging from -1 to 1. In the first case the CS will be near 1 while in the second case it will be near -1. All instances in the database are then ranked according to their CS; those at the top are most strongly correlated to the query signature, and those at the bottom are most strongly anti-correlated. An overview of the method is presented in Figure 3.8.

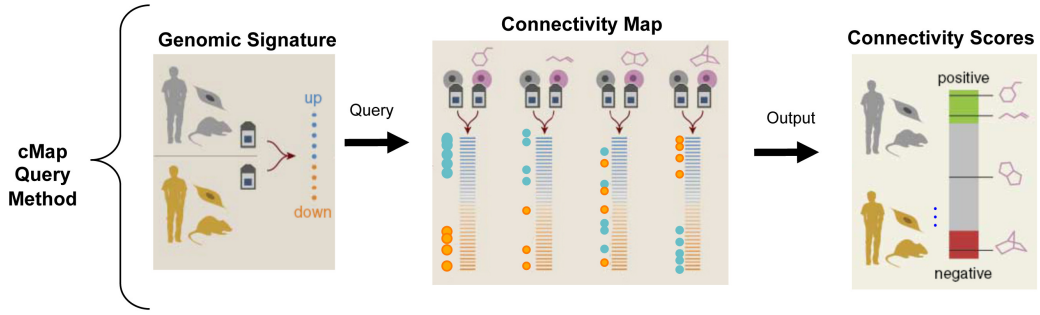

Figure 3.8: cMap query method - Overview of the cMap query system to link the GEPs to a well defined genomic signature.

The cMap query system has been successfully used in a valuable number of recent works and for different purposes (56, 58, 124).

3.4 A first pilot study

In a first pilot study we sought to probe a first release of the cMap dataset (164 small molecules for a total of 453 instances) in order to selectively compare all the contained drugs. We did this by using traditional methods for identifying small molecules with

similar effects on the basis of gene-expression profiles, such as hierarchical or partitional clustering with genome-wide correlation as similarity metric. The result was that these methods were not able to group (i.e. to consider "similar") GEPs following treatments with the same drug and the dominant detected structure in the correlation induced space was related to cell type and batch effects (similarity among cells of the same type grown at the same time).

We assessed this property by means of Positive Predicted Value (PPV) analysis: For each differential GEP A in the cMap dataset we sorted all the other GEPs according to their correlation with A (in decreasing order) then we computed for each k "most correlated to A" GEPs the percentage of them that were obtained with the same drug of A, for each $k = 1, \ldots, 453$. Results of this assessment are shown in Figure 3.10 (green curve and green area). Only 35% of the GEPs obtained by treating a cell line with drugs with an average number of 3.2 treated hybridization per batch had as closest neighbor (according to the correlation measure) another GEP obtained by treating cells with the same drug. This percentage is equal to 21% for drugs with an average number of 0.8 treated hybridization per batch and 12% for those with an average number of 0.4 treated hybridization per batch.

On the contrary the percentage of GEPs for which the closest neighbor was a GEP obtained by treating the same cell line was always 100% (see the green curve and the green area in Figure 3.11) while the percentages of those for which the closest neighbor was a GEP obtained in the same batch was equal to 82%, 91% and 91% respectively (see the green curve and the green area in Figure 3.11).

Since, in the cMap online tool there is a mechanism to build internal signatures, deriving set of genes, from the GEPs of the cMap itself, we built a signature for each GEP and we used these signatures as input to the cMap query system. Then we considered the obtained CSs as pair-wise similarities between individual GEPs. As shown in the figures 3.10 and 3.11 (blue curves and blue areas) the performance that we obtained with this method are comparable with those obtained with classical correlation.

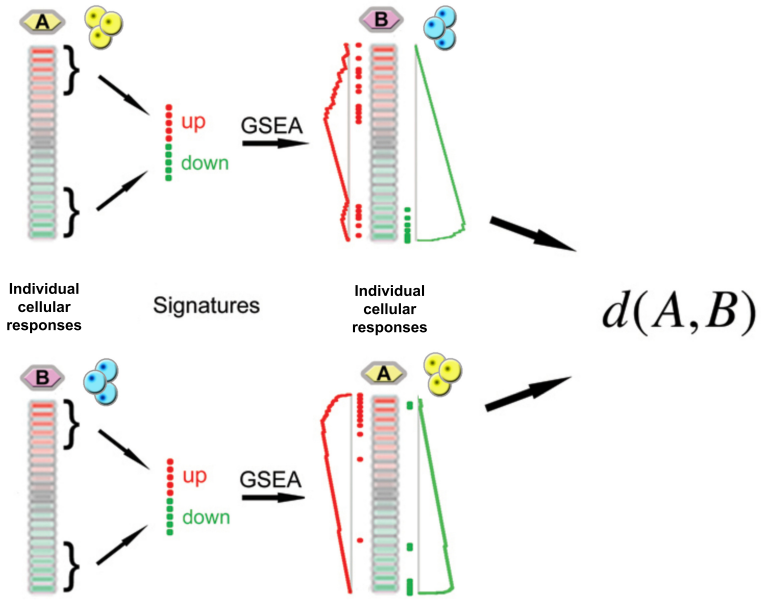

Figure 3.9: Profile-wise GSEA - Profile-wise CS as a measure of similarity between individual cellular responses to cMap compounds

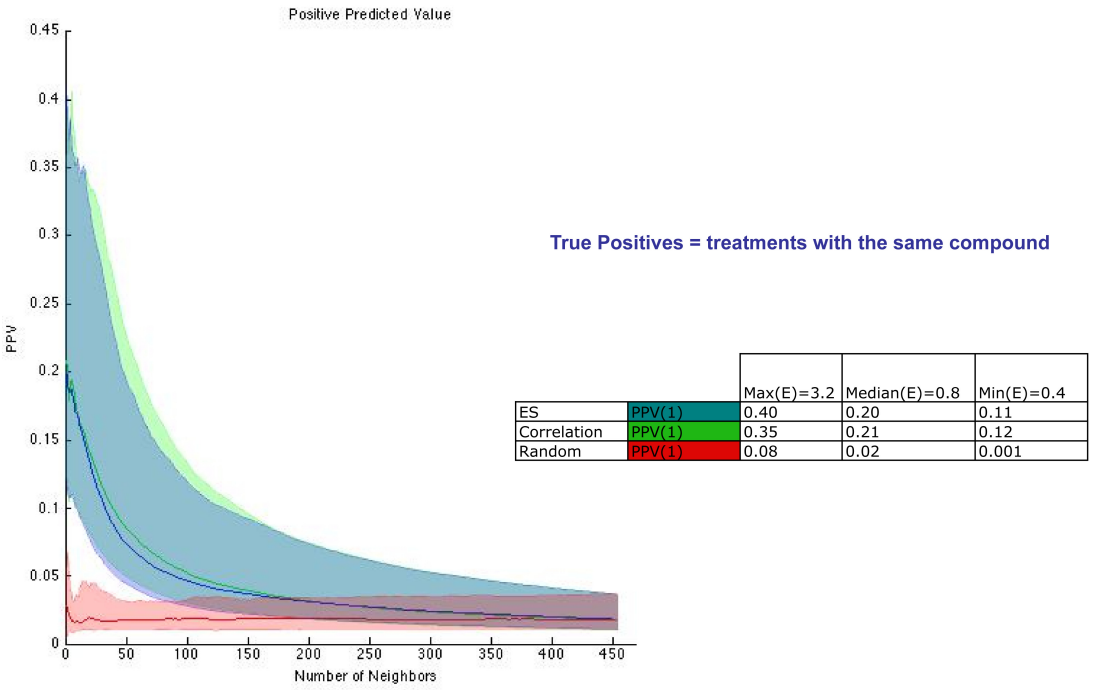

Figure 3.10: Distance performances - Profile-wise similarity performances

True Positives = treatments on the same cell line

		Max(E)=3.2	Median(E)=0.8	Min(E)=0.4
ES	PPV(1)	0.99	0.99	0.99
Correlation	PPV(1)	0.99	0.99	0.99
Random	PPV(1)	0.66	0.63	0.63

True Positives = treatments in the same batch experiment

		Max(E)=3.2	Median(E)=0.8	Min(E)=0.4
ES	PPV(1)	0.83	0.92	0.92
Correlation	PPV(1)	0.82	0.91	0.91
Random	PPV(1)	0.1	0.03	0.03

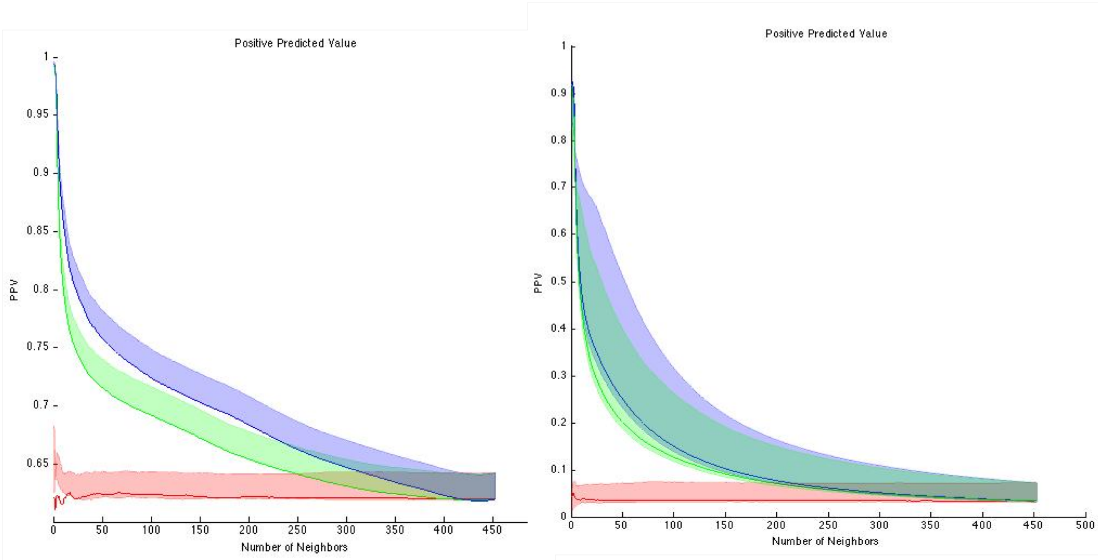

Figure 3.11: Distance performances - Profile-wise similarity performances

4

A novel computational framework for Drug Discovery

4.1 Introduction

We developed an automatic and robust approach that exploits similarity in the gene expression profiles of the cMap dataset to predict similarities in drug effect and MoA. We constructed a Drug Network (DN) of 1,302 nodes (drugs) and 41,047 edges (indicating similarities between pair of drugs) then applied network theory, partitioning drugs into groups of densely interconnected nodes (i.e., communities). These communities were significantly enriched for compounds with similar MoA and can be used to identify the compound-targeted biological pathways. New compounds can be integrated into the network to predict their therapeutic and off-target effects.

An overview of the method is depicted in Figure 4.1. At the heart of the approach is a novel definition of distance between two drugs. This is computed by combining gene expression profiles obtained with the same compound, but in different experimental settings, via an original rank-aggregation method, followed by the application of an established method for the analysis of gene sets along genome-wide ranked lists (Figure 4.1 (**a**)).

A DN is then generated by considering each compound as a node, and adding a weighted edge between two compounds if their similarity distance is below a given significance threshold (Figure 4.1 (**b**)). By using a novel clustering based procedure, we identified topological modules in our DN termed "communities" and we organized them into hi-

erarchical topological structures called rich-clubs. We observed that the whole topology of the resulting DN reflects hierarchy of similarities among the composing drugs and that communities were enriched for drug with similar MoA.
The DN is a powerful and meaningful classifier for novel drugs; previously undescribed compounds can be integrated in the DN by simply using gene expression data following treatment (Figure 4.1 (**c**)); then, the unknown MoA can be revealed by analyzing the surrounding drug communities and their enriched MoAs.
In this chapter we explain how we computed a synthetic consensual response, which is sufficiently independent from the treated phenotype, for each of the drug in the cMap (Section 4.2). The drug distance computation is detailed in Section 4.3 while the assessment of its effectiveness is the topic of Section 6.4. The method used to build the DN, based on our definition of distance between drugs, is explained in Section 6.5, whereas the community identification and the topological analysis description is provided in Section 6.6. Network topology is assessed and related to similarities and differences in drug MoAs in Section 4.7. Particularly, in the subsection 4.7.6 we describe a method based on fuzzy-logic that we conceived in order to additionally characterize each of the communities by analyzing the functional annotations of a corresponding set of consistently modulated genes.
The final goals of the DN and its ability in providing the basis for a general drug classification algorithm are described in the final section.

4.2 Synthetic Consensual Responses to Drugs

As discussed in Chapter 3, an increasing number of published methods builds on the idea that a given set of genes (i.e. a gene-signature) is sufficient to summarize a biological state or condition (108). Gene-signatures and their combined patterns of expression have been successfully used to classify cancers (87), for predicting survival rates in the progression of a disease (20) and for explaining drug resistance or susceptibility (39).
Attempts to summarize the general effect of a drug with a gene signature and to classify drugs on expression patterns alone have so far met with limited success. This happened because selecting genes whose differential expression is a marker of the general-effect of a drug (i.e. it is independent from the specificities in the response of the treated cell line) is non-trivial. Actually, the similarity in gene expression profiles due to unrelated

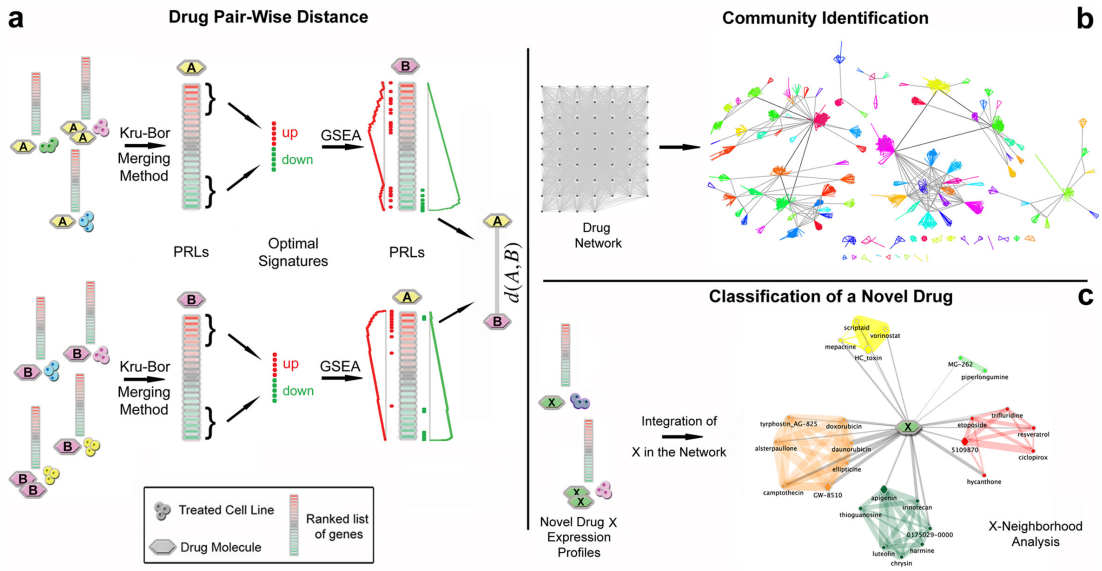

Figure 4.1: Methodology overview - (**a**) A distance value for each couple of drugs is computed. (**b**) Each drug is considered as a node in a network with weighted edges (proportional to distances) connecting pairs of drugs. Network communities are identified. (**c**) Ranked list of differentially expressed genes, following treatment with a novel drug X are merged together, and the distance $d(X, Y)$ is computed for each drug Y in the reference dataset. X is connected to drugs whose distance is below a significant threshold.

stimuli in cells grown at the same time (81) (also called "batch effect") is a phenomenon that needs to be overcome.

Inspired by these considerations, we developed an approach that is able to derive, for each drug in the reference dataset (i.e. the cMap), a consensus synthetic transcriptional response. This response summarizes the transcriptional effect of the drug across multiple treatments on different cell lines and/or at different dosages.

As starting point, for each compound in the cMap, we considered all the transcriptional responses following treatments, across different cell lines and/or at different concentrations. Each of these transcriptional responses was represented as a list of genes ranked according to their differential expression. We then computed a single synthetic ranked list of genes, the Prototype Ranked List (PRL), by merging all the ranked lists referring to the same compound (Figure 4.2).

In order to equally weight the contribution of each of the cell lines to the drug PRL, rank merging was achieved with a procedure we conceived, which is based on a hierarchical majority-voting scheme, where genes consistently over-expressed/down-regulated across the ranked lists are moved at the top/bottom.

We observed that 78% of the compounds contained in the cMap dataset were tested on, at least, three different cell lines (out of five) and just 6% of them were tested on a single cell line. Therefore, for the majority of the compounds in the cMap dataset, we have multiple treatments suitable for being merged together in order to compose a synthetic and general cellular response to the drug. For the minority of drugs (6%) that were tested on a single cell line at a single concentration, we had a single ranked list of genes, and therefore we used this single list as the cellular response to the drug.

In the rest of this chapter we will make use of the following notation:

- P: a set containing all the Microarray Probe-set Identifiers (MPI);
- m: the total number of MPI $= |P|$ (note that, for the microarray platform used in our reference dataset, $m = 22.283$);
- D: the set of all the possible permutations of the same set of m MPI;
- X: a set of ranked lists of MPI computed by sorting, in decreasing order, the genome-wide differential expression profiles obtained by treating cell lines with the same drug, $X \in D$;

- $r : P \times D \to [1, \ldots, m]$: a function with values in the interval $[1, \ldots, m]$, assigning to the couple composed by a MPI $i \in P$ and a ranked list $d \in D$ the rank position of i in d;
- $\delta : D^2 \to N$: the *Spearman's Footrule* distance associating to each pair of ranked lists in X, a natural number quantifying the similarity between them;
- $B : D^2 \to D$: the *Borda Merging Function* associating to each pair of ranked lists in X a new ranked list obtained by merging them with the *Borda Merging Method*.

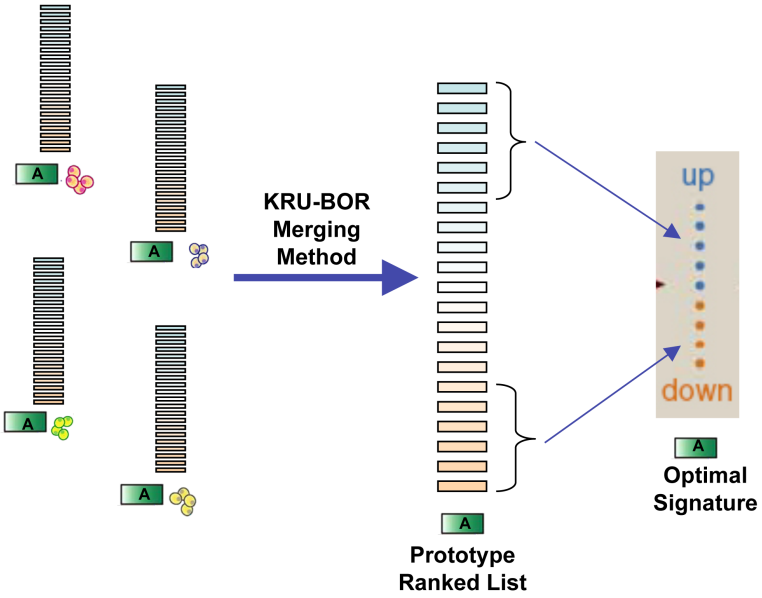

Figure 4.2: Synthetic consensual response to the drug - A drug PRL is obtained by merging together all the GEPs following treatments with a given drug A on a variety of human cell lines with different concentrations.

4.2.1 How to merge ranked lists of objects

We chose to use a famous and simple method to pair-wisely merge ranked lists of genes: the *Borda Merging Function*. This function is defined as $B : D \times D \to D$ and it associates to a couple of ranked lists of the same objects (i.e. two permutations of the same genome-wide set of MPIs) a third list that summarizes the sorting order of both the previous ones. So, $B(x, y) = z$ with $x, y, z \in D$. This function simply implements

a majority voting scheme by computing the list of values $P = [p_1, p_2, \ldots, p_m]$ in which $p_i = r(i, x) + r(i, y), i = 1, \ldots, m$. Where r is the function previously defined.
Finally a new ranked list of probes z is obtained by sorting them according to the corresponding values in P, in increasing order.

4.2.2 Adaptive weighting of individual cell responses

The cMap contains some experiments in which several treatments with a single drug on a particular cell line are available, but only one, or few of them, have been performed on the other cell lines. In this case, applying a single majority voting method (i.e., the Borda Merging Method) will lead to a final merged list in which the response of these cell lines with a smaller number of treatments are poorly represented. In order to avoid an over-representation of the most treated cell line and to equally weight the contribution of each of the cell specific responses, we designed a novel method that is capable to implicitly and adaptively compute these weights while it merges the lists. This novel algorithm builds a PRL for each drug by combining the following tools: a measure of the distance between two ranked lists (Spearman's Footrule), a method to merge two or more ranked lists (the Borda Merging Method) and an algorithm to obtain a single ranked list from a set of them in a hierarchical way (the Kruskal Algorithm) (22, 33, 86). For this reason we named this algorithm the KRUBOR merging method. Similarly to a hierarchical clustering method, it first computes the pair-wise Spearman's Footrule distances between all the ranked lists obtained with the same drug. Then it merges the two closest lists according to this distance with the Borda Merging Method, obtaining a new ranked list. Then this new list replaces the former two (that have been merged together) and the Spearman's Footrule distances are recomputed. This procedure is repeated until only one ranked list remains.

4.2.3 Spearman's Footrule

We computes the Spearman's Footrule by neglecting normalization terms (as m is fixed for all the pairs of ranked lists), as follows:

$$\delta(x, y) = \sum_{i=1}^{m} |r(i, x) - r(i, y)| \tag{4.1}$$

where, $x, y \in X \subseteq D$.

4.2.4 The KRUBOR algorithm

A pseudocode description of the method is in algorithm 1

Algorithm 1 KRUBOR merging method

PRL=**KRUBOR**(X)
input: X, a set of ranked lists of genes.
output: PRL, the prototype ranked list of genes for the drug used to produce X.
1. $n \leftarrow |X|$
2. while $n > 1$
3. find $i, j : x_i, x_j \in X$ and $\delta(x_i, x_j) = \min_{p,q=1,\ldots,n:p\neq q} \delta(x_p, x_q)$
4. $y = B(x_i, x_j)$
5. $X = (X/\{x_i, x_j\}) \cup y$
6. $n \leftarrow |X|$
7. endwhile
8. $PRL \leftarrow x : x \in X$
9. return PRL

The input of the algorithm is X (i.e. the set of all the ranked lists obtained by treating with a given drug). Following the Kruskal Algorithm (22) strategy, the algorithm first searches for the two ranked lists of MPI in X with the smallest Spearman's Footrule distance [line 3]. Then it merges them using the Borda Merging Method [line 4], obtaining the new ranked list of MPI y. In the next step [line 5], the two merged lists are removed from X and the new one is added to it. This process iterates until only one list remains in X and it is deemed as the drug PRL, which is given in output [lines, 8 and 9].

An example is provided in Figure 4.3. In this example we start from the pair-wise Spearman's Footrule distances computed among all the ranked lists obtained by treating a set of different cell lines with alvespimycin, an inhibitor of the Heat Shock Protein 90 (Hsp90) protein. In the figure, each node of the tree is a ranked list of genes, and the euclidean distances between the nodes the Spearman's Footrule distances between the ranked lists that they correspond to. The first two lists that the algorithm merges are those represented by nodes 1 and 2 (the closest ones, i.e. most similar, according to the Spearman's Footrule). These two lists are merged with the Borda Merging Method,

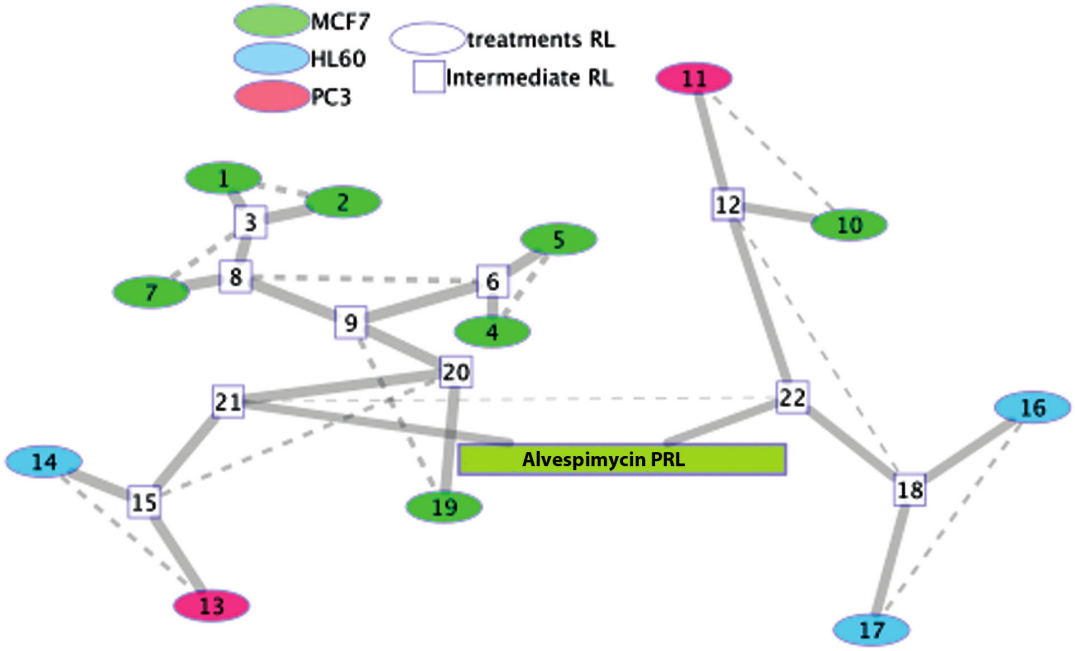

Figure 4.3: Cellular response variability three - Each leaf (ellipses) of this tree represents a ranked list obtained by treating a cell line with alvespymicin in a given batch experiment. Colors specify the treated cell line. Each internal node (square) represents an intermediate ranked list (RL) obtained by merging the two lists represented by its children nodes. The width of an edge connecting two nodes is inversely proportional to the Spearman's Footrule distance between the ranked lists represented by those nodes. The root of the tree is the final PRL of alvespimycin (large green rectangle). Solid lines indicate childhood relationships, while nodes connected by a dashed line are siblings.

yielding the new list represented by node 3. The algorithm continues by merging nodes 4 and 5 (the second closest pair in the set), yielding the list in node 6. This process iteratively continues until the lists represented by nodes 21 and 22 remains. Then they are merged, in the last iteration, and the alvespymicin PRL is obtained. If applied to the cMap, this approach is able to correctly merge ranked lists of differentially expressed genes obtained by multiple treatments with the same drug, adaptively weighting the contribution of each cell specific response to the final PRL. It is possible that, with this algorithm, an outlier (i.e., a ranked list of genes coming from a hybridization with a systematic error) could be overweighed. This could have been prevented by adding a pre-filtering step to outliers. However, even without this time-consuming pre-filtering step, the final results show that our method is robust when applied to the real dataset.

4.3 Drug distance Measure

4.3.1 Drug Optimal Signature

Once a PRL has been obtained for each drug in the cMap, we extract a gene signature p, q, where $p, q \in P$ and $|p| = |q| = 250$, from each of them. To this end, we selected the top-ranked 250 genes from a PRL and the bottom-ranked 250 ones (p and q, respectively). We deem this gene signature to be a synthetic short descriptor summarizing the general cellular response to the drug. In other words, we isolate sets of genes that seem to consistently vary in response to the drug across different experimental conditions (e.g., different cell lines and different dosages). Considerations about the chosen signature size can be found in Section 4.5.4.

4.3.2 Computation of the distance between two drugs

Given the optimal signature of the drug d, with $p = \{p_1, \ldots, p_{250}\}$ (up-regulated genes) and $q = \{q_1, \ldots, q_{250}\}$ (down-regulated genes), we define the Inverse Total Enrichment Score (ITES) of the drug d signature $\{p, q\}$ with respect to the PRL of drug x, as follows:

$$ITES_{d,x} = 1 - \frac{ES_x^p - ES_x^q}{2}. \tag{4.2}$$

Here, ES_x^y (with $y = p$ or q) is the ES (defined in the Section 3.3.1) of the up-regulated part of the signature (resp. the down-regulated one) with respect to the PRL of the drug x. As detailed in Section 3.3.1, ES_x^y ranges in $[-1, 1]$, it is a measure based on the Kolmogorov-Smirnov (KS) statistic, and it quantifies how much a set of genes tends to group at the top of a ranked list (140). The closer this measure is to 1, the closer the genes are to the top of the list. The closer to -1, the closer the genes are to the bottom of the list. $ITES_{d,x}$ ranges in $[0, 2]$, it takes as inputs a signature $\{p, q\}$ and a ranked list of genes x, and it quantifies how much the genes in the p set tend to be grouped at the top of the x PRL and how much the genes in the q set tend to be grouped at its bottom. The closer these two statements are to the truth, the closer to 0 is the value of $ITES_{d,x}$. We defined two different distance measurements among drugs as follows: Given two drugs A and B,

- **AES distance**:

$$aes(A, B) = \frac{ITES_{A,B} + ITES_{B,A}}{2}; \tag{4.3}$$

- **MES distance**:

$$mes(A, B) = \frac{\min(ITES_{A,B}, ITES_{B,A})}{2}. \tag{4.4}$$

As shown in the following section, we verified that the AES distance is more stringent than the MES distance, whereas the MES distance is more sensitive to weak similarities (Tables 6.5 and 6.7).

We computed a distances for each pairs of drug in the cMap, for a total number of $\binom{1,309}{2} = 856,086$ values, by using both definitions. The empirical Probability Density Function (pdf) of the AES distance and the MES distance on the whole cMap are provided in Figure 4.4 and Figure 4.5 respectively.

4.4 Distance assessment

4.4.1 Gold-Standard Definition

In a first assessment of our drug distance we tested the ability of our measure to consider "closer" to each others those pairs of drugs sharing a therapeutic application, or

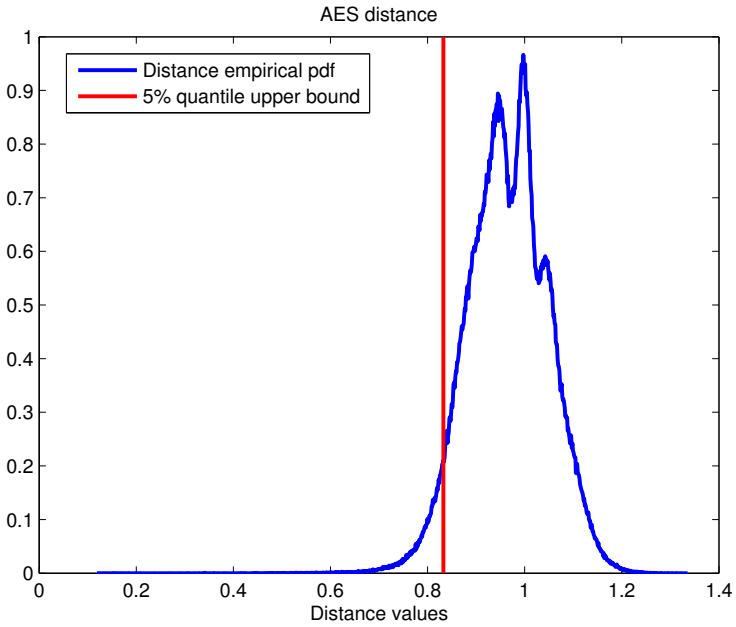

Figure 4.4: AES distance empirical pdf - The empirical probability density function of the AES distance on the whole cMap dataset

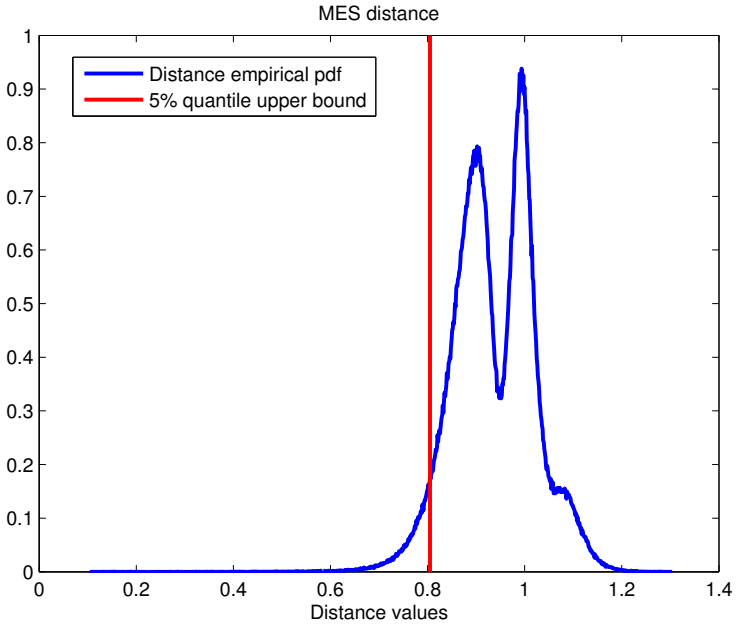

Figure 4.5: MES distance empirical pdf - The empirical probability density function of the MES distance on the whole cMap dataset

a MoA. To this end, we chose as "Gold-Standard" (i.e. a benchmark composed by true predictions) the "drug ontology" of the Anatomical Therapeutic Chemical (ATC) classification system (112, 126).
The ATC classification system is a method for drug cataloguing controlled by the World Health Organization Collaborating Centre (WHOCC) for drug statistics methodology. In this system each drug is coded with an alphanumerical identifier and the drug/code mapping yields a "drug ontology" in which compounds are grouped according to the organ, or system, on which they act and/or their therapeutic and chemical characteristics. Briefly this ontology is composed by a "forest of trees" in which the position of a drug is defined by the corresponding ATC code and ATC code prefixes of different lengths define the 5 different levels of the ontology, i.e. the depth of the trees. An example of ATC code is provided in Figure 4.6.

Each bottom-level ATC (corresponding to a leaf in one of the trees) stands for a therapeutic chemical substance in a single indication (or use), implying that more than one ATC code can be assigned to the same drug: for example, acetylsalicylic acid has A01AD05 as a drug for *local oral treatment*, B01AC06 as a *platelet inhibitor*, and N02BA01 as an *analgesic and antipyretic*.
The first level of the ATC code indicates the anatomical main group and consists of one letter. There are 14 of these main groups (reported in Table 4.1). In the example in Figure 4.6, the first letter of the ATC code of tamoxifen (an estrogen receptor antagonist used in the treatment of breast cancer), is L, which correspond to the tree containing *Antineoplastic and Immunomodulating agents*.

The second level of the ATC code denotes the therapeutic main group and consists of two digits. (in the figure: L02 corresponds to the sub-tree containing compounds used for *Endocrine Therapy*).
As examples, the complete second level of the B tree (i.e. *Blood and blood forming organs*) and L tree (i.e. *Antineoplastic and Immunomodulating agents*) are provided in Tables 4.2 and 4.3 respectively

The third level of the code indicates the therapeutic/pharmacological subgroup and consists of one letter (in the figure: L02B corresponds to the sub-tree containing *Hormone antagonists and related agents*).

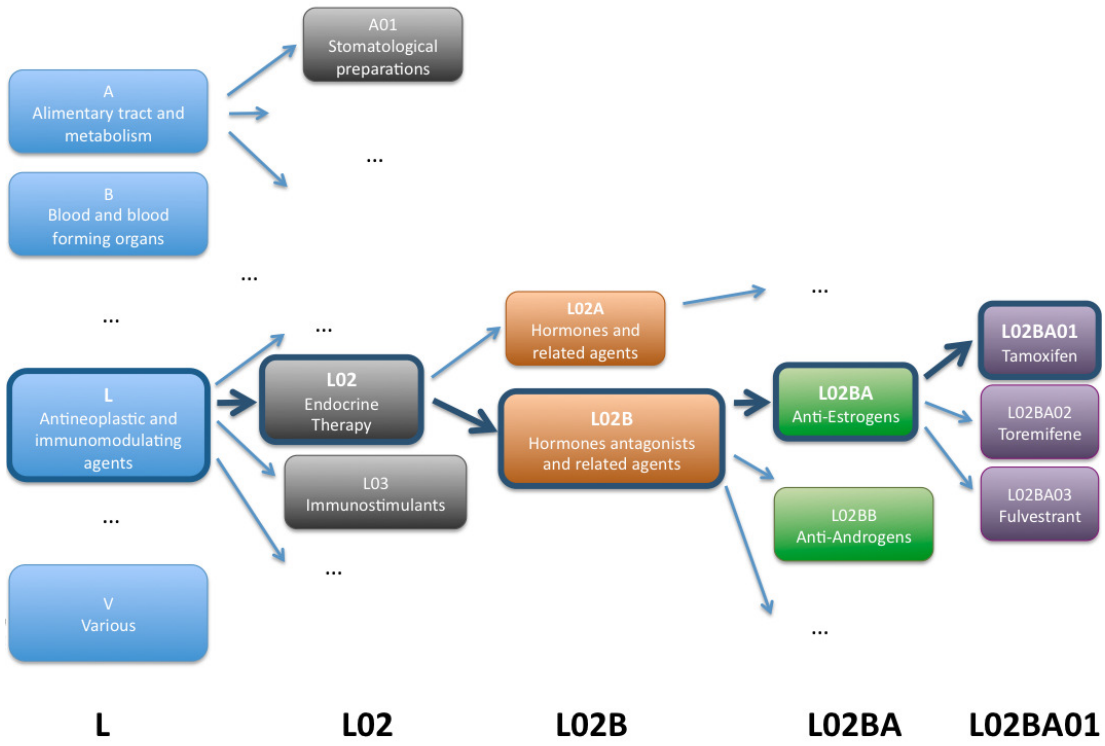

Figure 4.6: ATC code example - An example of ATC coding: Tamoxifen is in the *Antineoplastic and immunomulating agents* tree (letter L at the first level); in the *Endocrine Therapy* subtree (rooted in 02 at the second level); in the *Hormones antagonists and related substances* subtree (rooted in B at the third level); in the *Anti-estrogens* subtree (rooted in A at the fourth level); in the leaf denoted by 01. The final ATC code is L02BA01.

Code	Group
A	Alimentary tract and metabolism
B	Blood and blood forming organs
C	Cardiovascular system
D	Dermatologicals
G	Genito-urinary system and sex hormones
H	Systemic hormonal preparations, excluding sex hormones and insulins
J	Antiinfectives for systemic use
L	Antineoplastic and immunomodulating agents
M	Musculo-skeletal system
N	Nervous system
P	Antiparasitic products, insecticides and repellents
R	Respiratory system
S	Sensory organs
V	Various

Table 4.1: ATC coding system, 1st Level - The 14 main groups defined by the first letter in the ATC coding system.

Code	Group
B01	Antithrombotic agents
B02	Antihemorragic agents
B03	Antianemic preparations
B04	Blood substitutes and perfusion solutions

Table 4.2: ATC coding system, 2st Level of the B tree - The 4 sub-trees in the B tree of the ATC ontology, defined by the second and third digits in the ATC coding system.

Code	Group
L01	Antineoplastic agents
L02	Endocrine Therapy
L03	Immunostimulants
L04	Immunosuppressants

Table 4.3: ATC coding system, 2st Level of the L tree - The 4 sub-trees in the L tree of the ATC ontology, defined by the second and third digits in the ATC coding system.

For example, the third level of the C tree (i.e. *Cardiovascular system*) in the C01 sub-tree (i.e. *Cardiac Therapy*) is reported in Table 4.4 while the third level of the L tree (i.e. *Antineoplastic and Immunomodulating agents*) in the L02 sub-tree (i.e. *Endocrine Therapy*) is reported in Table 4.5.

Code	Group
C01A	Cardiac glycosides
C01B	Antiarrhythmics, class I and III
C01C	Cardiac stimulants excl. cardiac glycosides
C01D	Vasodilators used in cardiac diseases
C01E	Other cardiac preparations

Table 4.4: ATC coding system, 3rd Level of the C tree in C01 sub-tree - The 5 sub-trees in the C tree and C01 subtree of the ATC ontology, defined by the fourth letter in the ATC coding system.

Code	Group
L02A	Hormones and related agents
L02B	Hormones antagonists and related agents

Table 4.5: ATC coding system, 3rd Level of the L tree in L02 sub-tree - The 2 sub-trees in the L tree and L02 subtree of the ATC ontology, defined by the fourth letter in the ATC coding system.

The fourth level of the code indicates the chemical/therapeutic/pharmacological subgroup and consists of one letter. (in the figure: L02BA corresponds to the sub-tree containing *Anti-estrogens*).
For example, the other trees of this 4th level (L tree, L02 sub-tree, L02B sub-tree) are reported in Table 4.6.

The fifth level of the code denotes the chemical substance and consists of two digits (in the figure: L02BA01 corresponds to the tamoxifen leaf). For example, the complete list of leafs in the fifth level of the G tree (i.e. *Genito urinary system and sex hormones*), G03 sub-tree (*Sex hormones and modulators of the genital system*), G03D sub-tree (*Progestogens*) and G03DC sub-tree (*Estrogen derivatives*) is reported in Table 4.7.

The ATC coding system can be queried at the following URL: `http://www.whocc.no/atc_ddd_index/`.

Code	Group
L02BA	Anti-estrogens
L02BB	Anti-androgens
L02BG	Enzyme inhibitors
L02BX	Other hormone antagonists and related agents

Table 4.6: ATC coding system, 4th Level in the L02B sub-tree - The 4 sub-trees in the L02B sub-tree of the ATC ontology, defined by the fifth letter in the ATC coding system.

Code	Group
G03DC01	allylestrenol
G03DC02	norethisterone
G03DC03	lynestrenol
G03DC04	ethisterone
G03DC06	etynodiol
G03DC31	mehtylestrenolone

Table 4.7: ATC coding system, 4th Level in the L02B sub-tree - The 4 sub-trees in the L02B sub-tree of the ATC ontology, defined by the fifth letter in the ATC coding system.

Given the features of the ATC coding system, is reasonable to assume that two drugs sharing a prefix of their ATC code could share their therapeutic application, hence could have a similar MoA. The specificity of this commonality depends on the length of this shared prefix. For this reason, we assessed the reliability of our drug distance by labeling all the compounds in the cMap according to their ATC code. Since only 768 out of the 1,309 cMap compounds are classified with an ATC code, we restricted our analysis on the subset composed by the $\binom{768}{2} = 294,528$ similarity distances among couples of ATC-coded drugs. Then we ranked these distances in ascending order, and computed the curves shown in Figure 4.7 by meaning of Receiver Operating Characteristic (ROC) analysis as explained in the following section.

4.4.2 Assessment Methodology

In a typical *Binary Classification Problem* the task is to assign the objects contained in a given set to one among two different categories: the positive and the negative one. The PPV, or *precision rate* is a measure of the classification performances in such a problem and it is given by the proportion of objects who are correctly assigned to positive category (i.e. the True Positive (TP) prediction) out of the total number of object assigned to this category without taking into account of the prediction correctness (i.e. the set composed by the TP and False Positive (FP) prediction). Formally:

$$PPV = \frac{|TPs|}{|TPs| + |FPs|}. \tag{4.5}$$

In order to evaluate the performances of our drug distance, we considered a binary classification problem in which the set of objects to classify was composed by all the possible couples of ATC-coded drugs in the cMap. The two categories to predict were:

1. couples of drugs sharing an ATC code (positive category);

2. couples of drugs without a common ATC code (negative category).

Finally, we considered as positive prediction (i.e. assigned to the positive category) the set of drug-pairs corresponding to the smallest k drug distances (i.e. the first k couples of closest drugs, according to our distance). In this way we quantified the tendency

of our drug distance to consider "close to each other" drugs with similar MoA (i.e. sharing an ATC code).

Particularly, we computed PPVs considering as positive predictions the drug-pairs corresponding to the first k smallest drug distance values, with $k = 1, \ldots, 280,875$. Each of these drug-pairs was considered as a TP prediction if the two composing drugs shared an ATC code prefix.

4.4.3 Results

For each $k = 1, \ldots, 280,875$, we computed the PPV as the percentage of true positives TP out of the total number of positive prediction (i.e. TP and FP, k). By considering as positive category the set composed by couples of drugs sharing an ATC code of length 1, 3, 4 and 5 (1st, 2nd, 3rd and 4th level of specificity in the ATC coding system), respectively, we obtained the 4 curves plotted in Figure 4.7 for the first $10,000$ smallest distances (first 100 in the magnification).

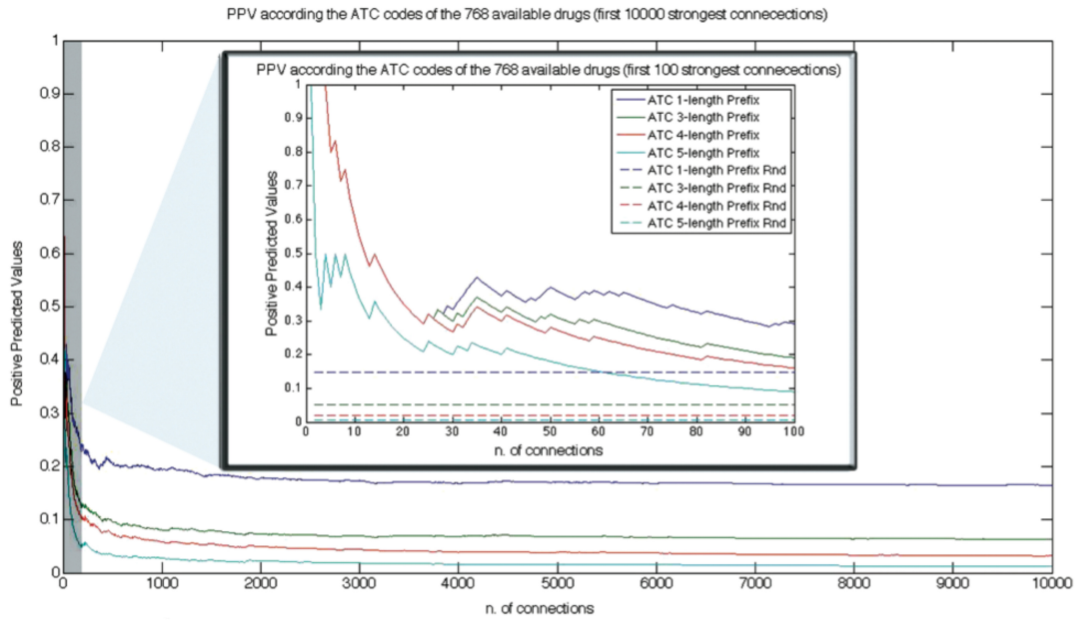

Figure 4.7: Drug distance performances - PPVs versus number of considered distances (from the smallest to the largest). A true positive prediction is defined as the distance between two drugs sharing an ATC code prefix. Different colors are for curves computed considering ATC code prefixes of different lenghts.

In order to analyze the significance of the obtained PPV curves we computed the expected PPV had the k considered drug distances been randomly chosen independently from their values (dashed line in Figure 4.7).
Given a set of s objects (i.e. ATC-coded drug-pairs) containing n objects of the positive category (i.e. drug-pairs sharing an ATC code), with $n \leq s$, and $s - n$ objects of the negative category (drug-pairs with no commonalities in their ATC codes), if we randomly select a subset of k of these objects, then the expected number of objects belonging to the positive category among the k selected ones can be computed considering the following hypergeometric distribution function

$$f(x, s, n, k) = \frac{\binom{n}{x}\binom{s-n}{k-x}}{\binom{s}{k}}, \tag{4.6}$$

which quantifies the probability of having x objects belonging to the positive category among the k randomly selected ones (with $x \leq k \leq n \leq s$). The average value of this function (i.e. the expected value of objects belonging to the positive category) is given by:

$$\frac{kn}{s}. \tag{4.7}$$

This implies that the expected PPV (i.e. the random PPV) is equal to the ratio between the expected number of correct predictions obtained by chance (the last formula) and k :

$$\frac{kn}{sk} = \frac{n}{s}. \tag{4.8}$$

By using this formula we computed the dashed lines in Figure 4.7 assessing that the classification performances achievable with our drug distance (i.e. its tendency to consider close to each other drugs with similar MoA) were clearly far from the random ones at each level of ATC-coding specificity.
In conclusion, according to our distance, drugs sharing a therapeutic application tends to be close to each other.

4.5 From pair-wise similarities to Drug-Network

4.5.1 Network Evolution

Once we computed a distance value for each pair of drugs of the cMap, we can build a DN by considering each of the drugs as a node and adding a weighted edge between

two drugs if their distance is below a given threshold. A first example of the results that we obtained by doing this is provided in Figure 4.8 (A).

In the first panel of the Figure (A) is represented the first sub-network obtained by linking two drugs with a weighted edge if their AES distance is under the very stringent threshold of 0.2 and considering only drugs with at least one incident edge.
This first sub-network contains two connected components. The first one consists of the following drugs (ATC codes are reported where available): digoxigenin, digoxin [C01AA05], digitoxigenin, lanatoside C [C01AA06] and helveticoside: cardiac glycosides from the plants *digitalis purpurea*, *digitalis orientalis* and *digitalis lanata*; ouabain and proscillaridin [C01AB51]: cardiac glycosides found in the ripe seeds of the african plant *strophantus* and from the plant *scilla maritima*, respectively. All of these drugs are mainly used to threat congestive heart failure and arrhythmia by increasing the force of contraction of the heart (52). One of the mechanisms through which this could be achieved is by increasing the availability of intracellular Calcium ions (Ca^{2+}) (123) and cardiac glycosides are effective to this because they inhibit the plasma membrane Sodium-Potassium pump (Na^+/K^+-ATPase) (123), leading to increased intracellular Sodium ion (Na^+) and Ca^{2+} and decreased intracellular Potassium ion (K^+) (74).
The second component consists of: cephaeline and emetine, two natural alkaloids used as anti-protozoal, as vomiting inducing agents and for blocking protein synthesis in eucaryotic cells. This last effect is due to the binding of these compounds to the 40S subunit of the ribosome, thus by blocking the elongation during protein synthesis (5).
This first result shows the efficacy of our approach in building a drug network from our drug distances since we used only GEPs without any prior knowledge about the MoA of the drugs or their chemical descriptors.
In the rest of this section we show how the topology of the DN grows coherently with the MoAs of the included drugs when the distance threshold for the edge inclusion increases.
By including connections between drugs whose AES distance was less than 0.3, we obtained the network depicted in Figure 4.8 (B).

With respect to the previous network, new connections appeared in the component containing the cardiac glycosides, whereas cicloheximide, an inhibitor of protein synthesis in eukaryotic organisms (38), joined the other protein synthesis inhibitors; a cluster

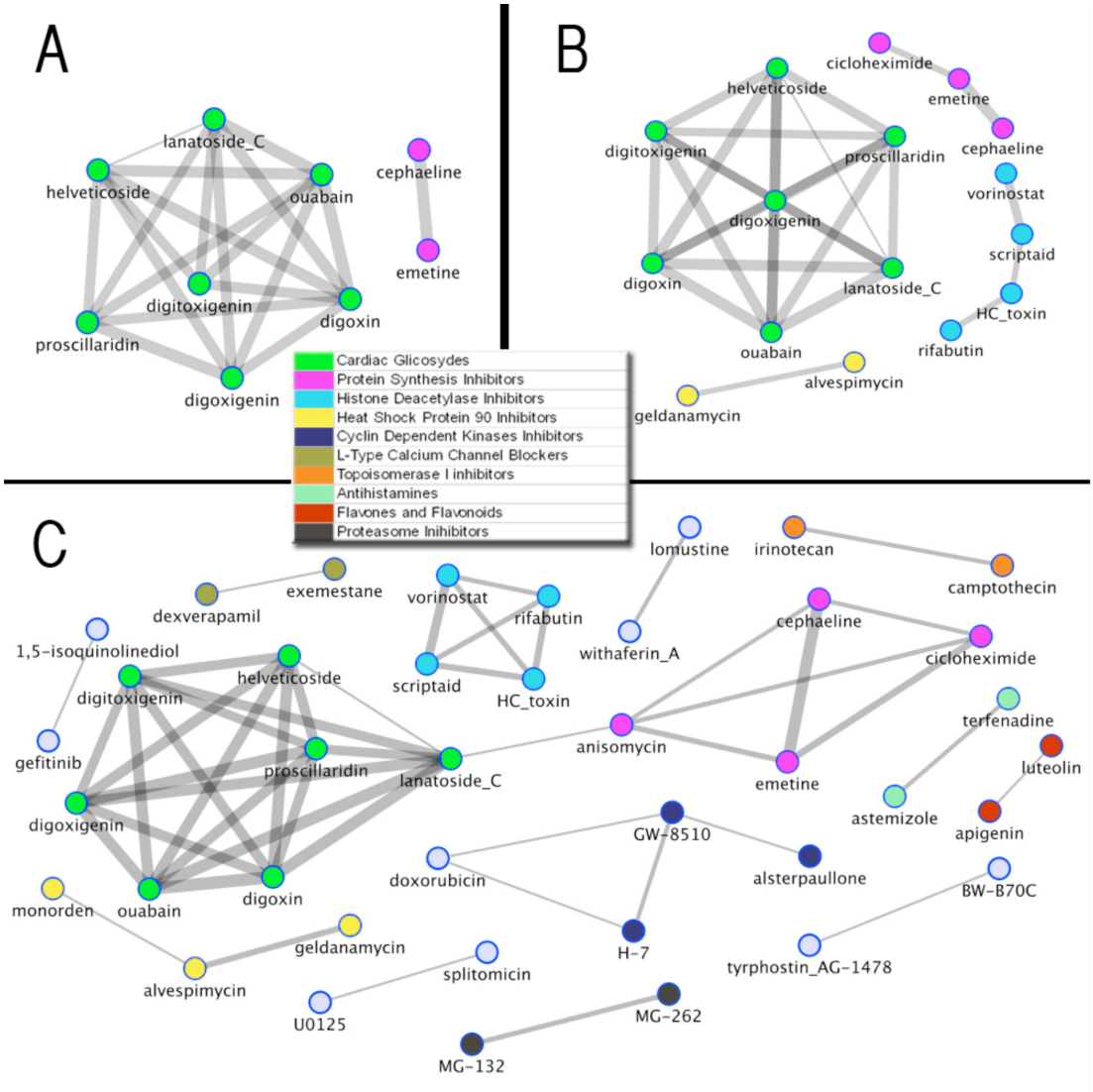

Figure 4.8: Network evolution - Each drug is a node and there is an edge between two drugs if their distance is less than a fixed threshold; edge thickness are inversely proportional to distance values; nodes are painted according to the known MoA of the corresponding drugs. (A) A first drug network obtained with distance threshold equal to 0.2; (B) Drug network obtained by linking two drugs if their distance is less than 0.3. (C) Drug network obtained by linking two drugs if their distance is less than 0.4.

of Histone deacetylases (HDAC) inhibitors (vorinostat, scriptaid, HC toxin), containing also the bactericidal antibiotic rifabutin [J04AB04], appeared, and two inhibitors of the Hsp90 protein (geldanamycin and alvespymicin) were linked together. Hsp90 is one of the most important "molecular chaperones" that plays a number of important roles, which include protein folding, intracellular transport, maintenance, and degradation of proteins as well as facilitating cell signaling.

Further increasing the distance threshold to 0.4 yielded the network in Figure 4.8 (C). Several drugs were added: monorden (an Hsp90 inhibitor) coherently joined the other two, whereas anisomycin, a protein synthesis inhibitor (51), was linked to the right component; a small cluster containing CDKs inhibitors (GW-8510, alsterpaullone and H-7) appeared, luteolin and apigenin (a flavone and a flavonoid, classes of plant secondary metabolites) became connected, as well as campthotecin and irinotecan [L01XX19] (TopoI inhibitors), astemizole [R06AX11] and terfenadine [R06AX12] (two antihistamines), MG-262 and MG-132 (two proteasome inhibitors), dexverapamil and examestane (two L-type calcium channel blockers), etc.
Interestingly, a weak edge between lanatoside C and anisomycin connected the component containing cardiac glycosides to protein synthesis inhibitors. This "second level" of similarity between these two drugs could be due to the fact that both of them inhibit caspase-3 (26, 115).
Increasing the distance threshold level to 0.6 yielded a network of 158 drugs with 379 edges.

4.5.2 Statistical significance of the Drug Distance

Because we had a huge number of pairwise distance values ($\binom{1,309}{2} = 856,086$) we assumed the empirical pdf estimable from these sample data as a good approximation of the real one and we used it to compute a statistically significant threshold level for the drug distance. Specifically, we chose as a significance threshold value the upper bound of the 5% quantile of the empirical Probability Density Function (pdf). These values were 0.8327 and 0.8065, respectively AES distance and MES distance.
In this way, given a distance value d, the corresponding empirical p-value can be computed by dividing the number of distances less than d in the whole set of all the possible ones by the cardinality of this set (i.e., 856,086).

Obviously, the empirical p-values of threshold values were equal to 0.05. Because we built the DN by linking with a weighted edge two drugs if their distance was below a statistically significant threshold, all of the included edges correspond to significant distances and the smaller the distance, the higher it is in statistical significance (additional considerations on the distance threshold impact on the network topology can be found in the following sections).

4.5.3 The final Drug Network

The final DN, obtained by linking two drug nodes if their MES distance was below the statistically significant threshold of 0.8065 is shown in Figure 4.9.

This network was composed by a giant connected component containing 1,302 drug nodes out of 1,309 and 41,047 edges, corresponding to 5% of a fully connected network with the same number of nodes (856,086 edges).
The average shortest path length was 2.5 and the average local clustering coefficient was 0.44 whereas the maximum shortest path length was 7.
The cumulative degree distribution of the DN is shown in Figure 4.10. In the plotting, the horizontal axis is the vertex degree k and the vertical axis is the cumulative probability distribution of degrees (i.e. the fraction of vertices that have degree greater than or equal to k). Clearly, this distribution does not follow the power-law and the network seems not to have a scale-free topology.

4.5.4 Network Robustness

We heuristically determined the size of the drug optimal signatures $\{p, q\}$ (whose computation is detailed in Section 4.3.1) guided by the following considerations. We tested optimal signatures of different length k and for each value of k, we computed distances among drugs and derived a drug network, always using the same distance significance threshold. We observed that the network obtained with the smallest k always contained, as a subnetwork, the networks obtained with larger k values. This means that, as the signature length k increases, the overall structure of the network does not change substantially. We chose $k = 250$ as a good compromise, which takes into account the number of considered genes (which should be sufficiently small), the edge density of the obtained network, and the network prediction performances (the network assessment

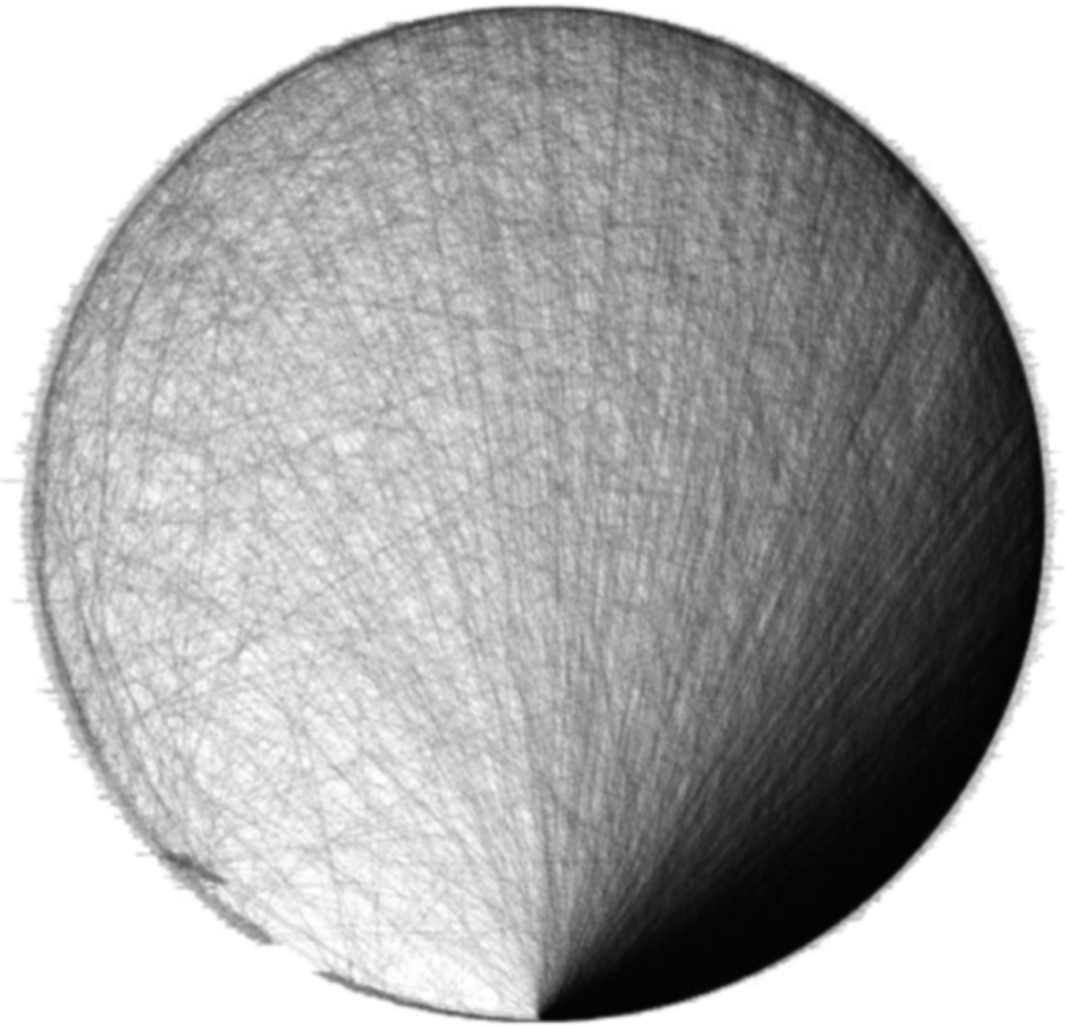

Figure 4.9: Final network - A final Drug Network obtained with a statistically significant threshold for the MES drug distance

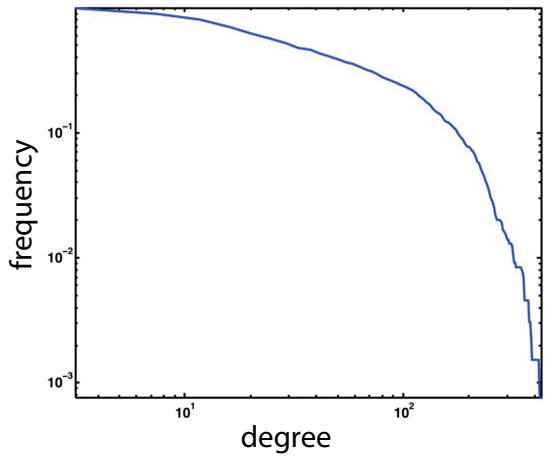

Figure 4.10: Network statistics - Cumulative degree distribution of the final DN

is detailed in the following sections).

Finally, we observed that the network topology, in terms of drug communities (whose computation is detailed in the following sections), did not change substantially if we chose a different significance threshold value. This happened because the community-finding algorithms we used were working on weighted edges (i.e., distances) and, therefore, were not very sensitive to the addition or removal of few edges, due to slightly different choices of the distance significance threshold.

4.6 Community Identification and Topological analysis

The final DN, which was obtained by considering only significant edges (see Figure 4.9), was very dense and hard to analyze and visualize. In order to "compress" it and make it easily analyzable we provided it with modularity by grouping its nodes in communities and rich-clubs with a novel hierarchical clustering algorithm.

We first applied a slightly modified version of the Girvan-Newman algorithm (48) to a drug network composed by 158 drug nodes and 379 edges, obtained by considering only AES drug distance values lower than 0.6 (which was quite far from the significant threshold of 0.8327), as a pilot study.

We decided to apply the same algorithm to the final DN (built on MES distances and considering the statistically significant threshold values) however it was too computationally expansive and hence infeasible.

Therefore, we turned to a novel clustering algorithm based on passing messages between data-points (44). By slightly refining its implementation for our purpose, we obtained the modular and hierarchical DN shown in Figure 4.14.

4.6.1 Girvan-Newman Algorithm for finding communities in complex networks

The Girvan-Newman algorithm (48) is a method for hierarchically clustering the nodes of a graph basing on its topology. The implemented method groups together nodes that are very interconnected among each other but with fewer connections with the rest of the graph (see the definition of community in Section 2.3). For this reason, this algorithm is deemed as a method to "identify" communities by considering modularity as an inherent property of a graph. It is based on the concept of "betweennes" (or "centrality") of the edges which are iteratively removed from the graph defining, at each iteration, a level of clustering in the hierarchy (in a top-down fashion).
The centrality of an edge is defined as the number of shortest paths between pairs of nodes that run through it. The rationale behind this strategy is that if a network is modular (i.e. contains communities) then it contains groups of nodes that are densely connected to each other but poorly connected with other groups by a few intergroup edges. Therefore all shortest paths between nodes lying in two distinct groups of these must run through one of these few edges. This imply that the edges connecting two nodes in different communities will have high betweenness. Hence, by removing these edges, the groups are separated and the underlying modular structure of the graph is revealed.
Informally, in each iteration, the algorithm first solves the all-pairs shortest path problem; for each edge e in the graph, it counts how many shortest paths include e, i.e. it computes the centrality of e; finally, the edge with the highest centrality is removed. At this point, if the number of connected components of the graph increases (meaning that one or more connected components disconnected into two sub-components) then the novel connected components are returned as communities in the current level of the hierarchy (which is updated) and the algorithm goes to the next iteration. If the graph does not become disconnected, then the algorithm goes to the next iteration directly, without updating the clustering hierarchy. Iterations end when no more edges remain in the graph.

As every other hierarchical clustering algorithm, this method gives in output a hierarchy of clusters (i.e. communities) containing the whole graph at the first level, a set of few communities at the second one which are divided in sub-communities in the third one and so on. At the bottom level the hierarchy contains single nodes. There is no way to compute the best level at which this hierarchy should be cut and set of disjointed communities considered. Moreover our DN is weighted and the Girvan-Newman Algorithm does not take this into account.
For this reasons we slightly modified this algorithm by including a heuristic way to automatically stop the "community desegregation" ending up into a set of disjointed groups of nodes, at a certain level of the hierarchy, and to take into account of the edge weights. We did this by adding the additional constraints, to each iteration:

1. Solve the weighted all-pairs shortest path problem;
2. If an identified community is a singleton (i.e. is composed by a single node) then remove it from the network;
3. If the number of identified community at the current level is lower than the previous one then stop the computation.

A pseudo-code of our version of the Girvan-Newman algorithm is provided in algorithm 2. Committing a little abuse of notation we will use $e \in p$ to denote that the edge e is part of the path p (i.e. the path p runs along e) and we will use $p \in G$ to indicate that p is an path on the graph G.

The algorithm takes in input a weighted graph and gives in output a hierarchy of network communities.
When the computation begins the connected components of the graph are computed and considered as the first level of the community hierarchy [lines 1-3]. Note that if the graph is connected then all its nodes form a unique community while if it is not connected then each connected component is considered as a first level community.
A cycle of iterations runs until the set of edges is emtpy [line 4]. In each of these iterations the minimum weight all-pairs shortest paths are computed and assigned to the set P [line 5]. Formally, P is the set of the admissible paths of G, p, for which does not exist any other path p' in G with the same extremity nodes and with a lower total weight.

Algorithm 2 Modified Girvan-Newman

$NCH = \mathbf{MGN}(G)$

input: $G = (V, E, \omega)$, a weighted network.

output: NCH, a network community hierarchy on the network G.

1. $h \leftarrow 1$
2. $\bar{C}^* = \left\{C : C \subseteq V, \forall (x,y) \in C^2 : \exists p \in G | s(p) = x, d(p) = y\right\}$
3. $NCH(h) \leftarrow \bar{C}^*$
4. while $E \neq \emptyset$
5. $\quad P \leftarrow \{p \in G | \nexists p' \in G : s(p) = s(p'),\ d(p) = d(p'), \bar{\omega}(p') < \bar{\omega}(p)\}$
6. $\quad$ find $i : \max\limits_{e_j \in E} |\{p \in P | e_j \in p\}| = |\{p \in P | e_i \in p\}|$
7. $\quad E \leftarrow E - \{e_i\}$
8. $\quad \bar{C} = \left\{C : C \subseteq N, \forall (x,y) \in C^2 : \exists p \in G | s(p) = x, d(p) = y\right\}$
9. $\quad$ for each $c \in \bar{C} : |c| = 1$
10. $\quad\quad \bar{C} \leftarrow \bar{C} - \{c\}$
11. $\quad$ endfor
12. $\quad$ if $|\bar{C}| > |\bar{C}^*|$
13. $\quad\quad h \leftarrow h + 1$
14. $\quad\quad NCH(h) \leftarrow \bar{C} - \bar{C}^*$
15. $\quad\quad \bar{C}^* \leftarrow \bar{C}$
16. $\quad$ endif
17. $\quad$ if $|\bar{C}| < |\bar{C}^*|$
18. $\quad\quad$ return NCH
19. $\quad$ endif
20. endwhile
21. return NCH

At this point the edge with the highest centrality is searched and deleted from the graph [lines 6 and 7]. After this removal, the connected components of the graph are recomputed [line 8] and assigned to the set $\bar{C}$. If some single nodes become disconnected from the graph (i.e. there are connected components containing a single nodes) then they are removed from the set $\bar{C}$ [lines 9-11].
If the number of connected components in $\bar{C}$ is greater than the number of those identified in the previous iteration then the novel connected components are considered as network communities and the hierarchy is updated [lines 12-16].
If the number of connected components in $\bar{C}$ is lower than the number of those identified in the previous iteration then the computation ends and the hierarchy of communities is returned in output.
In a pilot study, we applied this algorithm to a drug network composed by 158 drug nodes and 379 edges, obtained by considering as significant the edges corresponding to AES distance values lower than 0.6. This drug network is depicted in Figure 4.11 and the 33 identified communities are highlighted in blue. The composing drugs are reported for some of them in Table 4.8.

These results further confirmed the efficacy of our drug distance in defining a "pseudo-metric" space in which drugs sharing a MoA are placed in close positions and drugs can be clustered in sufficiently homogeneous groups with unsupervised learning methods.
Particularly, some of the communities in Table 4.8 and 4.9 contained drugs targeting the same proteins (communities 2, 4, 5 and 8), having the same therapeutic application (communities 1, 6 and 9) or with similar chemical characteristics (communities 3, 11, 12, 13) and modulating the activity of the same biological pathways (communities 7 and 10).
Applying this algorithm to the complete DN (Figure 4.9) obtained by considering significant edges according the threshold levels described in subsection 4.5.2 would have been too computationally expansive. For this reason we implemented a novel method based on a recent and more efficient clustering algorithm.

4.6.2 Clustering by Passing Messages between datapoints

As explained in Section 2.3.1, one of the problems linked to cluster analysis is that is not possible to automatically compute the number of clusters. Clustering data by iden-

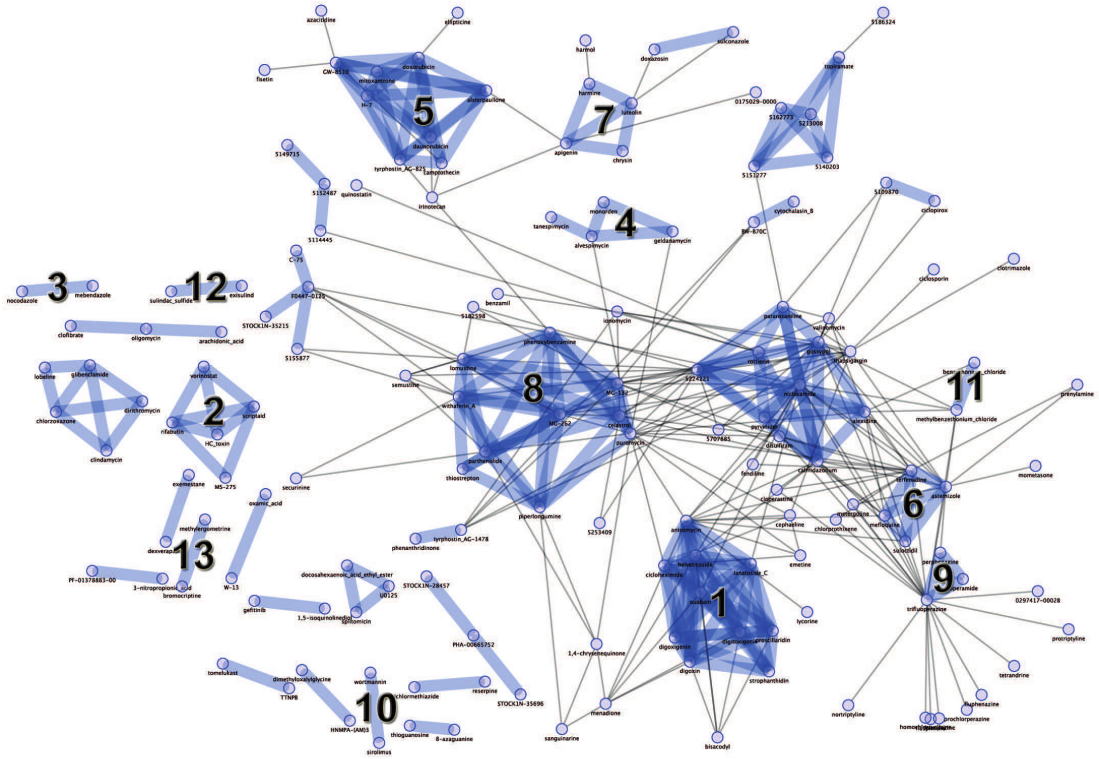

Figure 4.11: Girvan-Newman network communities - Results of a first community identification pilot study on a DN obtained by filtering out edges corresponding to AES drug distances greater than 0.6. This network contained 158 drug nodes and 379 edges. 33 network communities were identified by our modified version of the Girvan-Newman algorithm. The composition of the numbered communities are reported in table 4.8.

Community n. 1			
anisomycin	cicloheximide	digitoxigenin	digoxigenin
digoxin	elveticoside	lanatoside C	ouabain
proscillaridin	strophanthidin		
Drug Commonality: Cardiac glicosydes and protein synthesis inhibitors			
Community n. 2			
HC toxin	MS-275	rifabutin	scriptaid
vorinostat			
Drug Commonality: HDAC inhibitors			
Community n. 3			
nocodazole	mebendazole		
Drug Commonality: Benzimidazoles			
Community n. 4			
alvespimycin	geldanamycin	monorden	tanespimycin
Drug Commonality: Hsp90 inhibitors			
Community n. 5			
GW-8510	H-7	alsterpaullone	camptothecin
daunorubicin	doxorubicin	mitoxantrone	tyrphostin AG-825
Drug Commonality: CDKs inhibitors, anthracyclines and Topo inhibitors			
Community n. 6			
astemizole	mefloquine	suloctidil	
Drug Commonality: Antihistamines, anticholinergics			
Community n. 7			
apigenin	chrysin	harmine	luteolin
Drug Commonality: PPAR-gamma modulators			
Community n. 8			
MG-132	MG-262	celastrol	lomustine
parthenolide	phenoxybenzamine	piperlongumine	puromycin
thiostrepton	withaferin A		
Drug Commonality: Proteasome inhibitors			

Table 4.8: Girvan-Newman Communities (a) - The composition of some of the 33 communities identified with our version of the Girvan-Newman algorithm on a DN generated with an AES distance threshold equal to 0.6.

		Community n. 9
loperamide	perphenazine	trifluoperazine
Drug Commonality:	Antipsychotics	
		Community n. 10
sirolimus	wortmannin	
Drug Commonality:	PI3K inhibitor	
		Community n. 11
benzethonium chloride	methylbenzethonium chloride	
Drug Commonality:	Quaternary ammonium compounds	
		Community n. 12
exisulind	sulindac sulfide	
Drug Commonality:	Sulindac metabolites	
		Community n. 13
bromocriptine	methylergometrine	
Drug Commonality:	Ergot alkaloids derivatives	

Table 4.9: Girvan-Newman Communities (b) - The composition of some of the 33 communities identified with our version of the Girvan-Newman algorithm on a DN generated with an AES distance threshold equal to 0.6.

tifying a subset of representative examples is important for processing sensory signals, detecting patterns in data. Such "exemplars" can be found by randomly choosing an initial subset of data points and then iteratively refining it, as the K-means algorithm does. However this works well only if that initial choice is close to a good solution and there is no way to determine the optimal number of exemplar. A recent method, based on "affinity propagation" (44) takes as input measures of similarity between pairs of data points, then it allows real-valued messages exchanging between data points until a high-quality set of exemplars and corresponding clusters gradually emerges. For each of the identified clusters the element whose features best interpolate the features of all the other points in the cluster (i.e. the cluster exemplar) is also given in output.
The Affinity Propagation Clustering (APC) algorithm requirement consists in a pairwise distance matrix and a set of probabilities, one for each node to be elected as exemplar of a cluster.
If this probability is assumed to be uniform then the algorithm automatically calculates it and infers the proper number of exemplars, hence of clusters, exclusively from the data. From this point of view, APC can be considered striking and revolutionary and probably this is the key aspect contributing to its great success.
How affinity propagation works is illustrated for two-dimensional data points in Figure 4.12, where negative euclidean distance was used to measure similarity. Each point is colored according to the current evidence that it is a cluster exemplar. The darkness of the arrow directed from point i to point k corresponds to the strength of the transmitted message that point i "prefers" k as exemplar of its cluster. This preferences are determined by the distances. The more i is close to k the more it will prefer k as exemplar of its cluster.

So, as a first step, "responsibilities" $r(i,k)$ are sent from data points to candidate exemplars and indicate how strongly each data point favors the candidate exemplar over other candidate exemplars. Then "availabilities" $a(i,k)$ are sent from candidate exemplars to data points and indicate to what degree each candidate exemplar is available as a cluster exemplar for the data point. The "availability" $a(i,k)$, sent from candidate exemplar point k to point i, reflects the accumulated evidence for how appropriate it would be for point i to choose point k as its exemplar, taking into account the support from other points that point k should be an exemplar. Basing on these messages, points

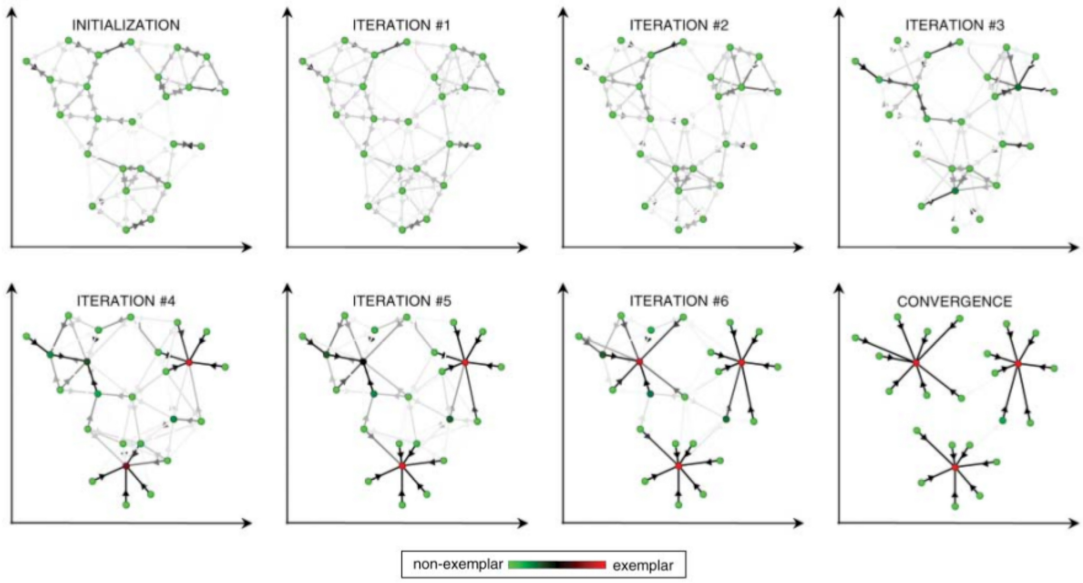

Figure 4.12: Affinity propagation algorithm - The affinity propagation is illustrated for two-dimensional data points, where negative euclidean distance was used to measure similarity. Each point is colored according to the current evidence that it is a cluster exemplar. The darkness of the arrow directed from point i to point k corresponds to the strength of the transmitted message that point i prefers to belong to the cluster whose exemplar is k.

are re-assigned to clusters, novel clusters arise or existing ones disappear.

When updating the messages, it is important that they be dampened to avoid numerical oscillations that arise in some circumstances. Each message is sent to λ times its value from the previous iteration plus $1-\lambda$ times is prescribed updated value, where the damping factor λ is between 0 and 1. Each iteration of affinity propagation consist of (i) updating all responsibilities given the availabilities, (ii) updating all availabilities given the responsibilities, and (iii) combining availabilities and responsibilities to monitor the exemplar decisions and terminate the algorithm when these decisions did not change for ten iterations. To begin availabilities are initialized to zero.

We used a slightly modified version of the APC algorithm for "providing" modularity to our DN. In fact, we used the whole MES pair-wise distance matrix without using any filtering threshold on the edges. In this way we treated the problem of identifying communities in our DN as an usual cluster analysis problem. As a consequence, the network modularity here is conceptually different with respect to the approach we used with the Girvan-Newman algorithm. In fact, we here dealt with an exhaustive set of pair-wise distances among drugs (rather than a network) and from this we built a modular and hierarchical network.
In order to turn the clusters obtained with the affinity propagation algorithm into network communities, to build a final modular and hierarchical connected network rather than a set of disjointed communities and to investigate how different network levels reflect the hierarchy of similarities in the drug MoAs, we conceived a novel hierarchical clustering algorithm based on the APC one.

4.6.3 Building a modular network by recursive affinity propagation clustering

We used the APC algorithm in a recursive and hierarchical fashion. In the first step we computed clusters of drugs by starting from the whole MES pair-wise distance matrix. We considered each of the drugs as a potential cluster exemplar. In the second phase, we focused on the cluster exemplars only (and the corresponding MES sub-matrix) and we clustered them again with the APC. Then we continued this procedure recursively until no more data-points were merged together.

The clusters obtained at each iteration step were turned out into disjointed network communities as follows. For each cluster C we built a community (i.e. a sub-network) $G_C = (N_C, E_C)$ in which $N_C = C$ and $\forall\,(x,y) \in N_C \times N_C : (x,y) \in E_C \Leftrightarrow mes(x,y) \leqslant 0.8065$. In other words we built a community by linking the nodes clustered in C through edges corresponding to statistically significant MES distances. For construction, most of the communities built in this way contain a hub node ("the rich guy") that is the exemplar of the corresponding cluster. For simplicity we will refer to it as the community exemplar.
The DN community obtained in the first iteration step (i.e. the first-level communities) are shown in Figure 4.13. Extending this method to cluster exemplars, we obtained communities of community exemplars, in other words we obtained communities of rich-guys (i.e. rich-clubs). The rich-clubs had their own exemplar so the procedure was iterated until no more nodes were clustered together and the final hierarchical and modular network depicted in Figure 4.14 was obtained.

A pseudo-code of our recursive procedure, named NeTwork by Recursive Affinity Propagation (N-TRAP) , is provided in algorithm 3. In that pseudo-code we denoted with D a set of n data-points, with M an $n \times n$ symmetric matrix in which $M_{i,j}$ is the distance between the data-points x_i and x_j.
Given the distance matrix M and a set of data-points $p \subseteq D$, we used the notation M_p to refer to the sub-matrix of M that contains the pair-wise distances among the points in p.
Additionally, in the pseudo-code we denote with $APC(M)$ the call to the APC algorithm by passing the distance matrix M to it and by denoting with $\{C, p'\}$ its output. We omitted the vector of probabilities of each nodes to be elected as cluster exemplar implicitly by deeming all of them eligible with the same probability. C is the set of clusters computed by the APC, which is a partition of D.
Formally $C = \{c | c \subset D\}$ such that

- $\forall i, j = 1, \ldots, |C| : h \neq k \Rightarrow c_i \cap c_j = \emptyset$;
- $\bigcup_{i=1}^{|C|} c_i = D$.

$p' \subseteq D$ is the set of cluster exemplars. Please note that when no data-points are merged together then $p' = D$ and $C = \{x | x \in D\}$, in other words each of the data-points belong to a distinct cluster and it is the exemplar of that cluster.

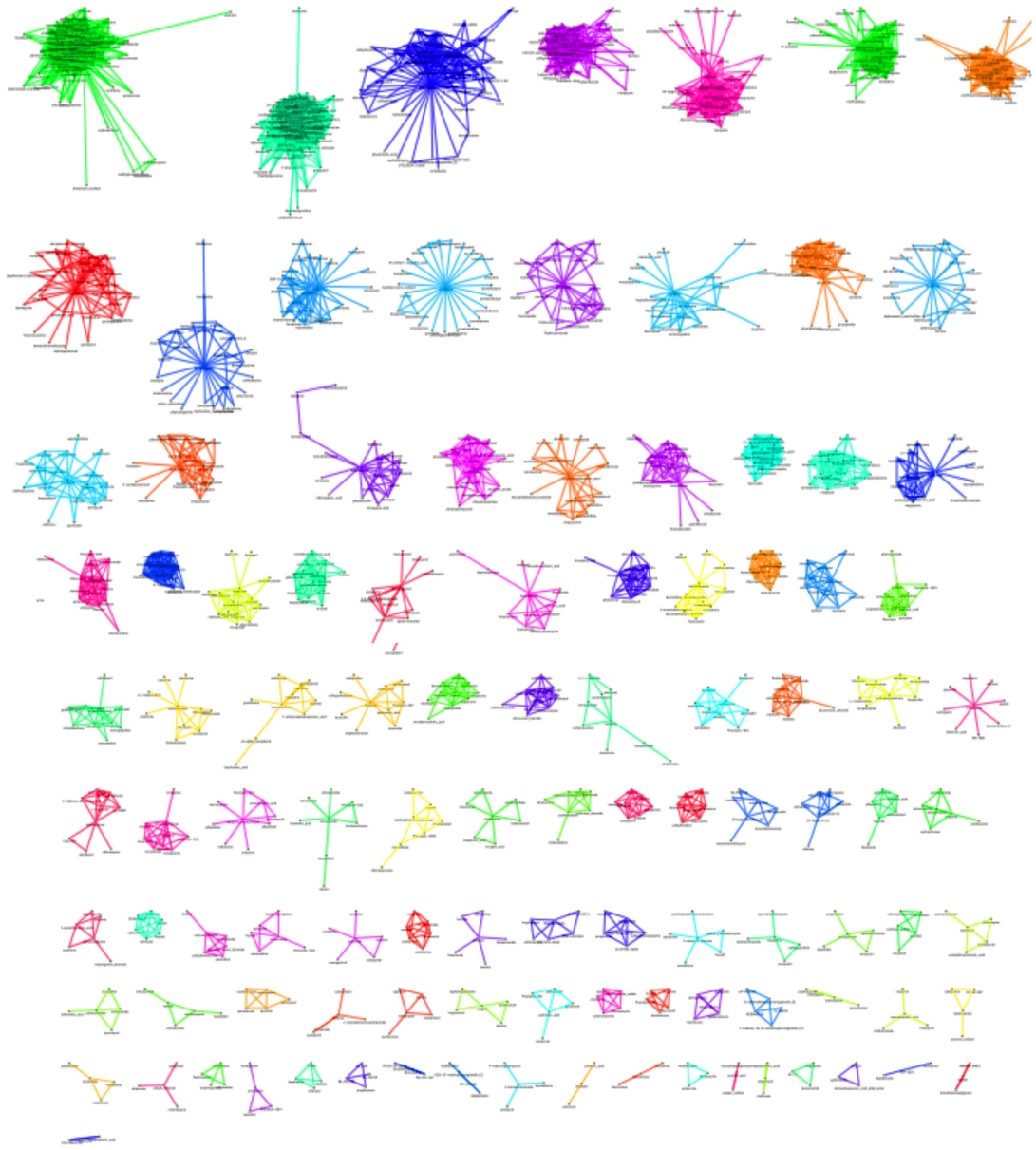

Figure 4.13: Drug communities - The first-level DN communities obtained with the N-TRAP algorithm on the whole MES distance matrix, assuming that each drug node is eligible as cluster exemplar. In order to turn clusters into communities, drugs in the same cluster have been linked through edges corresponding to statistically significant MES distances. Communities have been coded with a numerical identifier and a color.

Algorithm 3 N-TRAP

$N =$ **N-TRAP**(M, p)

input: p, a set of data-points;
M, a matrix containing the distances among all the data-points in p

output: $N = (V_N, E_N, \omega_N)$, a weighted network which contains communities and rich-club structure.

1. $V_N \leftarrow p$
2. $E_N \leftarrow \emptyset$
3. $\omega_N : E_N \rightarrow 0$
4. $N \leftarrow (V_N, E_N, \omega_N)$
5. if $p = \emptyset$
6. return N
7. endif
8. $\{C, p'\} \leftarrow APC(M)$
9. if $|p'| = |C|$
10. $p' = \emptyset$
11. else
12. for each $c_i \in C$
13. for each $(x, y) \in c_i^2$
14. if $M_{x,y} < 0.8065$
15. $E_N \leftarrow E_N \cup (x, y)$
16. $\omega((x, y)) = M_{x,y}$
17. endif
18. endfor
19. endfor
20. endif
21. $N' =$N-TRAP$(M_{p'}, p')$
22. $E_N \leftarrow E_N \cup E'_N$
23. for each $(x, y) \in E'_N$
24. $\omega((x, y)) = \omega'((x, y))$
25. endfor
26. $N \leftarrow (V_N, E_N, \omega_N)$
27. return N

The algorithm takes in input a set of data-points (i.e. p) and a matrix containing the distances among all the data-points in p (i.e. M). The output is a weighted network containing communities and rich-club structure that are obtained by using the APC algorithm in a hierarchical fashion and by adding significant edges among nodes in the same cluster.
The first 7 lines of the pseudo-code are for the termination condition. If the parameter p is an empty set [line 5] then an empty network (built in [lines 1-4], i.e. a set of nodes without connections) is given in output and the computation ends.
If it is not the case, then the data-points are clustered with the APC algorithm [line 8]. If no data-points are clustered together (i.e. each data-point is the exemplar of its own cluster) [line 9] then the set of exemplars is emptied and the computation continues to the next recursion, which ends immediately after the first 6 lines of code giving in output an empty network. This network is added to that of the previous recursion level [lines 22-26], which is given in output terminating the computation. Alternatively, if the APC algorithm clusters together at least two data-points then the instruction block contained between lines 12 and 19 is executed. Each cluster is considered in turn [line 12]. If the distance between two data-points in the considered cluster is less than the significant threshold [line 14], then the nodes corresponding to those data-points are linked through an edge in the current recursion level network [line 15] and the weight of that edge is equal to the distance between the two data-points [line 16].
Finally, the algorithm is called again recursively on the set of cluster exemplar and the corresponding sub-matrix of M [line 21]. When the new level of recursion ends, the novel edges given in output in that level are added to the previous recursion level network [line 22] with the proper weights [lines 23-25] and the algorithm ends giving in output the network.
When applied to the matrix containing all the pair-wise MES distances the N-TRAP algorithm gave in output a set of 106 drug communities at the first level of recursion. These communities are shown in Figure 4.13. The average number of drugs per community was equal to 11.62. The largest community was n. 90, containing 79 drugs, while the smallest ones were communities n. 17, 18, 24, 45, 51, 56, 65, 78, 94 and 103. At the second level of recursion the 106 community exemplars were clustered in 9 rich-clubs containing an average number of 9 community exemplars. At the third level rich-club 6 exemplars were clustered into two super-rich-clubs containing 4 and 2

rich-club exemplars respectively. Finally the exemplars of these 2 super-rich-clubs were linked together yielding the final hierarchical network depicted in Figure 4.14. The composition of all the drug communities and rich-clubs together with some the public available information for each drug (from the DrugBank (148) and ChemBank (127) repositories) is contained in the SDD [**SDD1-CommRichClubs.xls**].

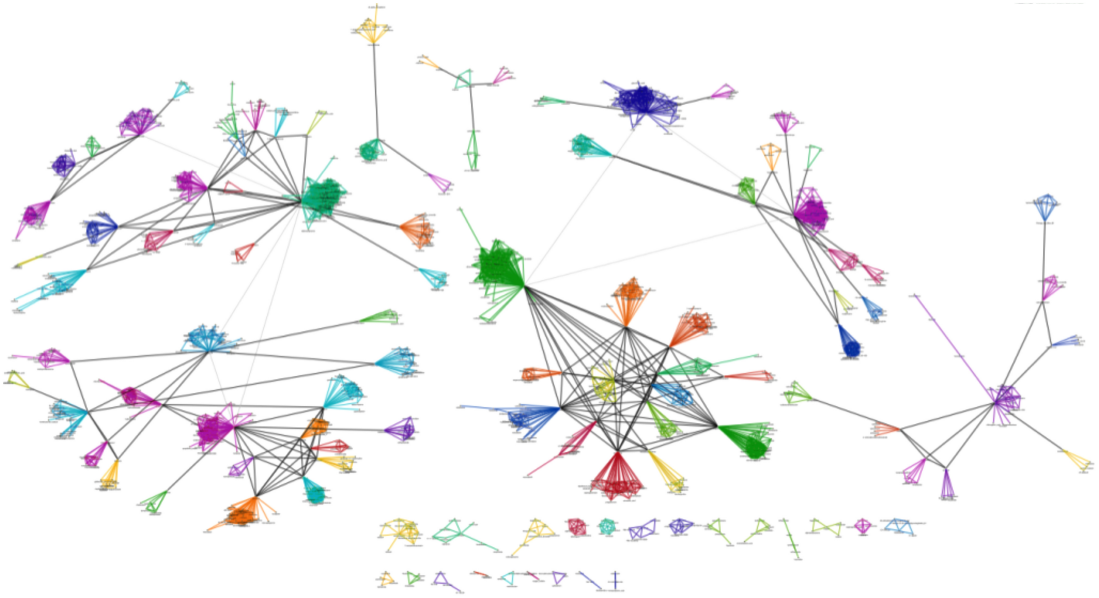

Figure 4.14: The Drug Network - The modular and hierarchial DN obtained by using the N-TRAP algorithm on the MES drug distance matrix.

The final DN contained 1233 drugs in 27 connected components (note that 76 drugs did not cluster with any other drugs at the first iteration so were excluded from the DN).

The average number of drug nodes in each component was 45.67. The largest component (the big subnetwork in Figure 4.14) contained 547 drug nodes whereas the smallest ones contained 3 drug nodes.

The total number of edges was equal to 5,403, i.e. 0.7% of a fully connected network with the same number of nodes.

The average shortest path length was equal to 3.82 and the average local clustering coefficient was equal to 0.65. Finally the longest shortest path contained 8 edges. By

comparing these characteristics with those of the network obtained by including all the significant edges (described at the end of subsection 4.5.3), it results that even if we significantly "compressed" the network ($\approx 1.5\%$ of the significant edges was included) we increased the average local clustering coefficient for $\approx 47\%$, increasing the average shortest path length only for a $\approx 52\%$. It means that this network contains few edges respect to the previous one but the neighbors of each node are more interconnected among each other and the degree of separation between each couple of nodes is only slightly increased. In conclusion we provided modularity to the network composed by all the significant edges. Additionally now the degree distribution is closer to a power-law (Figure 4.15).

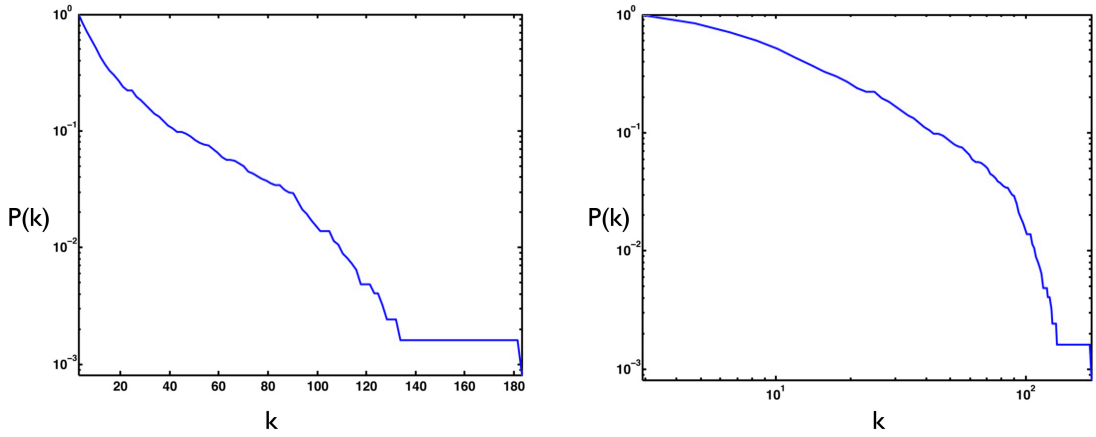

Figure 4.15: Post-processed network statistics - The degree cumulative distribution of the network obtained with the N-TRAP algorithm. Linear scale (left plot) and logarithmic scale (right plot).

4.7 Network Assessment

Similarly to what we observed for the communities obtained with the our version of the Girvan-Newman algorithm, also the N-TRAP communities contain drugs with similar effects. Some examples are reported in Figure 4.16.

In order to formally assess this property and, more generally, how the topological properties of our DN reveals similarities and differences in the MoA of the composing drugs we assessed, first of all, that the tendency of our method to group drugs in the same community was not exclusively due to trivial chemical commonalities (as

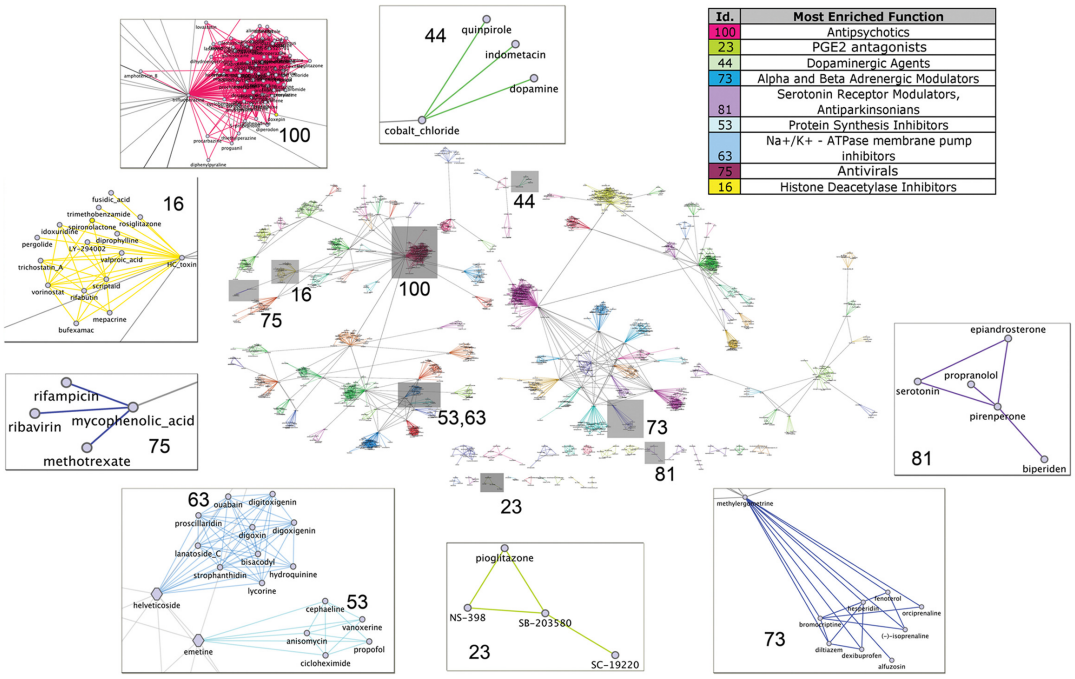

Figure 4.16: N-TRAP communities contain similar drugs - The communities obtained with the N-TRAP algorithm tend to contain drugs with similar effects. In the insets some communities are magnified, and the enriched MoAs are provided in the legend.

detailed in subsection 4.7.5). We next determined whether drugs within a community shared a common MoA. For this reason, we collected for each drug the ATC code, the known direct target genes (from DrugBank (148) and ChemBank (127)), and other literature-based evidences. As explained in Section 4.4.1, ATC codes (112, 126) are alphanumerical strings assigned by the WHOCC to group drugs according to their therapeutic and chemical profiles. ATC codes were available for 59% of the drugs (768 out of 1,309). We retrieved the known target genes for 535 out of 1,309 (41%) drugs from the public repositories, DrugBank (148) and ChemBank (127). We thus assigned a known MoA to 804 drugs out of 1,309 (61%).

For each community, we counted the number of contained drugs with the same MoA. We then divided this number by the number one would expect had the drugs been randomly grouped, to compute odds ratios and p-values.

We further checked if compounds in the same community impinge on common biological pathways. To this aim we developed a Fuzzy-Logic based approach to identify a common set of genes that was consistently up-, or, down-regulated in the PRLs of the compounds in the same community. We thus associated over-represented gene functional annotations, i.e. GO terms (7), to the drug communities by performing a GO enrichment analysis on the common set of genes.

We finally assessed the opposite tendency, i.e. whether compounds characterized by the same MoA end up in the same drug community.

4.7.1 Statistical Testing

We validated each community by checking if ATC codes or target genes were surprisingly overrepresented among those associated to its composing drugs (or vice versa checking if drugs with similar MoA, i.e., same ATC codes or target gene, were found in the same communities). In a similar way, we searched for enriched GO terms when we analyzed sets of genes that were differentially expressed after treatments with all drugs in a community.

In both cases we had to analyze frequencies of terms (ATC codes/target genes and GO terms, respectively) within given sets (drug communities and set of genes, respectively). Therefore, we performed the same statistical test in both analyses.

In order to test the enrichment significance of each ATC code/target gene in a drug community, and to quantify it through a p-value assignment, we had to compute the

probability of counting, by chance, at least k occurrences of a given ATC code/target gene among those associated to the n drugs within a community. If we know that, in the total set D of N drugs, m of them are associated to the given ATC code/target gene, then that probability follows the hypergeometric distribution and is given by

$$\Pr\{X \geqslant k\} = \sum_{x=k}^{\infty} \left(\begin{array}{c} m \\ x \end{array} \right) \left(\begin{array}{c} N-m \\ n-x \end{array} \right) \Big/ \left(\begin{array}{c} N \\ n \end{array} \right). \tag{4.9}$$

In the same way, p-values were computed for assessing the significance of a given GO term enrichment within those associated to genes in a given set. Finally, correction for multiple hypothesis testing was applied to the obtained p-values. The odds ratio (number of observed terms divided by the expected value) was computed as follows:

$$\frac{k}{E(X)} = k\frac{N}{nm} \tag{4.10}$$

4.7.2 Community enrichments

We found that 52 out of 95 assessable communities (i.e., those containing at least two compounds with known MoA) were significantly enriched (p-value < 0.05) for compounds with similar MoA. Specifically, 3 communities were enriched for a direct target gene, 28 for one *ATC* code, whereas 21 were enriched for both a direct target gene and an *ATC* code. Additionally, by searching the literature for supporting evidences, we found 43 communities including several compounds with similar MoA, 9 of which were composed by compounds with no ATC codes and no known target genes. So the total number of enriched communities was 61 $(52 + 9)$ (as reported in Figure 4.17). This number goes up to 77, considering as significant communities, those with a corresponding significant odds ratio greater than 1.
The whole lists of communities enriched for a given MoA (literature based evidence, ATC codes and direct target gene) are provided in appendices B.1, B.2 and B.3, together with the computed p-values and odds-ratios.

4.7.3 Mode of Action enrichments

To assess the opposite tendency, i.e. how compounds characterized by the same MoA end up in the same drug community, we considered in the set of 804 compounds with known MoA a subset of 698 drugs (i.e., with an ATC code or a known target gene).

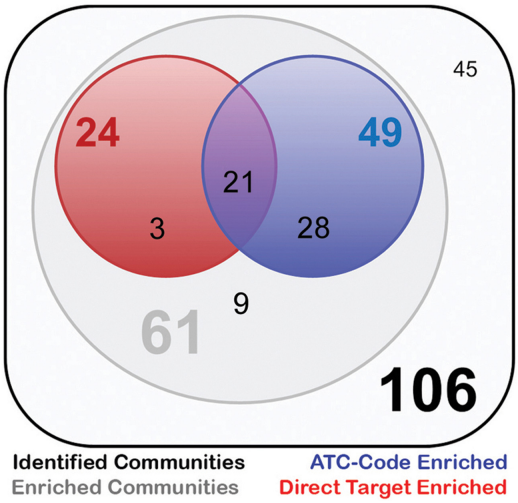

Figure 4.17: Community enrichments - Overview of the community enrichments

This subset contained only the drugs sharing their MoA with at least another drug and was divided in 429 groups (not mutually disjointed) of drugs with the same MoA.
We verified that the MoA of 512 drugs (out of 698) was enriched for a specific community (p-value < 0.05). This number goes up to 586 drugs, considering those with a significant odds ratio greater than 1 (Figure 4.18). The whole lists of MoA (i.e. ATC codes and direct target genes) enriched for occurrences in a given community are provided in appendices C.1 and C.2, together with the computed p-values and odds-ratios.

4.7.4 Network hierarchy reflects different degrees of similarity

The DN has a community structure, which reflects different levels of similarities in the MoA of the composing drugs. In Figure 4.19, an example of this property is reported. The rich-club in Figure 4.19, contains a group of communities whose exemplars are interlinked. This group contains community n. 28 (enriched for Hsp90 inhibitors), community n. 53 (enriched for inhibitors of elongation during protein synthesis), community n. 40 (enriched for proteasome inhibitors and Ubiquitin Proteasome System (UPS) modulators), community n. 104 (enriched for UPS modulators). Even if acting on different intracellular targets, all these classes of drugs produce as a down-stream effect a stress in the cellular environment due to an increased presence of unfolded/misfolded protein. In fact Hsp90 is a molecular chaperone, the UPS system is responsible for the degradation of misfolded protein and a premature stop of the elongation during the

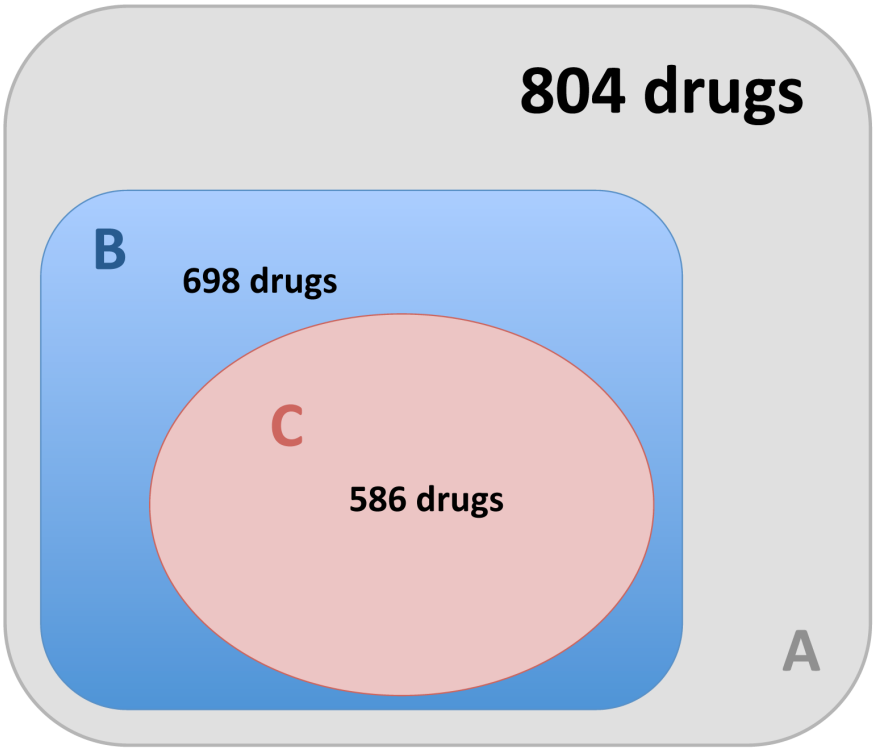

Figure 4.18: MoAs enrichments - Overview of the MoA enrichments

protein synthesis do not allow the polypeptide chains to fold correctly. Consequently, all these drugs cause the up-regulation of genes involved in the response to this stress. So also this secondary effect is detectable at a transcriptional level and reflected by the hierarchy of the network.

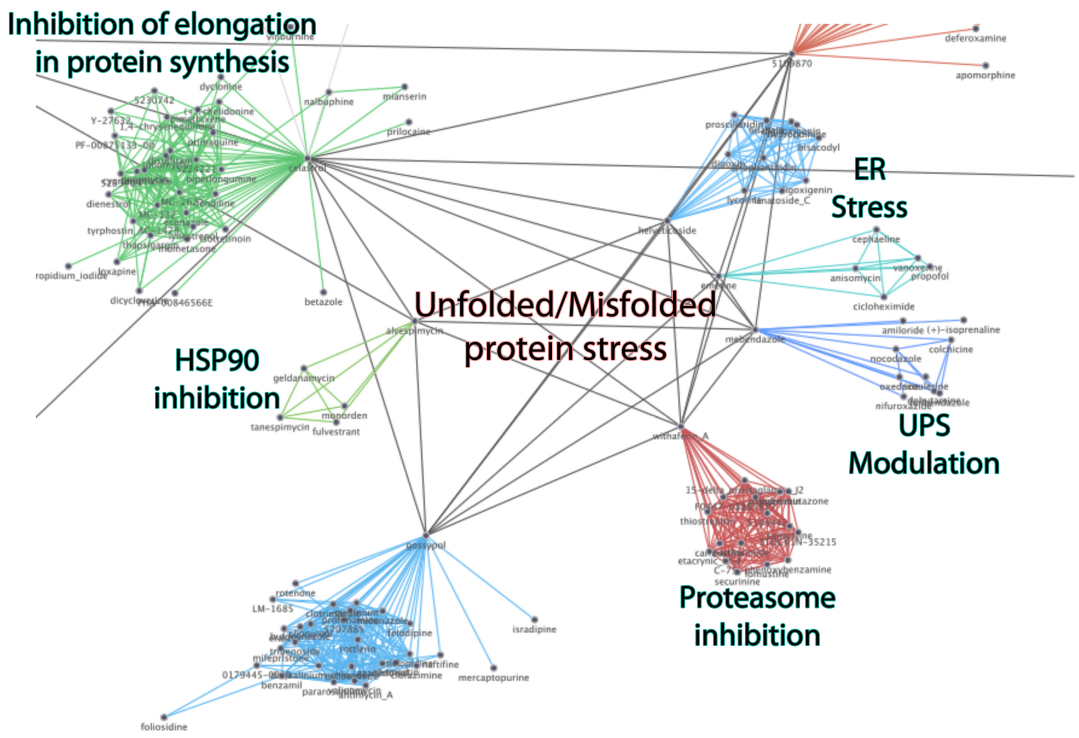

Figure 4.19: Hierarchy of similarities and topology - Hierarchies of similarity in MoA are reflected by the network topology.

4.7.5 Influence of Chemical Commonalities on drug distance and network topology

In order to test whether drugs that are found to be similar according to our drug distance and the network topology could have also been identified simply by looking at their chemical similarities, we first collected the canonical Simplified Molecular Input Line Entry Specification (SMILES) (146) describing the chemical structure of the cMap drugs, and we then computed chemical similarities among them. Finally, we checked if any correlation between chemical similarity and our drug distance was present.
A SMILES is a specification for unambiguously describing the structure of chemical

molecules using short text strings. SMILES were available on the DrugBank database (148) for 579 cMap drugs (out of 1,309).

We focused on this subset of drugs by computing $\binom{179}{2} = 167,331$ pair-wise chemical similarities with two different methods (both working on SMILES): The first one was based on a definition of distance between molecular Electrotopological States (ESF) (53, 54), whereas the second one is based on comparisons between extended-connectivity fingerprints and, making use of a software tool from SciTegic®, computes a property distance inversely proportional to chemical similarity (applications can be found in (63, 70, 102, 122)).

In Figure 4.20, each point represents a pair of drugs for which both the SMILES were available. The first coordinate of each point is equal to the MES distance between the two drugs (DN distance). The second coordinate is equal to 1 minus the ESF similarity between the SMILES of the two drugs.

As apparent, there is no significant correlation between our distance and the ESF similarity (Pearson correlation coefficient (121) between these two measurements is equal to 0.04).

In the same way, there is no significant correlation between the MES of distance and the extended-connectivity fingerprints Property Distance. Also in this case, both the correlation plot and the Pearson Correlation Coefficient (0.05) show that there is no significant correlation between these two distances. This is a first evidence that chemical commonalities between two drugs have no significant influences on their DN distance. As a matter of fact, in very few cases (i.e., points on the figure) with the MES distance less that 0.5 (which is a value lower than the selected significance threshold of 0.8065), there is a tendency for chemical distance and DN distance to both be small, but for the majority of the cases (i.e., those with a MES distance below the 0.8065 threshold) the chemical similarity does not correlate at all with the MES distance.

In addition, also the opposite effect happens; that is, drugs with very small chemical distance have very high MES distance. Therefore, the two measures are not correlated, although there are a few cases where very small chemical distance corresponds with small MES distance.

Moreover, we measured the tendency of our network communities to group together drugs that are similar by the chemical point of view. To this end we considered the

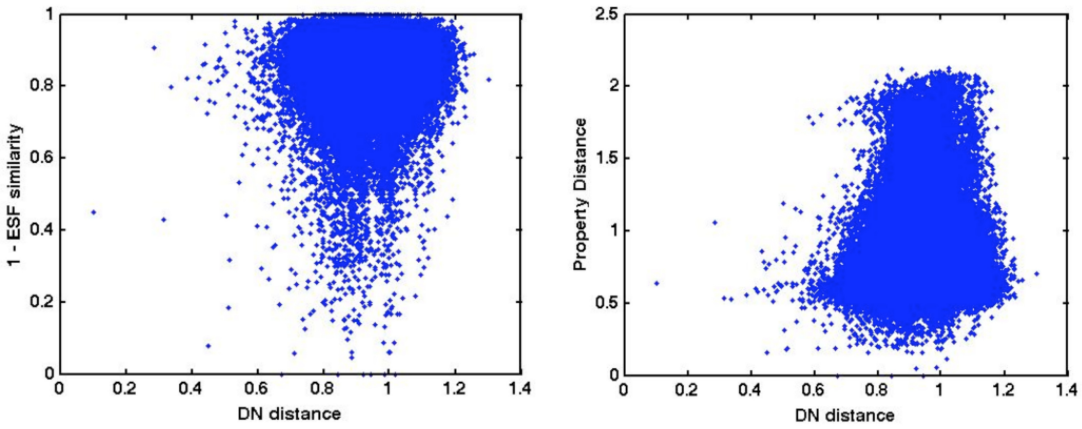

Figure 4.20: Correlation with chemical similarity - Correlation plots between the MES distance and two measures quantifying the chemical similarities among the cMap compounds: ESF similarity (left plot) and property distance (rigth plot)).

empirical pdf of the pairwise ESF similarity, computed on the whole set of drugs with a SMILES. Then we considered the pairwise ESF similarity computed only between drugs in the same community.
Finally, we tested the null hypotesis that this set (similarities in the same community) was sampled from the first distribution. The obtained p-value was equal to 1, meaning that the composition of our communities is not significantly influenced by chemical similarities.
In Figure 4.21, we can observe that the empirical pdf of the pairwise ESF similarity computed between drugs in the same community (red line) almost perfectly overlaps the pdf of the pairwise ESF computed on the whole set of drugs with a SMILES (blue line).
Very similar results were obtained by considering the Property Distance measures (in the same figure).

Finally, we computed the average ESF similarity for all the communities that are enriched for a given MoA (appendices B.1, B.2 and B.3) and containing at least two drugs with an available SMILES. Results are contained in appendix D and show that just for few communities the average ESF is significantly greater than the average value (i.e. 1.7).
In the table of appendix D the first column contains the community identifiers, the

second one contains the community enrichment (Literature evidence/ATC-code/Direct Target Gene), the third one contains the fraction of drugs in the community for which chemical descriptors were available, and the last column contains the average ESF similarity for the community.

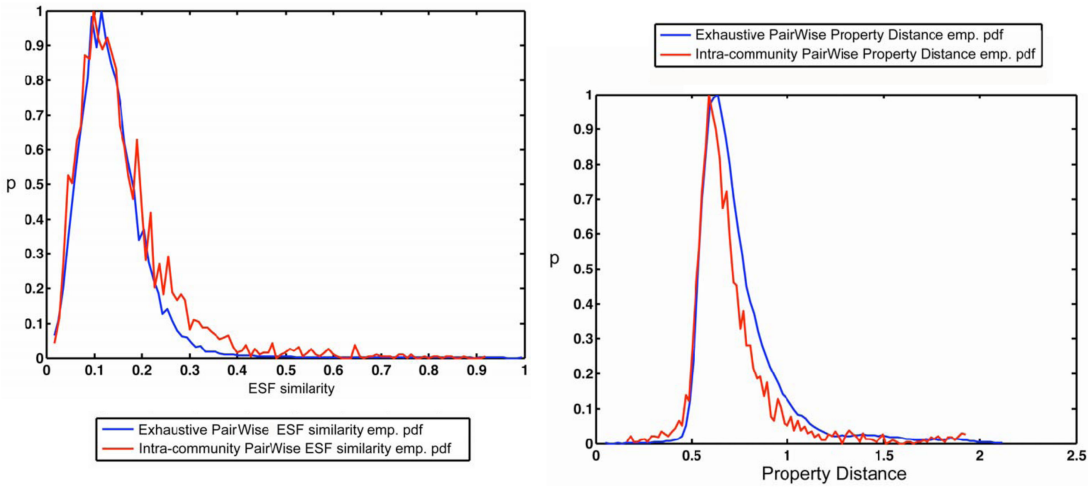

Figure 4.21: Influence of chemical similarity on drug distance - Empirical pdfs of the ESF similarity (left plot) and the property distance (right plot) computed between each couple of cMap compounds (blue curves) and only between couples of compounds clustered in the same community (red curves).

4.7.6 Gene Ontology Fuzzy-Enrichment analysis of the communities

We developed a Fuzzy-Logic (155) based approach to identify a common set of genes that was consistently up-, or, down-regulated in the PRLs of the compounds in the same community. We thus associated significant GO terms statistically over-represented in these sets to 57 drug communities. For most of these communities, the associated fuzzy-enriched GO terms were strictly linked to the mode of action of the composing drugs. We considered this a very interesting result, which is objective, completely unsupervised and obtained with a general method.

We describe the method that we developed to perform this analysis in this section and we report also on some of the obtained results.

The Gene Ontology (7) is a hierarchical vocabulary describing the roles of genes and proteins in eucaryotes organisms, which is accessible by the web. It is composed by

ontology of terms (GO:terms) defining, in a proper standard way, gene properties and it covers three different domains: biological processes, molecular functions and cellular components. In order to obtain information about a given gene it is sufficient to analyze the associated GO:terms. Analysis of GO:term enrichments (78) has become a widely recognized way to quickly gain insights about the biological condition represented by a set of genes and many bioinformatics tools have been developed for this aim. As an example, pretend to have a set of genes describing a condition of interest (i.e. genes that are differentially expressed after a drug treatment or that are selectively expressed in a disease). If a GO:term is surprisingly recurrent among those associated to the genes in that set, then we can conclude that the selection criterion through which the set of genes was composed (i.e. the condition of interests) is biologically and semantically connected to the property described by the GO:term.

In our case the condition of interest is the transcriptional response to the drugs contained in the same community. We computed fuzzy-sets (154) of differentially expressed genes for each community and we performed a GO enrichment analysis on them.

Differently from a normal set (also said a "crisp set") elements can belong to a fuzzy-set with different degree of confidence, which is defined by a membership function with values in $[0, 1]$.

Formally a fuzzy-set is defined as a couple (A, m), where A is a set and $m : A \to [0, 1]$ is said "membership function".

For each $x \in A$, $m(x)$ is called the degree of membership of x in A. If $m(x) = 0$ then x is called not included in the fuzzy-set (A, m) whereas x is called fully included if $m(x) = 1$. Finally, x is called fuzzy member if $0 < m(x) < 1$.

The set $\{x \in A | m(x) > 0\}$, composed by all the fuzzy members of A, is called the "support" of (A, m) and the set $\{x \in A | m(x) = 1\}$, composed by the elements that are fully included in A, is called its "kernel".

In our method, we first collected the PRLs for each of the drug in a community then we gave all the PRLs in input them to the algorithm GO:Fuzzy-Enrichment-Analysis. The algorithm computed two fuzzy-sets of differentially expressed genes (up-regulated and down-regulated, respectively) and two "fuzzy-intersections" (defined in the following) of genes (up-regulated and down- regulated, respectively) by heuristically determining an optimal threshold value for the fuzzy-sets membership functions.

The output of the algorithm was composed by the computed fuzzy-intersections, the

optimal value of the membership function threshold and the GO:terms enriched in the fuzzy-intersections.

Note that to compute p-values in the GO:enrichment analysis we used the considerations and formulas introduced in subsection 4.7.1.

Let us consider a community C and the set of PRLs associated to the drugs contained in it $\{d_1, d_2, \ldots, d_n\}$. First of all, for each d_i we selected the top-ranked 2,000 genes by composing with them the set Up_i and the bottom-ranked 2,000 ones by composing with them the set $Down_i$.

We then compute the following fuzzy-sets: (U_{UP}, m_{UP}) and (U_{DOWN}, m_{DOWN}) where:

$$U_{UP} = \bigcup_{i=1}^{n} Up_i \tag{4.11}$$

and

$$U_{DOWN} = \bigcup_{i=1}^{n} Down_i. \tag{4.12}$$

For each gene x in U_{UP} (respectively U_{DOWN}) the membership function $m_{UP} : U_{UP} \to [0, 1]$ (respectively $m_{DOWN} : U_{DOWN} \to [0, 1]$) was defined as follows:

$$m_{UP}(x) = |\{i : x \in Up_i\}|/n \tag{4.13}$$

respectively

$$m_{DOWN}(x) = |\{i : x \in Down_i\}|/n. \tag{4.14}$$

Obviously, the following relations always hold:

$$\forall x \in U_{UP} : 1/n \leqslant m_{UP}(x) \leqslant 1 \tag{4.15}$$

and

$$\forall x \in U_{DOWN} : 1/n \leqslant m_{DOWN}(x) \leqslant 1. \tag{4.16}$$

This means that the support set of (U_{UP}, m_{UP}) (respectively, (U_{DOWN}, m_{DOWN}) is equal to U_{UP} (respectively, U_{DOWN}) since it contains the genes that belong to at least one Up_i (respectively, $Down_i$).

The kernel set of (U_{UP}, m_{UP}) (respectively, (U_{DOWN}, m_{DOWN}) contains the genes that are in the top-ranked (bottom-ranked, respectively) 2,000 positions in the PRL of all the drugs of the community under consideration.

Now, fixing threshold level k for the membership function values, such that $1/n \leqslant k \leqslant 1$,

we define the fuzzy-intersection of up-regulated genes, in the drug community C, with membership k, as follows:

$$F_{UP}(C,k) = \{x \in U_{UP} | m_{UP}(x) \geqslant k\}. \tag{4.17}$$

Note that $F_{UP}(C,1)$ is equal to the traditional intersection (called also "crisp" intersection) $\bigcap_{i=1}^{n} Up_i$. In the same way, the fuzzy-intersectio of down-regulated genes, in the community C, with membership k, is defined as

$$F_{DOWN}(C,k) = \{x \in U_{DOWN} | m_{DOWN}(x) \geqslant k\}. \tag{4.18}$$

Also in this case $F_{DOWN}(C,1)$ is equal to the crisp intersection $\bigcap_{i=1}^{n} Down_i$.
In order to describe our algorithm through a pseudo-code, let us introduce the following additional notation: we denote with $GO_{UP}(k)$ the set of GO:terms that are statistically over-represented (i.e. enriched) among those associated to the genes in $F_{UP}(C,k)$ and with $GO_{DOWN}(k)$ the set of GO:terms over-represented among those associated to the genes in $F_{DOWN}(C,k)$.
In our algorithm we used a heuristic approach to fix an appropriate value of k, in order to maximize it and the cardinalities of these two sets as well.
The input of the algorithm is the drug community C. The output is composed by the two sets $GO_{UP}(k)$ and $GO_{DOWN}(k)$ together with the compute value of k.
The pseudo-code is provided in algorithm 4.

When the computation begins, k is set to 1 [line 1]. The cardinality of the two fuzzy-intersections is set to zero [line 2] and the set of fuzzy enriched GO:terms is set to the empty set [line 3].
Then a cycle iterates until at least one of the two fuzzy-sets contains more than 2,000 genes [line 4]. In each of the iterations, the fuzzy intersections and the sets of fuzzy enriched GO:terms are recomputed [lines 5 and 6], according to the current value of k. If the total number of fuzzy-enriched GO:terms does decrease [line 7], then the current sets of fuzzy enriched GO:terms together with the value of k, which has been computed in the previous iteration, are given in output and the algorithm ends [line 8]. Otherwise (if the total number of fuzzy enriched GO:terms does not decrease) [line 9], then the variables are updated [lines 10 and 11] and the membership threshold value is decreased [line 12]. The remaining code [lines 13 to 15] is executed if the total number of fuzzy

Algorithm 4 GO:Fuzzy-Enrichment-Analysis

$\{k, GO_{UP}(k), GO_{DOWN}(k)\}$ = **GO:Fuzzy-Enrichment-Analysis**(C)

input: C, a drug community

output: k, the optimal threshold value for the membership functions

$GO_{UP}(k)$, the GO:terms enriched in the up-regulated fuzzy-intersection of C

$GO_{DOWN}(k)$, the GO:terms enriched in the down-regulated fuzzy-intersection of C

1. $k \leftarrow 1$
2. $nUp \leftarrow nDown \leftarrow 0$
3. $totalGO = \emptyset$
4. while $nUp < 2,000$ and $nDown < 2,000$
5. compute $F_{UP}(C,k)$ and $F_{DOWN}(C,k)$
6. compute $GO_{UP}(k)$ and $GO_{DOWN}(k)$
7. if $|GO_{UP}(k)| + |GO_{DOWN}(k)| < |totalGO|$
8. then return $\{k + 1/n, GO_{UP}(k + 1/n), GO_{DOWN}(k + 1/n)\}$
9. else
10. $totalGO \leftarrow GO_{UP}(k) \cup GO_{DOWN}(k)$
11. $nUp \leftarrow |GO_{UP}(k)|$, $Down \leftarrow |GO_{DOWN}(k)|$
12. $k \leftarrow k - 1/n$
13. endif
14. endwhile
15. return $\{k + 1/n, GO_{UP}(k + 1/n), GO_{DOWN}(k + 1/n)\}$

enriched GO:terms never decreases, while decreasing k, and the total number of genes in the two fuzzy intersection is greater than 2,000.

The results that we obtained with the GO:Fuzzy-Enrichment-Analysis on the 106 communities of our DN are provided in the SDD [**SDD2-GOFuzzyEnrichements.xls**]. We report in the rest of this section on some of the most representative ones.
For community n. 28 (enriched for Hsp90 inhibitors) our algorithm gave in output an optimal threshold level for the membership functions (i.e. k) equal to 0.8 (meaning that the computed fuzzy-intersections were composed by genes that were significantly differentially expressed when treating with 4 among 5 drugs in this cluster). The fuzzy-intersection of up-regulated genes contained 209 genes while the down- regulated one 236. Fuzzy enriched GO:terms for this cluster of drugs are shown in table 4.10.

Community n. 28: Hsp90 inhibitors

Up-regulated fuzzy-intersection			
Biological Process Enriched GO:Terms	***p*-value**	**Molecular Function Enriched GO:Terms**	***p*-value**
response to unfolded protein	4.42×10^{-29}	Unfolded protein binding	1.63×10^{-13}
response to protein stimulus	4.42×10^{-29}	TPR domain binding	4.53×10^{-08}
protein folding	2.26×10^{-24}	heat shock protein binding	3.30×10^{-05}
response to biotic stimulus	3.73×10^{-13}	nitric-oxide synthase regulator activity	3.07×10^{-04}
response to chemical stimulus	3.59×10^{-05}	chaperone binding	4.18×10^{-04}
regulation of nitrogen compound metabolic process	5.05×10^{-05}	macrolide binding	2.37×10^{-02}
positive regulation of nitrogen compound metabolic process	7.59×10^{-05}	FK506 binding	2.37×10^{-02}
protein refolding	1.13×10^{-04}		
regulation of nitric oxide biosynthetic process	2.36×10^{-04}		
response to stress	3.43×10^{-04}		
positive regulation of nitric oxide biosynthetic process	1.80×10^{-03}		
protein metabolic process	4.02×10^{-02}		
Down-regulated fuzzy-intersection			
Biological Process Enriched GO:Terms	***p*-value**	**Molecular Function Enriched GO:Terms**	***p*-value**
tRNA processing	2.57×10^{-06}		
ribosome biogenesis and assembly	6.44×10^{-04}		

Table 4.10: GO:Fuzzy enrichment analysis results for Community n. 28

Hsp90 is a chaperone protein responsible for the correct folding, stabilization, and function of multiple proteins (141). Inhibition of Hsp90 increases the amount of unfolded client proteins in the cellular environment. This leads to a stress condition for the cell, resulting in the activation of a proper response via the activation of several pathways, as those involved in the ubiquitin-proteasome degradation system. Looking at the GO:terms enriched in the up-regulated fuzzy intersection (table 4.10), for this

community, highlights the response induced in the cell by these compounds.
The genes contained in the fuzzy-intersections of this cluster were differentially expressed when treating with the analyzed drugs in the following proportions: 98% alvespimycin, 95% geldanamycin, 89% monorden, 84% tanespimycin, 45% fulvestrant, meaning that the drugs in this cluster are represented by the computed fuzzy- intersections with different levels of specificity.
Interestingly, this specificity is approximately proportional to the relation occurring between these drugs and the MoA characterizing this cluster (i.e. inhibition of the Hsp90 protein). This because alvespimycin, and geldanamycin directly bind the Hsp90 protein inhibiting its cytosolic chaperone function and they are very similar by a chemical point of view; monorden is a wider Hsp90 inhibitor with effects on Topo I and II also; fulvestrant binds the estrogen receptor, dissociates Hsp90 and triggers its intracellular degradation so it indirectly inhibits the chaperone functionality in the cell.
For community n. 14 (CDK2 inhibitors and Topo inhibitors) our algorithm gave in output an optimal threshold level for the membership functions equal to 0.8 (i.e. fuzzy-intersections were composed by genes that were significantly differentially expressed when treating with 12 among 15 drugs in this community). The fuzzy-intersection of up-regulated genes contained 8 genes (and no enriched GO terms) while the down-regulated one contained 75 genes and the enriched GO:terms reported in table 4.11.

Community n. 14: CDK2 and Topo inhibitors [L01D, L01DB, L01, L] [GSK3B, TOP2A]

Down-regulated fuzzy-intersection	
Biological Process Enriched GO:Terms	***p*-value**
cell division	2.07×10^{-03}
Mitosis	2.04×10^{-02}
M phase of mitotic cell cycle	2.28×10^{-02}
M phase cell cycle phase	3.68×10^{-02}
cell cycle phase	4.75×10^{-02}

Table 4.11: GO:Fuzzy enrichment analysis results for Community n. 14

Cyclin-Dependant Kinases (CDKs) are key regulators of cell cycle progression. CDK2 and CDK4 are responsible for phosphorylation of the Retinoblastoma (RB) protein, causing the release and activation of the E2F transcription factors, resulting in transcription of genes involved in cell cycle progression (94).
Also in this case, it is possible to hypothesize the effects of this community of drugs on the transcription just by looking to the GO:fuzzy-enriched terms reported in table 4.11.

Additionally, as in the previous case, the proportions of genes of the fuzzy-intersections that are differentially expressed when treating with each of the drugs reflect the specificity of the described drug effect ($> 80\%$ for the CDK2 and Topo inhibitors, $< 40\%$ for the other drugs).
For the community n. 63 (Sodium/Potassium membrane pump blocker), our algorithm gave in output an optimal threshold level for the membership functions equal to 0.91 (i.e. fuzzy-intersections were composed by genes that were significantly differentially expressed when treating cells with 10 among 11 drugs in this community). The fuzzy-intersection of up-regulated genes contained 40 genes (and the fuzzy enriched GO:terms reported in table 4.12) while the down-regulated one contained 39 genes but no enriched GO:terms.

Community n. 63: Na^+/K^+-ATPase membrane pump inhibitors [C01A, C01AA, C01, C], [ATP1A1]

Up-regulated fuzzy-intersection	
Biological Process Enriched GO:Terms	***p*-value**
biogenic amine biosynthetic process	3.60×10^{-04}
amino acid derivative biosynthetic process	6.11×10^{-04}
biogenic amine metabolic process	1.40×10^{-02}
regulation of epidermis development	2.23×10^{-02}
ethanolamine metabolic process	2.81×10^{-02}
phosphatidylethanolamine biosynthetic process	2.81×10^{-02}
ethanolamine biosynthetic process	2.81×10^{-02}
amino acid derivative metabolic process	3.90×10^{-02}

Table 4.12: GO:Fuzzy enrichment analysis results for Community n. 63

The obtained fuzzy-enriched GO:terms are linked to a specific effect of cardiac glycosides (the majority the drugs in this community): the enhancement of some heart phosphatides (i.e. ethanolamine and phopshatidylethanolamine) activity (98). The majority of the genes contained in the computed fuzzy-intersections were differentially expressed in all of the PRL of the cardiac glycosides in this cluster ($> 90\%$).
Finally, for community n. 43 (estrogen and estrogen inhibitors), our algorithm gave in output an optimal threshold level for the membership functions equal to 0.44 (meaning that genes in the fuzzy-intersections were differentially expressed when treating cells with 4 among 9 drugs in this cluster). The fuzzy intersection of up-regulated genes contained 425 genes while the down-regulated one contained 335 genes. The fuzzy enriched GO:terms (reported in table 4.13) hilighted the interactions between estrogens and the Golgi apparatus (50, 116) and the down-regulation of genes involved in metabolic processes of organic compounds interacting with estrogens (cobalamin, por-

phyrin and others). So, also in this case the fuzzy enriched GO:terms were strictly linked to the MoA of the analyzed drugs.

Community n. 43: Estrogens and estrogen inhibitors

Up-regulated fuzzy-intersection

Biological Process Enriched GO:Terms	***p*-value**	**Molecular Function Enriched GO:Terms**	***p*-value**
ER to Golgi vesicle-mediated transport	4.36×10^{-05}	cadmium ion binding	3.54×10^{-06}
protein transport	7.38×10^{-05}	protein disulfide isomerase activity	9.61×10^{-06}
establishment of protein localization	7.38×10^{-05}	intramolecular oxidoreductase activity & transposing S-S bonds"	9.61×10^{-06}
Golgi vesicle transport	8.41×10^{-05}	intramolecular oxidoreductase activity & interconverting keto- and enol-groups"	1.63×10^{-05}
protein folding	1.30×10^{-04}	isomerase activity	$1.80 \times 10^{-}04$
protein localization	2.17×10^{-03}	neutral amino acid transmembrane transporter activity	3.59×10^{-02}
cell redox homeostasis	5.36×10^{-03}	cystine:glutamate antiporter activity	3.92×10^{-02}
macromolecule localization	1.03×10^{-02}	unfolded protein binding	4.37×10^{-02}
Golgi organization	1.77×10^{-02}		

Down-regulated fuzzy-intersection

Biological Process Enriched GO:Terms	***p*-value**	**Molecular Function Enriched GO:Terms**	***p*-value**
biotin biosynthetic process	2.78×10^{-07}	adenosylmethionine-8-amino-7-oxononanoate transaminase activity	1.02×10^{-07}
cobalamin metabolic process	2.78×10^{-07}	biotin synthase activity	1.02×10^{-07}
cobalamin biosynthetic process	2.78×10^{-07}	dethiobiotin synthase activity	1.02×10^{-07}
biotin metabolic process	3.79×10^{-06}	8-amino-7-oxononanoate synthase activity	1.02×10^{-07}
vitamin biosynthetic process	3.95×10^{-04}	cobyrinic acid a & c-diamide synthase activity	1.02×10^{-07}
porphyrin biosynthetic process	3.72×10^{-03}	sulfurtransferase activity	1.39×10^{-06}
tetrapyrrole biosynthetic process	3.72×10^{-03}	glutaminase activity	3.88×10^{-06}
water-soluble vitamin biosynthetic process	3.72×10^{-03}	cyclo-ligase activity	9.54×10^{-06}
		2 iron & 2 sulfur cluster binding	8.42×10^{-05}
		4 iron & 4 sulfur cluster binding	4.95×10^{-04}
		transaminase activity	2.10×10^{-03}
		iron-sulfur cluster binding	6.35×10^{-03}
		metal cluster binding	6.35×10^{-03}
		transferase activity & transferring nitrogenous groups	2.14×10^{-02}
		transferase activity & transferring acyl groups other than amino-acyl groups	4.12×10^{-02}

Table 4.13: GO:Fuzzy enrichment analysis results for Community n. 43

4.8 Goals of a drug network with modular and characterized topology

The N-TRAP algorithm performs a proper pruning of the edges of the network in order to make it modular. Most of the identified communities are enriched for a given mode

of action, others have been characterized through a GO enrichment analysis, as shown in the previous section. This means that, if one would be able to integrate a compound under investigation into the DN, then he could make hypothesis on its MoA by looking at the communities to which it is connected and to their characterizations. This is the leading idea of the classification algorithm that is described in the next chapter and represents a striking and original improvement to existing methods (79, 80).
In conclusion we show that the N-TRAP pruning is able to keep the "right" connections and to eliminate the "wrong" ones. In other words, we measure how the tendency of drugs with a similar MoA of being linked together changes after removing edges with the N-TRAP algorithm. We ranked the edges of both the pruned network and the original one, in ascending order (according to the associated distance value), and we computed the percentage of edges that connect drugs sharing an ATC prefix of length 3 following the assessment methodology explained in subsection 4.4.2 and obtaining the results shown in Figure 4.22. As we can see in Figure 4.22, in the pruned network the edges connecting similar drugs tend to be kept and the performances of the network improve.

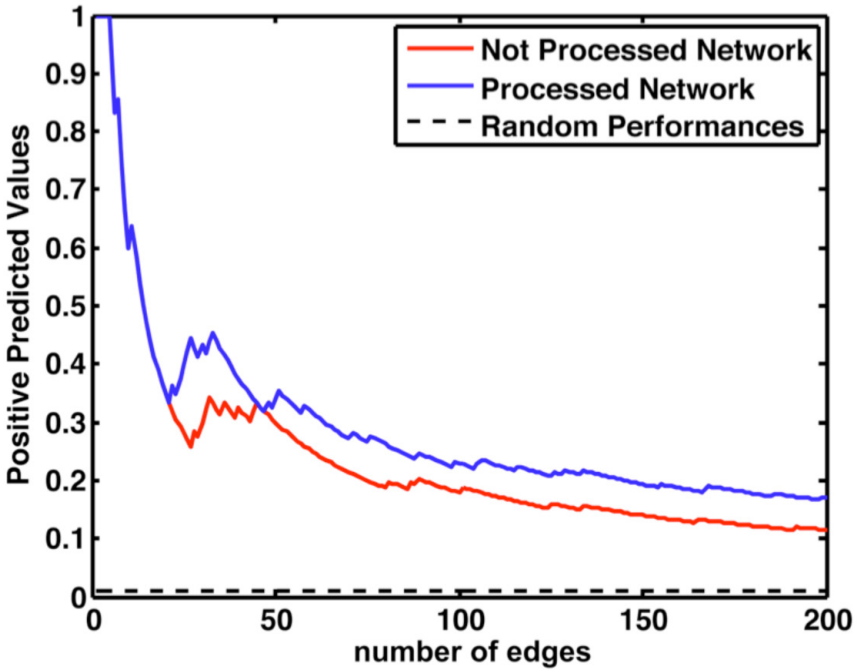

Figure 4.22: Modularity and performances - Impact of the community identication on the performances. PPV curve obtained by sorting the edges of the whole significant according to their weight (in decreasing order) in the whole significant network (red) and the processed network (i.e. the modular network built with the N-TRAP algorithm) (blue).

5

MANTRA: Mode of Action by NeTwork Analysis

5.1 Introduction

In this Chapter we describe a web-tool that we developed in order to allow users to explore the DN and query it for classification of previously undescribed compounds.
We named this tool Mode of Action by Network Analysis (MANTRA) and made it available online for the scientific community at the following web-site `http://mantra.tigem.it`).
With this online software unreported similarities among drugs can be easily explored by users via an interactive analysis of the DN topology. Hence safe and FDA approved drugs can be easily proposed for a repurposing and novel drugs can be integrated in the DN revealing their MoA
When a new drug is added to the network, by drawing the significant connections among this drug and the other drugs in the different communities, it is possible to check to which drug communities the drug is similar to, and which are the closest communities in terms of distance.
We imagined this to be an interactive approach in which the user looks at his drugs and the communities to which the drug is connected to in order to make hypothesis on the drug MoA. Thus realizing a semi-automated approach. Although the tool can be used also in a full-automatic way thanks to a definition of a "drug-to-community distance" that we conceived. This is a score to rank the communities closest to the drug. The

score between a drug and a community is detailed in Section 5.2. The smaller the score, the closer the total distance between the drug and the community.
In the next chapter we show that thanks to this score we can automatically predict the closest community without the need to look at the network, although we believe this to be useful.
In Section 5.3 of this chapter a detailed description of the classification algorithm is provided while in the Section 5.4 we report on the development of MANTRA illustrating how it can be used for exploring the drug network and its modules, to search for candidates for drug repositioning and to classify novel drugs basing on gene expression data only.

5.2 Drug-to-Community Distance

We defined the Drug-to-Community distance as follows: Let x be the testing drug and C a network community containing a subset C_x of, at least, two drugs that are connected to x through significant edges (i.e., through edges corresponding to distances that are below the significance threshold). Then we define the distance between x and C as

$$\bar{D}(x, C) = \sqrt[|C_x|]{\prod_{y \in C_x} MES(x, y)/|C_x|}. \tag{5.1}$$

So, the distance between the testing drug x and the network community C is given by the ratio between the geometric mean of the significant distances between drugs in C and x and the cardinality of this set of distances. If $|C_x| < 2$, then we assume that the distance between C and x is equal to ∞.

5.3 Classification Algorithm

A description of the classification procedure implemented in our online tool is provided in algorithm 5.

The algorithm takes in input the DN N, the set containing the PRLs for each drug in the DN, the set containing the node communities of the DN C_N and a set X of ranked lists containing all the MPI (Affymetrix®HG-U133a chip). The MPI in these lists are sorted according to their differential expression values in MicroArray experiments in which the drug that is going to be classified has been tested on. In these

Algorithm 5 Drug-Classification

$\{N^*, \eta, \chi\}$ = **Drug-Classification**(X, N, PRL_N, C_N)

input: X, a set of permutations of the m MPI of the specific microarray platform;
$N = (V_N, E_N, \omega_N)$, the DN;
PRL_N, the set of the PRLs of the drugs in the DN,
such that $\forall y \in V_N : PRL_y$ is the PRL of the corresponding drug;
C_N, the set of the communities in N.

output: $N^* = (V_N^*, E_N^*, \omega_N^*)$, the updated DN;
η = a sorted list of drug nodes;
χ = a sorted list of drug communitites.

1. $x \leftarrow$ the novel drug nodes
2. $V_N \leftarrow V_N \cup \{x\}$
3. $PRL_x \leftarrow \text{KRUBOR}(X)$
4. for each $y \in V_N \setminus \{x\}$
5. $d_{x,y} = d(PRL_x, PRL_y)$
6. if $d_{x,y} < th$
7. $E_N \leftarrow E_N \cup \{(x, y)\}$
8. $\omega(x, y) = d_{x,y}$
9. endif
10. endfor
11. $\eta = \{y \in V_N | (x, y) \in E_N\}$
12. turn η into a list by sorting its nodes y basing on the values $\omega((x, y))$ in ascending order
13. $\chi = \{C \in C_N | \bar{D}(x, C) < \infty\}$ in η
14. turn χ into a list by sorting its communities C basing on the values $\bar{D}(x, C)$ in ascending order
15. if $|\eta| = 0$
16. $V_N^* \leftarrow V_N^* \setminus \{x\}$
17. endif
18. $V_N^* \leftarrow V_N$
19. $E_N^* \leftarrow E_N$
20. $\omega_N^* \leftarrow \omega_N$
21. $N^* \leftarrow (V_N^*, E_N^*, \omega_N^*)$
22. return $\{N^*, \eta, \chi\}$

experiments differential expression values should be computed with respect to the untreated hybridizations of the corresponding cell lines following the cMap composition scheme.

The output of the algorithm is the updated version of the DN, N^* in which the node corresponding to the testing drug has been added together with the significant connections; the drug nodes lying in the neighborhood that the novel drug node x has in N^*, sorted in ascending order according to the weights of the edges connecting them to x (i.e. the distances for the corresponding drugs from the novel one); the drug communities with a finite community-distance from the novel drug, sorted according to this distance in increasing order.

When the computation begins a new node associated to the testing drug is created [line 1] and temporarily added to the set of the drug nodes in the DN [line 2].

The PRL for the novel drug is computed with the KRUBOR algorithm starting from the ranked lists of MPI in input [line 3].

For each node different from x in the DN [line 4] the distance between the corresponding drug and the novel one is computed [line 5]. Note that at this stage the d function can compute the distance via one of the two functions (aes or mes, respectively for the AES drug distance and the MES drug distance) introduced in Section 4.3. We defined these two functions by the mathematical point of view only but it is clear that their implementation computes the drug optimal signatures by retrieving them from the drug PRLs (as explained in Section 4.3) so the input they should receive is represented by a pair of PRLs only (and this is the input of the function d as well).

At this point, if the distance between x and the drug under consideration is less than the statistically significant threshold th (0.8065 or 0.8327, for MES drug distance and AES drug distance, respectively) [line 6] than x is connected to that drug (i.e. the edge (x, y) is added to the set of the edges of the DN) [line 7]; the weight of this new edge is equal to the distance between the two drugs [line 8].

When the cycle ends, the set of drug nodes with a connection with x (i.e. the η, the x-neighborhood built with the code in [line 11]) is sorted according to the connection weights in ascending order (i.e. the set η is turned out in a ranked list according to these weights) [line 12].

The code in [line 13] builds the set χ of communities with a finite distance from x (i.e. the communities containing at least two drugs connected to x).

This set is sorted according the community distances from x [line 14].
With the code in the [line 15 to 17] x is removed from the set of the drug nodes if it is not connected to any other drug nodes. Note that if it is the case than the set χ (that now is a ranked list) is for sure empty.
The rest of the code [lines 18 to 22] updates the DN and terminates the computation returning the proper output.

5.4 MANTRA web-tool

The MANTRA web-tool implements the algorithm described in the previous section and the method that we published in (66).
Briefly, the web-tool allows users to visually explore the DN we built among the cMap compounds in a user-friendly environment (see Figure 5.1) providing, for each of the drugs, information about biochemical interactions, therapeutic indications, known MoA, pharmacology and targeted proteins.

These data come from the public available databases DrugBank (148) and ChemBank (127) and are displayed in a pop-up window when the user moves the mouse pointer over a drug node (yellow box in Figure 5.1).
By exploring the DN with MANTRA, users can identify unexpected similarities between drugs acting on different direct intra-cellular targets and search for "repositionable" drugs (i.e. drugs for which novel and previously unrecognized therapeutic applications could be hypothesized).
Finally, the user can choose to integrate its own drug in the DN and to classify its MoA by uploading up to 6 GEPs.
MANTRA has been implemented as a Java Applet by customizing the Applet version of Medusa (60), a front end to the STRING protein interaction database (144), which can be also used as a general graph visualization tool. The algorithm for classifying novel drugs is embedded in the applet and the whole system has been implemented in a Tomcat server. Use of MANTRA web-tool is free to academic, government and non-profit users for non-commercial use only. With MANTRA it is possible to predict established and FDA approved drugs that could be repositioned by finding previously unreported MoAs.
To this aim, it is sufficient to search for "interesting outliers" in the drug communities.

Figure 5.1: MANTRA - According to the Hindu tradition a "Mantra" is a sound, syllable, word or group of words capable of creating "spiritual transformation". Our MANTRA is capable of creating a transformation: it turns the information hidden in a microarray experiment in a meaningful landscape of drugs providing an "enlightened" view of them. MANTRA is a web-tool for the analysis of novel drugs and for the "repositioning" of known and FDA approved drugs by the assignment of previously unrecognized putative therapeutic applications (hence the Hippocrates Symbol in this logo).
The MANTRA analysis is based on a novel similarity measure among the cellular responses elicited by a huge set of compounds in human (hence the central man figure in the logo). These responses are summarized by genome-wide gene expression profiles, hence the DNA aura surrounding the man.

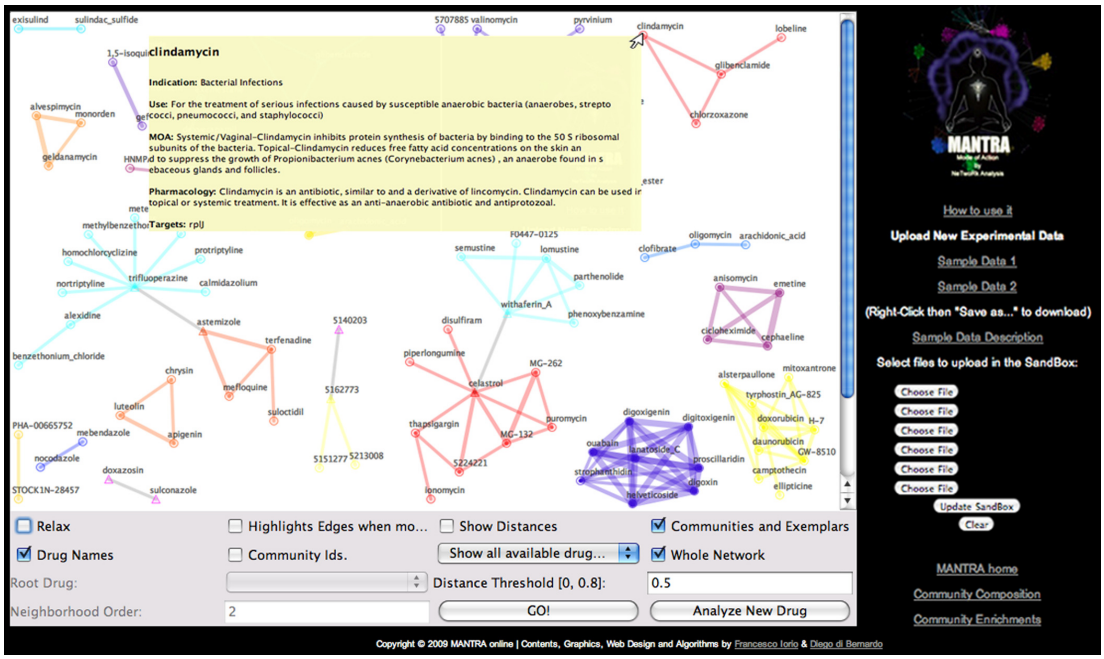

Figure 5.2: MANTRA interface - The graphical user interface of MANTRA

In fact, if a given drug x is contained in a community C that is enriched for a given MoA and that MoA has never been linked to x before then it could be interesting to test whether x shares that MoA too.

If this is the case, and x is a safe and approved drug, then it can be proposed for a "repurposing" to treat conditions in which the novel discovered MoA has a therapeutic effect.

Alternatively, users can choose a drug with a desired MoA and searching in its neighborhood for other safe and approved drugs that were not previously linked to that MoA. Examples are reported in (66) and (67).

In order to integrate a novel drug into the DN and to classify its MoA, it is sufficient to submit to MANTRA up to 6 text files containing all the 22,283 MPI of the Affymetrix®HG-U133a microarray platform (human), sorted according to their differential expression (in decreasing order) following treatments with the drug on a sufficiently heterogeneous (i.e. with different genetic background) set of human cell lines.

Once the drug is integrated in the DN, according to the algorithm we introuduced, users can make hypotheses on its MoA by studying drug nodes and communities in its

neighborhood. Examples are reported in (66).
In conclusion, MANTRA is a web-tool for the drug mode of action discovery and drug repositioning based on a network of consensual transcriptional responses to drugs.
It allows the classification of novel drugs by simply uploading gene expression profiles following treatment and the topology of its network has an incredible potential (easily exploitable by users) in finding novel applications for a huge number of approved drugs.

6

Experimental validation of MANTRA predictions using known and novel chemotherapeutic agents

6.1 Introduction

In this chapter we show how we experimentally assessed the ability of our method in classifying novel compounds by using gene expression data only. In principle, we probed our method with two different classes of compounds, containing both known and novel drugs, observing that they were correctly integrated in the DN and connected to the right drug communities.

By studying these MoAs we were able to make hypotheses on the effects of the novel drugs and to discover, for the first time, a strong transcriptional similarities between chemotherapeutic agents acting on two distinct molecular targets.

A description of the tested compounds, the experimental design, and the microarray data that we produced are provided in the following section while the classification results are reported in Section 6.3, together with the discussion about the new experimental data that we generated in order to further investigate the transcriptional similarity that we discovered.

The rationale behind the predictions of our method is explored in Section 6.4 while in

Section 6.5 classification performances are numerically evaluated and compared with those of other existing methods.
In 6.6 the impact of our rank merging strategy on the classification performances is assessed and we show that a set of microarray data coming from a sufficiently heterogeneous set of cell lines treated with a drug provides a general summary of the drug effect.
A discussion about the attained goals and the limitations of our method is provided in the last section of the chapter.

6.2 A "blind" classification test

We assessed the ability of the DN to predict the MoA of anticancer compounds whose gene expression profiles were not included in the original cMap dataset as well as the ability in classifying established drugs contained in the cMap.
As summarized in Figure 4.1 (C), we measured expression profiles derived from different cell lines treated with anticancer compounds still being studied, developed at Nerviano Medical Science (NMS) and reference drugs already present in the cMap dataset.
Nine compounds were considered for a total amount of 39 microarray hybridizations.
We computed a PRL for each of the tested compounds, and their distances from the 1,309 drugs in the cMap dataset. We then integrated the compounds in the DN by connecting them to the other drugs, if their distance was below the significant threshold.
Additionally, we computed a drug-to-community distance, which quantifies how close the tested compound is to each of the communities.
This was a "blind" classification test because the NMS provided us the microarray data without specifying the compound used nor the treated cell lines.

6.2.1 Experimental Setting and protocols

Drugs tested were chosen among well-known compounds, already present in the cMap dataset, and new generation compounds. They included Hsp90 inhibitors tanespimycin (132), NVP-AUY922 (35), NMS-E973 (42); Topo inhibitors SN-38 (75) and doxorubicin (6); CDKs inhibitors flavopiridol (128), PHA-848125 (14), PHA-690509, and PHA-793887 (13).
The rationale behind the choice of the test compounds and the cell lines to treat is explained in the following sections.

Human ovarian cancer cell line (A2780) cells were treated with flavopiridol (0.3 μM), PHA-848125 (1 μM), PHA-690509 (3 μM), and PHA-793887 (1 μM), whereas Human breast cancer cell line (MCF7) was treated with PHA-848125 (8.5 μM), PHA-793887 (6.0 μM), tanespimycin (0.5 μM), NVP-AUY922 (0.07 μM), NMS-E973 (2 μM), SN-38 (0.165 μM), and doxorubicin (1.5 μM).

Additional data were collected by treating Human glioma cell line (U251) and Human glioblastoma cell line (SF539) with PHA-848125 (3 μM), to assess the impact of the merging of data coming from different settings on the classification performances as explained in Section 6.6.

A2780 and MCF7 from European Collection of Cell Cultures were seeded in T-75 tissue culture flasks (Corning), 25.000 cells/cm2 in RPMI medium 1640 (Gibco), pH 7.4, 10% FBS (EUROCLONE Australia-USDA approved), 2 mM L-Glutamine (Gibco), 1× penicillinstreptomycin (Gibco), and maintained in 5% CO_2 at 37 C with 96% relative humidity.

After 24 hours, cells were treated with different compounds at a dose equal to 5× the IC_{50} for 6 hours and collected using Qiagen RNeasy Lysis Buffer (Qiagen cat no. 79216). Total RNA was extracted using Qiagen RNeasy kit (Qiagen cat. no. 74104), starting from total cell lysates.

The RNA was purified following manufacturer instructions. During the process, any genomic DNA contaminations were removed by DNAse treatment. Quantity and purity of the extracted RNA were assessed by spectrophotometric evaluation of light absorbance at 260 and 280 nm; after extraction, RNA was stored at -80C. Biotin-labeled, fragmented cRNA probes were prepared starting from 1.5 μg of total RNA per replicate sample, using the One-Cycle Target Labeling and Control Reagents (Affymetrix®) according to the protocols included in the Affymetrix GeneChip Expression Analysis Technical Manual (`http://www.affymetrix.com`).

Samples were hybridized onto Affymetrix GeneChip®Human Genome U133 Plus 2.0 Arrays and processed as per manufacturers instructions using GeneChip®Hybridization, Wash, and Stain Kit components (Affymetrix). Scanned images were first inspected for quality control (QC) using a variety of built-in QC tools from the Bioconductor package [`http://www.bioconductor.org`] of R, the open source environment for statistical analysis.

Feature intensity values from scanned arrays were normalized and reduced to expression summaries using MAS5 implemented in the R statistical environment.
A ranked list of genes was obtained for each compound treatment by sorting the microarray probe-set identifiers according to the differential expression values with respect to the untreated hybridization. These ranked lists composed the starting point of our classification test.
Data are available at Gene Expression Omnibus database (GEO), `www.ncbi.nlm.nih.gov/geo` (accession no. GSE18552).

6.2.2 Hsp90 Inhibitors

Heat Shock Protein 90 (Hsp90) is one of the most abundantly expressed molecular chaperone in the cell (27). Chaperones are proteins responsible for the folding or unfolding and the assembly or disassembly of other macromolecular structures, but not occurring in these structures when they are performing their normal biological functions.
Hsp90 is a member of the heat shock protein family which are up-regulated when the cell is exposed to elevated temperatures or in response to other kind of stress (32).
Heat shock proteins, as a class, are among the most highly expressed cellular proteins across all species. As their name implies, heat shock proteins protect cells when stressed by elevated temperatures. They account for 1 - 2% of total protein in unstressed cells, increasing to 4 - 6% when cells are heated.
Hsp90 is one of the most common of the heat related proteins. The protein is named "HSP" for obvious reasons whereas the "90" comes from the fact that it weighs roughly 90 kiloDaltons. A 90 kDa size protein is considered a fairly large for a non-fibrous protein.
Hsp90 is part of the cell's powerful network of chaperones to fight the deleterious consequences of protein unfolding caused by nonphysiological conditions. In the absence of stress, however, Hsp90 is an obligate component of fundamental cellular processes such as hormone signaling and cell cycle control. In this context, several key regulatory proteins, such as steroid receptors, cell cycle kinases, and p53, have been identified as substrates of Hsp90. Recently, Hsp90 was shown to be the unique target for geldanamycin, a potent new anti-tumor drug that blocks cell proliferation.
Interestingly, under physiological conditions, Hsp90 seems to perform its chaperone function in a complex with a set of partner proteins, suggesting that the Hsp90 complex

is a multi-chaperone machine specialized in guiding the maturation of conformationally labile proteins. Clinical studies have demonstrated that disruption of many client proteins chaperoned by Hsp90 is achievable and associated with significant growth inhibition, both in vitro and in tumor xenografts (49). The regulation of key signaling molecules of the cell by the Hsp90 machinery is a stimulating new concept emerging from these studies, and Hsp90 has become a promising new drug target (125) for several therapeutic applications (111).

Additionally, Hsp90 is capable of suppressing protein aggregation, solubilizing protein aggregates and targeting protein clients for degradation. Induction of the heat-shock response by small molecules may facilitate the clearance of toxic aggregates responsible for neurodegenerative diseases and, consequently, Hsp90 has emerged more recently as a target for the treatment of neurodegenerative diseases that result from misfolded and aggregated proteins (136).

Our DN contains a community enriched for Hsp90 inhibitors (n. 28) containing, sorted according to their distance from the community exemplar, alvespimycin (the exemplar), geldanamycin (MES distance from the exemplar = 0.28), monorden (0.35), tanespimycin (0.52), fulvestrant (0.77). 4 of these drug are known Hsp90 inhibitors (alvespimycin, geldanamycin, monorden and tanespimycin) while fulvestrant is a selective estrogen receptor down-regulator. Note that the distances from the exemplar reflects the specificity of these compounds in inhibiting the Hsp90 (see subsection 4.7.6). The sub-community containing only the 4 Hsp90 inhibitors is a fully connected component with edges corresponding to MES distances less than 0.59, which is a very significant value and at least monorden has a chemical structure that significantly differs from those of the other three. Moreover, as mentioned in subsection 4.7.6, even if it does not inhibit Hsp90 directly, fulvestrant binds the estrogen receptor, dissociates Hsp90 and triggers its intracellular degradation so it indirectly inhibits the chaperone functionality in the cell, which translates in a down-stream effect on the transcription that is similar to that elicited by the Hsp90 inhibitors.

In conclusion, this class of compounds elicit a well defined transcriptional response that should be easily recognizable so we chose to perform a first classification test by using known and novel Hsp90 inhibitors.

6.2.3 Cyclin-Dependent kinase (CDK) 2 Inhibitors

As second set of testing compounds we chose four Cyclin-Dependent kinase (CDK) 2 inhibitors.

As introduced in Section 2.2 and Section 4.7.6, CDKs are key regulators of cell cycle progression: CDK2 and CDK4 are responsible for phosphorylation of the Retinoblastoma (RB) protein, causing activation of the E2F transcription factor and transcription of genes involved in G1/S transition and initiation of DNA replication (94).

From human tumors and mouse models, it is clear that misregulation of G1 CDK activity by either over-expression of cyclins or loss of CDK inhibitory proteins almost invariably leads to hyperproliferative defects and eventually to tumor development. In particular, activation of the CDK4/6 pathway seems to dramatically decrease the requirements that allow cells to enter the cell cycle and participate in tumor development (95).

Similarly, activation of CDK2 and perhaps CDK1, through over-expression of E-, A-, or B-type cyclins or $p27^{Kip1}$/$p21^{Cip1}$ inactivation, seems to force the entry into S phase and commit cells to progress through the mitotic cell cycle. These data have been obtained from multiple research efforts including molecular analysis of human tumors, molecular and cellular biology, and the characterization of knock-out and knock-in mice. These data have stimulated the design and development of small-molecule CDK inhibitors as new drugs for cancer therapy. In the last few years, a plethora of CDK inhibitors have been analyzed in vitro, in mouse models, or in clinical trials (129, 131, 133).

CDK2 activity is deregulated in human cancer primarily through over-expression of cyclin E and cyclin A and inactivation of the CDK inhibitor $p27^{Kip1}$ (95).

Given the relevance of these alterations in human cancer, CDK2 has been considered an important target for cancer therapy. Numerous CDK2 inhibitors have been described and their crystallographic structures either in complex with CDK2 or CDK2-cyclin A have been broadly analyzed(145) including flavopiridol, currently in Phase III clinical trials (128). Many of these inhibitors also inhibit CDK1 and in certain cases other kinases such as CDK5, CDK7, CDK9, Glycogen synthase kinase (GSK) 3β, Mitogen-activated protein kinase (MAPK), and Extracellular signal-regulated kinases (ERK) (8), complicating their biochemical profiling. Therefore, further studies need to be accomplished to depict whether the anti-tumor effects are mainly because of the CDK2

inhibition or the synergism with other kinases.

Genetic evidence has shown that CDK2-cyclin E activity is not essential for cell progression through the cell cycle and may be compensated by another kinases, possibly CDK4, CDK6, or CDK1. In addition, CDK2 inhibition by RNA interference fails to arrest proliferation of osteosarcoma cells and pRB-negative cervical cancer cells (143). These results suggest that CDK2 may not be a good target for inhibition by small molecules intended to treat cancer.

This finding, along the fact that most efficient CDK2 inhibitors also inhibit other kinases, have shifted attention back toward CDK4 (92, 93) or CDK1 as the primary cell-cycle target for cancer drug discovery. Actual efforts are directed to obtain more specific CDK2 inhibitors.

The cMap includes a limited number of molecules whose MoA is associated with the inhibition of CDK2. These compounds are clustered in two different communities (n. 14 and n. 32) that are part of the same rich-club. This means that the effect on the transcription of this class of compounds is wider with respect to that of the Hsp90 inhibitors and it could be harder to "recover" them in the DN in a classification test. Therefore, we sought to probe the DN, through our classification algorithm, with the transcriptional profile of flavopiridol, as well as those of PHA-690509, PHA-793887, and PHA-848125, three ATP-competitive CDK inhibitors developed at NMS, with different selectivity profiles within the CDK family, which have completed Phase I clinical trials (13, 14).

Table 6.1 reports a selectivity profile of the four CDK inhibitors. In this table we reported for each tested compound and for a set of kinases the average Half maximal inhibitory concentration (IC_{50}) concentration. The IC_{50} is a measure of the effectiveness of a compound in inhibiting biological or biochemical function. This quantitative measure indicates how much of a particular drug or other substance (inhibitor) is needed to inhibit a given biological process (or component of a process, i.e. an enzyme, cell, cell receptor or microorganism) by half. In other words, it is the half maximal (50%) inhibitory concentration of a substance (50% IC, or IC_{50}).

Enzyme	**PHA-793887** Average IC50 (μM)	**PHA-848125** Average IC50 (μM)	**PHA-690509** Average IC50 (μM)	**Flavopiridol** Average IC50 (μM)
CDK1	0.06	0.398	0.16	0.034
CDK2	0.008	0.045	0.031	0.04
CDK4	0.062	0.16	>10	0.09
CDK5	0.005	0.265	0.09	0.102
CDK7	0.01	0.15	nt	0.754
CDK9	0.138	1.112	0.141	0.025
GSK3	0.079	>10	1.9	0.971
TRKA	>10	0.053	nt	nt

Table 6.1: Selectivity profiles of the tested CDK inhibitors.

6.3 Classification results

Figure 6.1 shows the position of the tested compounds in the DN whereas the 10 closest neighboring drugs and communities in the DN for each of the tested compounds, according to the Drug-Classification algorithm, are listed in table 6.2 and table 6.3, respectively. The whole neighborhoods are listed in the appendix F.

Particularly, the closest community to the three tested Hsp90 inhibitors is n. 28, composed by the Hsp90 inhibitors present in cMap, as well as the anti-estrogen drug fulvestrant, known to bind the estrogen receptor, dissociate HSP90, and trigger its intracellular degradation. The second closest community common to all the three compounds (n. 40) is enriched for proteasome inhibitors, ubiquitin proteasome system modulators (celastrol, MG-132, MG-262, thapsigargin, disulfiram, mometasone), and protein synthesis inhibitors (puromycin and primaquine). Another interesting community is n.104, which contains the proteasome/NF-kB inhibitors withaferin A, parthenolide, thiostrepton, and etacrynic acid. Weaker edges connect two of the three tested compounds to community n. 63, consisting of Na^+/K^+-ATPase membrane pump inhibitors. This proximity might be explained by the fact that inhibition of Na^+/K^+-ATPase by cardiac glycosides has been shown to affect NF-kB signaling (150). As introduced in Section 4.7.6, fuzzy GO:term enrichment analysis showed that genes involved in the response to unfolded proteins are up-regulated in community n. 28 and community n. 104, whereas community n. 40 is enriched for GO:terms relative to endoplasmatic reticulum overload and stress.
Therefore, the DN approach correctly predicted, with multiple evidences, the MoA of the tested compounds by identifying them as Hsp90 inhibitors.
All four CDK inhibitors were positioned in the DN in close vicinity to community n.

NMS-tanespimycin		**NMS-E973**		**NVP-AUY922**	
MES	*Compound*	*MAS*	*Compound*	*MAS*	*Compound*
0.4360	alvespimycin*	0.4436	alvespimycin*	0.6084	alvespimycin*
0.4913	geldanamycin*	0.4891	geldanamycin*	0.6391	monorden*
0.5176	monorden*	0.5294	monorden*	0.7123	geldanamycin*
0.6315	tanespimycin*	0.6568	tanespimycin*	0.7506	puromycin
0.6533	puromycin	0.6723	puromycin	0.7608	tanespimycin*
0.7178	trifluoperazine	0.7308	trifluoperazine	0.7756	gefitinib
0.7542	parthenolide	0.7638	disulfiram		
0.7561	thiostrepton	0.7842	methylbenzethonium chloride		
0.7608	withaferin A	0.7850	parthenolide		
0.7724	disulfiram	0.7903	lanatoside C		
NMS-doxorubicin		**SN38**		**flavopiridol**	
MES	*Compound*	*MAS*	*Compound*	*MAS*	*Compound*
0.5587	daunorubicin*	0.3215	irinotecan*	0.4540	alsterpaullone*
0.6495	GW-8510	0.5641	camptothecin*	0.4857	GW-8510*
0.6536	hycanthone	0.6158	apigenin*	0.5374	apigenin*
0.6555	ellipticine*	0.6251	phenoxybenzamine	0.5534	0175029-0000
0.6689	irinotecan	0.6363	etoposide	0.5789	daunorubicin
0.6900	camptothecin	0.6596	luteolin*	0.5966	doxorubicin
0.6921	etoposide*	0.6675	tyrphostin AG 825	0.5976	camptothecin
0.6926	mycophenolic acid	0.6877	daunorubicin	0.6196	ellipticine
0.6996	phenoxybenzamine	0.6882	thioguanosine	0.6270	H-7*
0.7175	doxorubicin*	0.6903	hycanthone	0.6301	tyrphostin AG 825
PHA-690509		**PHA-793887**		**PHA-848125**	
MES	*Compound*	*MAS*	*Compound*	*MAS*	*Compound*
0.3838	GW-8510*	0.4715	0175029-0000	0.6212	0175029-0000
0.4613	doxorubicin	0.4846	GW-8510*	0.6352	apigenin*
0.4794	alsterpaullone*	0.5145	alsterpaullone*	0.6504	harmine*
0.5001	H-7*	0.5370	apigenin*	0.6672	thioguanosine
0.5593	daunorubicin	0.5694	daunorubicin	0.6711	GW-8510*
0.5873	camptothecin	0.5976	doxorubicin	0.6746	luteolin*
0.5956	ellipticine	0.6014	ellipticine	0.6795	daunorubicin
0.6048	mitoxantrone	0.6353	tyrphostin AG 825	0.6828	irinotecan
0.6144	tyrphostin AG 825	0.6582	luteolin*	0.6877	camptothecin
0.6274	fisetin*	0.6607	camptothecin	0.6886	piperlongumine

*True Positives: drugs sharing the mode of action with the testing one

Table 6.2: First ten neighbors of the tested compounds in the drug network

NMS-tanespimycin		**NMS-E973**		**NVP-AUY922**	
D	*Community*	*D*	*Community*	*D*	*Community*
0.1285	28*	0.1310	28*	0.1285	28*
0.1296	104	0.1996	63	0.1296	40
0.1329	63	0.2481	40		
0.1863	40	0.2566	100		
0.2567	100	0.2640	104		
NMS-doxorubicin		**SN38**		**flavopiridol**	
D	*Community*	*D*	*Community*	*D*	*Community*
0.0978	14*	0.0888	32*	0.0480	14*
0.1458	3	0.1174	14	0.0603	90
0.1900	16	0.1434	3	0.0625	32*
0.2374	32	0.2581	89	0.0954	89
0.3955	40	0.3798	75	0.1929	52
				0.1995	85
				0.2527	40
				0.2564	63
				0.3781	104
				0.3874	61
PHA-690509		**PHA-793887**		**PHA-848125**	
D	*Community*	*D*	*Community*	*D*	*Community*
0.0300	90	0.0527	14*	0.0721	14*
0.0464	14*	0.0916	32*	0.0845	32*
0.0585	32*	0.0947	63	0.0927	63
0.0639	89	0.1927	3	0.2550	89
0.1283	85	0.3830	104	0.2590	104
0.1299	52			0.3762	69
0.1931	74			0.3763	100
0.1933	61			0.3847	3
0.2561	13				
0.3837	40				

*True positives: communities enriched for the mode of action of the testing drug.

Table 6.3: Closest ten drug network communities in the neighborhood of the tested compounds

14, which includes a mixture of CDKs and Topo inhibitors, altogether accounting for about 80% of this community (Figure 6.1 (c)).
The other closest community was n. 32, also containing several CDK and/or Topo inhibitors, such as the CDK2 inhibitors chrysin, harmine, harman, and harmol, the CDK2/Topo II inhibitor apigenin, the CDK2/Topo I inhibitor luteolin, and the Topo I inhibitors irinotecan and skimmianine.
The intermixing of CDK and Topo inhibitors in communities n. 14 and n. 32, as well as the identification of several Topo inhibitors as the closest neighbors of the CDK inhibitors, implies a similarity of their effects at the transcriptional level, despite their different intracellular protein targets. To confirm this transcriptional similarity, we probed the DN with in-house generated transcriptional profiles following treatment with two known Topo inhibitors as detailed in the following section.

6.3.1 Topoisomerase Inhibitors

Topoisomerase inhibitors are agents designed to interfere with the action of Topoisomerase (Topo) enzymes (Topo I and II), which are enzymes that control the changes in DNA structure by catalyzing the breaking and rejoining of the phosphodiester backbone of DNA strands during the normal cell cycle (see Section 2.2.2).
In recent years, topoisomerases have become popular targets for cancer chemotherapy treatments. It is thought that topoisomerase inhibitors block the ligation step of the cell cycle, generating single and double stranded breaks that harm the integrity of the genome. Introduction of these breaks subsequently lead to apoptosis and cell death. Topoisomerase inhibitors can also function as antibacterial agents (104) (quinolones have this function (41)).
We decided to probe our classification method with SN-38, the active metabolite of irinotecan (a prototypic Topo I inhibitor) and with doxorubicin (a prototypic Topo II inhibitor). The tested doxorubicin will be denoted in the following text with NMS-doxorubicin in order to avoid ambiguities with the counterpart in the cMap dataset
As shown in Figure 6.1 (b) and Tables 6.2 and 6.3, SN-38 and NMS-doxorubicin were positioned, as expected, close to communities n. 14 and n. 32, containing their counterparts in the database. Hence, also in this case the compounds were correctly classified. Additionally, the 10 closest neighbors for both compounds included a mixture of CDKs and Topo I or II inhibitors. Suggesting that these two class of compounds elicit a

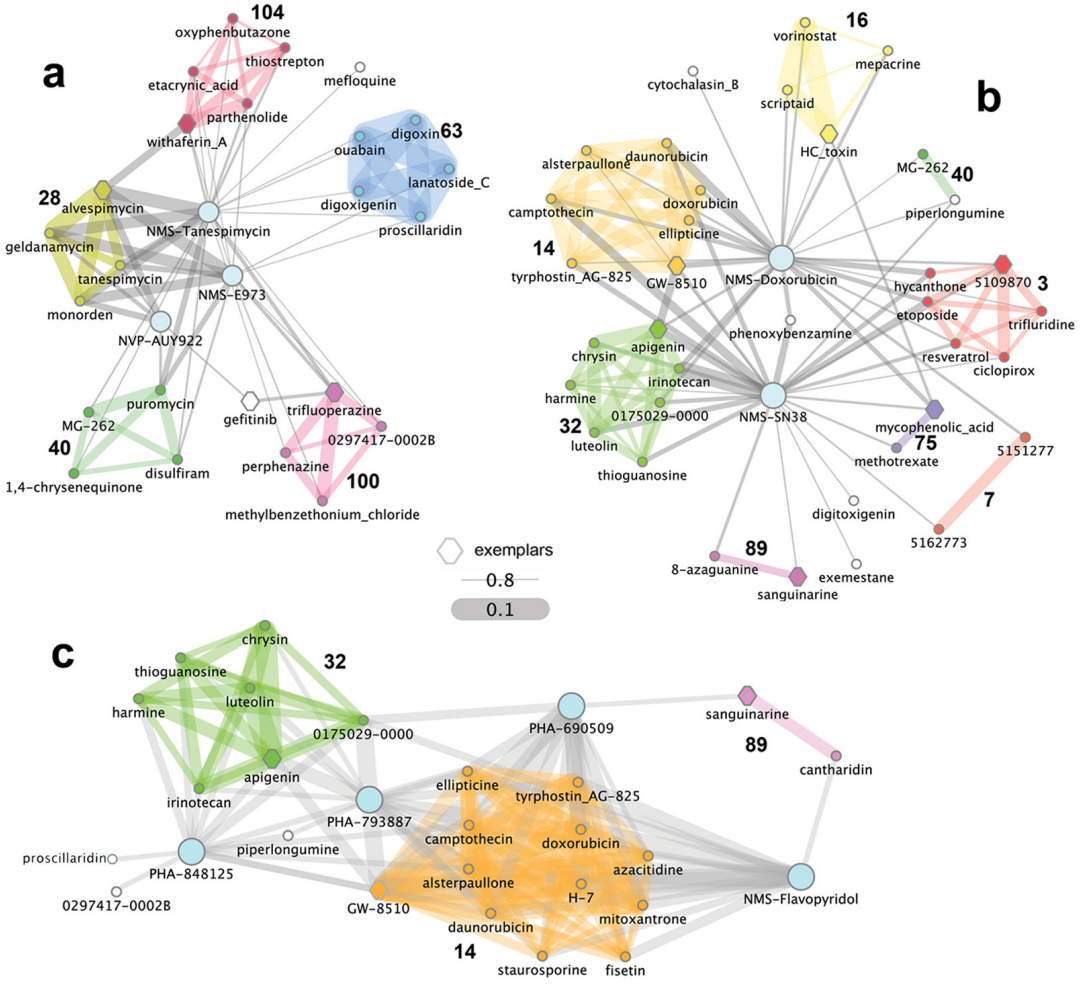

Figure 6.1: Classification results - Subnetworks connected to the tested compounds (cyan nodes) once they have been integrated in the DN. For clarity we included only compounds whose distances from the tested compounds were less than 0.8 (A and C) or 0.72 (B). Edge thickness is inversely proportional to the distance between the drugs; edge and node colors indicate communities. Hexagonal-shaped nodes represent community exemplars. (a) HSP90 inhibitors; (b) Topo inhibitors; (c) CDK inhibitors.

similar transcriptional response even if acting on different intracellular targets.
We studied the cause of this similarity discovering a strong rationale behind this classification outcomes. Conclusions of these studies are summarized in the following section.

6.4 MANTRA highlights previously unreported similarities

Whereas most CDK inhibitors act by competitively binding to the ATP pocket of kinases, and given that Topo II uses ATP hydrolysis for its function, we verified that there was no direct biochemical inhibition of CDKs by SN-38 and doxorubicin, and that flavopiridol was not able to interfere with the ATPase activity of Topo II (Figure 6.2). Another possible way to induce functional inhibition of CDKs is through the

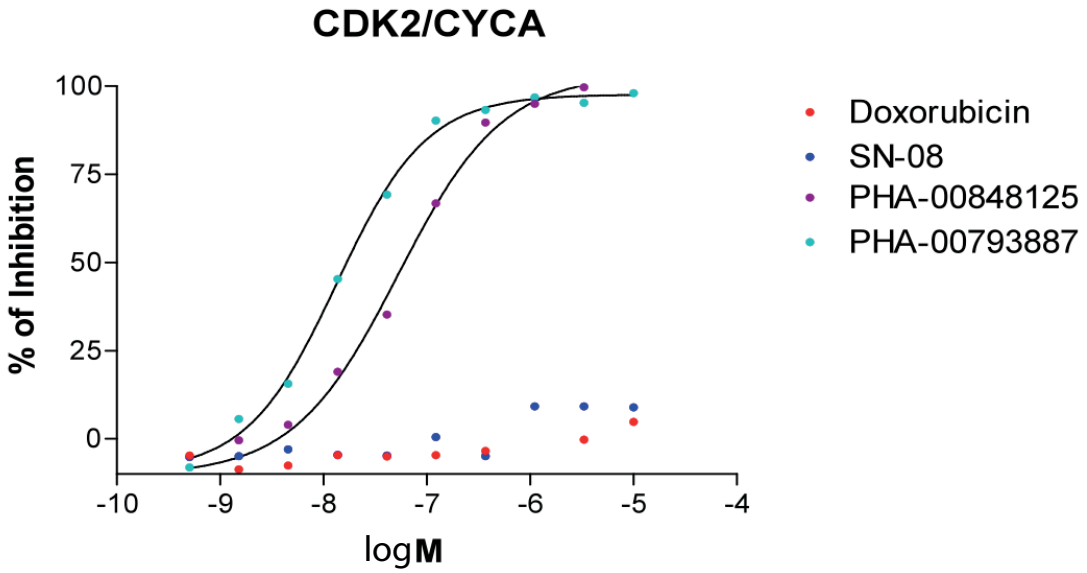

Figure 6.2: Inhibition of CDKs by doxorubicin and SN-38 - Inhibition of CDKs by doxorubicin and SN-38: Biochemical Essay. Inhibition of CDK2/cyclinA complex (CDK2/CYCA) activity by two topoisomerase inhibitors (doxorubicin and SN-38) and two CDK inhibitors (PHA-00848125 and PHA-00793887) developed at NMS, used as controls, tested in a biochemical assay. Compound concentration is on the x axes, expressed in Moles (in logarithmic values), whereas percentage of inhibition is on the y axes. Different colors represent different compounds. No biochemical inhibition of CDKs by SN-38 and doxorubicin could be observed.

induction of their universal inhibitor p21. Indeed, DNA damage induced by Topo inhibitors causes p21 up-regulation activating both p53-dependent and independent apoptosis (1, 88).
We hypothesized that p21 inhibition of the endogenous CDKs, and in particular CDK2, elicited an effect on the RB-mediated transcription and might thus explain the similarity at the gene expression level (as summarized in Figure 6.3).

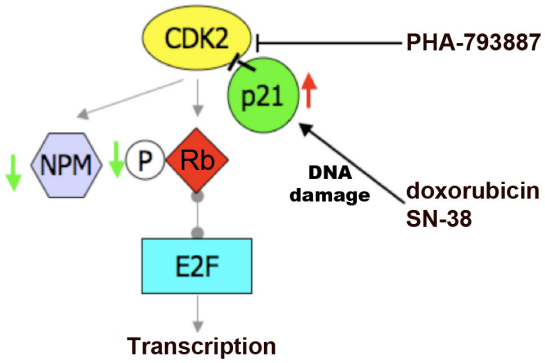

Figure 6.3: Down-stream effects of CDK2 and Topo inhibitors - Summary of the common down-stream effect of CDK2 and Topo inhibitors on E2F mediated transcription, as elucidated by MANTRA

To confirm this, we treated MCF7 cells for 6 hours with PHA-793887 (used as reference CDK inhibitor), doxorubicin, or SN-38, at the same doses previously used, and analyzed the protein cell lysates by Western Blot (WB).
Following treatment with both Topo inhibitors, we observed induction of p21 resulting in inhibition of CDK2, as measured by decreased phosphorylation of the CDK2 substrates, RB and nucleophosmin (Figure 6.4).

Although we cannot exclude that induction of other genes, such as p27, in addition to p21, may also contribute to this effect.
It was recently proposed that camptothecin (a natural analog of irinotecan) treatment would directly inhibit CDK9 activity by disrupting its complex with the activating cyclin T partner, inducing a functional effect similar to that observed after ATP-competitive inhibition of CDK9 by flavopiridol (4). To test this hypothesis, we analyzed the protein cell lysates used in the previous experiment for inhibition of RNA polymerase II as measured by decreased phosphorylation of its carboxyterminal domain and diminished Myeloid Cell Leukemia sequence (MCL)1 levels.

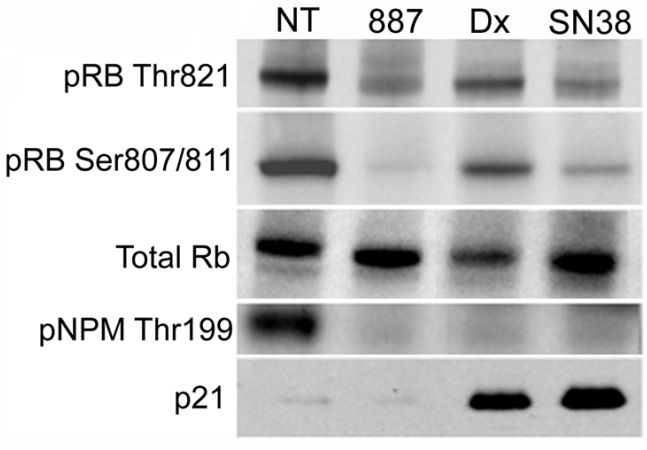

Figure 6.4: Effects on p21 and CDK2 substrates - Western blot of total MCF7 cell lysates following 6 hours of treatment with doxorubicin (Dx), SN-38 (SN38), and the CDK inhibitor PHA-793887 (887). Induction of p21 coupled to decreased phosphorylation of the CDK2 substrates Retinoblastoma (Rb) and Nucleophosmin (NPM) by the Topo inhibitors Dx and SN38 is observed.

After treatment with PHA-793887 (CDK7 inhibition IC_{50} 10 nM; CDK9 inhibition IC_{50} 140 nM), a decrease of phosphoserine 5, and to a minor extent also of phosphoserine 2, was detected and resulted in diminished levels of MCL1. However, no effect on RNA Polymerase II phosphorylation or MCL1 levels was observed after treatment with the Topo inhibitors, suggesting that this pathway was not affected (as shown in figure 6.5).

Taken together, these data prove that the transcriptional effects observed with the Topo I and Topo II inhibitors are due to an (indirect) inhibition of CDK2 (and possibly other CDKs such as CDK4) mediated by p21 induction, highlighting a previously unreported similarity that provides a strong rationale for the DN classification results.

6.5 Classification Performance assessment and comparison with other tools

In order to compare the classification results achievable with method with those provided by the cMap online tool (79, 80), we computed a signature of differentially expressed genes in a "traditional way" (i.e., list of significant genes according to t-test corrected with false discovery rate False Discovery Rate (FDR)) for each microarray

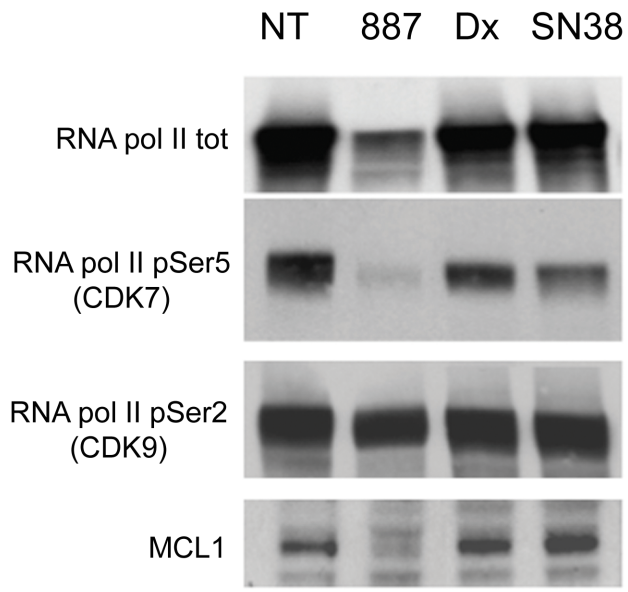

Figure 6.5: Effects on RNA pol II - Western blots following treatments with CDK2 and Topo inhibitors. Decreased phosphorylation of RNA polymerase II (RNA Pol II) on Serine 5 (RNA Pol II pSer 5, a CDK7 substrate) and to a lesser extent on Serine 2 (RNA Pol II pSer2, a CDK9 substrate) by PHA-00793887 (887), a CDK inhibitor developed at Nerviano Medical Science, coupled to loss of MCL1 protein. Minor effects on Serine 5 are observed also with doxorubicin (Dx) and SN-38 (SN38), but they do not affect MCL1 levels.

experiment, as follows.

Scanned microarray images were first inspected for Quality control (QC) using a variety of built-in QC tools from the Bioconductor (47) package of R, the open source environment for statistical analysis.

Feature intensity values from scanned arrays were normalized and reduced to expression summaries using the Robust Multiarray Algorithm (RMA) and normalized by the quantiles method (68, 149).

To assess differential expression, we used a moderated t-test together with a FDR correction of the p-value (135, 147). Thus, the list of differentially expressed genes was generated using a FDR ≤ 0.05 together with an absolute fold-change threshold of 2 (i.e., $|\log_2(\text{fold change}) \geqslant 1|$) and composed the signatures that were used to query the cMap online tool.

The experiments were relative to four groups of related drugs (see table 6.4).

Tested Compound Set	**True Positives [Hsp90 inhibitors]**
NMS-tanespimycin NVP-AUY922 NMS-E973	geldanamycin, alvespimycin, rifabutin, monorden tanespimycin
Tested Compound Set	**True Positives [Topo I inhibitors]**
SN-38	irinotecan, camptothecin, luteolin, kaempferol, suranin-sodium, vidarabine, proscillaridin, apigenin, cinoxacin, skiammianine
Tested Compound Set	**True Positives [Topo II inhibitors]**
NMS-doxorubicin	daunorubicin, podophyllotoxin, mitoxantrone, genistein, ellipticine, oxilinc-acid, etoposide, doxorubicin, nalidic-acid, ofloxacin, enoxacin, novobiocin, ciprofoxacin, apigenin
Tested Compound Set	**True Positives [CDK2 inhibitors]**
PHA-848125 PHA-690509 flavopiridol PHA-793887	alsterpaullone staurosporine, GW-8510, H-7, apigenin, harmine, harmol, luteolin, chrysin, fisetin, sanguinarine, thyrpostin AG 825

Table 6.4: Compounds analyzed with the cMap online tool and corresponding sets of true positives.

We used the computed signatures to query the cMap online tool. We then compared the results obtained with our approach with those provided by the cMap online tool by means of ROC analysis.

The cMap tool provided in output a list of drugs connected to each of the input signatures. In these lists, we filtered out the drugs that were predicted to be negatively connected to the input signature, and we considered each of the remaining drugs as true positives if they belonged to at least one of four different reference golden standard sets (sets of true positives in table 6.4).

These reference golden standard sets included both the counterpart of the tested drugs (if they were present in the cMap) and drugs known to have the same MoA as the tested drugs (respectively, Hsp90 inhibitors, Topo I inhibitors, Topo II inhibitors, and CDK2 inhibitors) according to either Drugbank (148) or ChemBank (127).

All of the signatures obtained with the traditional approach, which have been used to query the cMap online tool, are available at `http://mantra.tigem.it` and in the SDD (in a unique compressed folder, containing each signature in the cMap .grp format).

File **`SDD3-Signatures-for-cMap.zip`**.

All the classification results obtained with the cMap online tool when queried with these signatures are shown in the appendix E.

The result assessment shows that our classification method performed comparably and, in many cases, better than the cMap classic online tool. The percentage of cases in which the first neighbor of a tested compound in the DN is a true positive is equal to 89% for the AES distance and 77% for the MES distance. This value raises to 100% if we consider the case in which there is at least a true positive among the first two neighbors of each tested compound, for both the distances (as depicted in table 6.5).

	N = 1	**N = 2**	**N = 3**	**N = 5**	**N = 10**	**N = 15**
cMap classic query system	56%	88%	88%	88%	88%	88%
AES distance	89%	100%	100%	100%	100%	100%
MES distance	77%	100%	100%	100%	100%	100%

Table 6.5: Distance performances - Percentage of tested compounds with at least one correct neighbor among the first n.

The ROC analysis results are provided in table 6.6.

6.5 Classification Performance assessment and comparison with other tools

a) cMap Classic Query System

		PPV when considering the first n neighbors							
Tested Compound	*Treated Cell Line*	$n = 1$	$n = 2$	$n = 3$	$n = 5$	$n = 10$	$n = 15$	$n = 20$	$n = 50$
PHA-848125	U251	0.00	0.50	0.33	0.40	0.30	0.20	0.20	0.10
flavopiridol	A2780	1.00	0.50	0.33	0.60	0.60	0.40	0.35	0.14
PHA-848125	A2780	0.00	0.00	0.00	0.40	0.20	0.13	0.10	0.08
PHA-690509	A2780	0.00	0.50	0.67	0.60	0.30	0.20	0.15	0.12
PHA-793887	A2780	1.00	0.50	0.67	0.60	0.40	0.33	0.25	0.12
PHA-793887	MCF7	1.00	0.50	0.33	0.60	0.30	0.20	0.15	0.12
NMS-tanespimicyn	MCF7	1.00	1.00	1.00	0.80	0.40	0.40	0.30	0.12
NMS-E973	MCF7	1.00	1.00	1.00	0.80	0.50	0.40	0.30	0.12
NVP-AUY922	MCF7	0.00	0.50	0.67	0.60	0.40	0.27	0.20	0.08
SN38	MCF7	1.00	0.50	0.33	0.20	0.20	0.13	0.10	0.08
NMS-doxorubicin	MCF7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04
Average Value:		0.54	0.50	0.42	0.50	0.32	0.24	0.19	0.10

b) cMap Classic Query System (Best Profile per Drug)

		PPV when considering the first n neighbors							
Tested Compound	*Treated Cell Line*	$n = 1$	$n = 2$	$n = 3$	$n = 5$	$n = 10$	$n = 15$	$n = 20$	$n = 50$
PHA-848125	U251	0.00	0.50	0.33	0.40	0.30	0.20	0.20	0.10
flavopiridol	A2780	1.00	0.50	0.33	0.60	0.60	0.40	0.35	0.14
PHA-690509	A2780	0.00	0.50	0.67	0.60	0.30	0.20	0.15	0.12
PHA-793887	A2780	1.00	0.50	0.67	0.60	0.40	0.33	0.25	0.12
NMS-tanespimicyn	MCF7	1.00	1.00	1.00	0.80	0.40	0.40	0.30	0.12
NMS-E973	MCF7	1.00	1.00	1.00	0.80	0.50	0.40	0.30	0.12
NVP-AUY922	MCF7	0.00	0.50	0.67	0.60	0.40	0.27	0.20	0.08
SN38	MCF7	1.00	0.50	0.33	0.20	0.20	0.13	0.10	0.08
NMS-doxorubicin	MCF7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04
Average Value:		0.56	0.56	0.56	0.51	0.34	0.26	0.21	0.10

c) DN, AES distance

		PPV when considering the first n neighbors							
Tested Compound	*Treated Cell Line*	$n = 1$	$n = 2$	$n = 3$	$n = 5$	$n = 10$	$n = 15$	$n = 20$	$n = 50$
PHA-848125	A2780, MCF7	1.00	1.00	0.67	0.40	0.40	0.33	0.25	0.25
flavopiridol	A2780	1.00	1.00	1.00	0.60	0.50	0.40	0.40	0.20
PHA-690509	A2780	1.00	1.00	0.67	0.60	0.40	0.40	0.35	0.18
PHA-793887	A2780, MCF7	1.00	1.00	0.67	0.60	0.40	0.40	0.38	0.12
NMS-tanespimicyn	MCF7	1.00	1.00	1.00	0.80	0.40	0.27	0.20	0.08
NMS-E973	MCF7	1.00	1.00	1.00	0.80	0.40	0.27	0.20	0.08
NVP-AUY922	MCF7	1.00	1.00	0.67	0.40	0.20	0.13	0.10	0.04
SN38	MCF7	1.00	1.00	0.67	0.40	0.40	0.27	0.20	0.08
NMS-doxorubicin	MCF7	0.00	0.50	0.67	0.40	0.20	0.13	0.10	0.04
Average Value:		0.89	0.94	0.78	0.56	0.37	0.29	0.24	0.12

d) DN, MES distance

		PPV when considering the first n neighbors							
Tested Compound	*Treated Cell Line*	$n = 1$	$n = 2$	$n = 3$	$n = 5$	$n = 10$	$n = 15$	$n = 20$	$n = 50$
PHA-848125	A2780, MCF7	0.00	0.50	0.67	0.60	0.40	0.30	0.30	0.12
flavopiridol	A2780	1.00	1.00	1.00	0.60	0.40	0.40	0.40	0.20
PHA-690509	A2780	1.00	0.50	0.67	0.60	0.40	0.47	0.35	0.18
PHA-793887	A2780, MCF7	0.00	0.50	0.67	0.60	0.40	0.40	0.25	0.16
NMS-tanespimicyn	MCF7	1.00	1.00	1.00	0.80	0.50	0.33	0.25	0.10
NMS-E973	MCF7	1.00	1.00	1.00	0.80	0.40	0.33	0.25	0.10
NVP-AUY922	MCF7	1.00	1.00	1.00	0.80	0.40	0.27	0.20	0.08
SN38	MCF7	1.00	1.00	1.00	0.60	0.40	0.27	0.20	0.08
NMS-doxorubicin	MCF7	1.00	0.50	0.33	0.40	0.40	0.33	0.25	0.04
Average Value:		0.77	0.77	0.81	0.64	0.41	0.34	0.27	0.13

Table 6.6: Classification performances ROC analysis. PPV values when considering the first n neighbors (according to our drug distances) and the connectivity scores of the cMap online tool queried with traditional signatures. The color of the n-th PPV is red if no TPs were found among the first n predictions, is green if at least one TP was found. Finally, is black if for in the considered case less than n predictions were significant.

We would like to point out that we used for this comparison 11 gene expression profiles for 9 compounds: for the testing compounds PHA-793887 and PHA-848125 we had two gene expression profiles since two different cell lines were treated with them. Our methods merges together multiple data for a given drug (with the rank merging procedure) but the cMap online tool does not. Consequently, in order to have comparable results, we used the cMap online tool also on a subset of 9 gene expression profiles including, for PHA-793887 and PHA-848125, only the best-classified profile between the two available ones (panel b in table 6.6). Although, even with this "supervised little help" the cMap online tool performed worse than both our distances.
Additionally, the particular case of the tested compound NMS-doxorubicin shows that our approach is able to correctly classify drugs with high precision and sensitivity where the cMap classic online tool clearly fails (see Table 6.7).

Interestingly, Table 6.7 shows also that the AES distance is generally more stringent and reliable whereas the MES one is more sensitive to weak similarities and provides a lower PPV but a higher "recall" (i.e. the ratio of true positives recognized among all the possible ones).
Moreover, the usefulness of our DN classification approach and its output format is demonstrated in the following example: When the NMS-tanespimycin signature, including the 142 maximally up-regulated and the 61 maximally down-regulated probe sets (available in the SDD, where previously specificed), was used to interrogate the cMap in the classic way, geldanamycin, tanespimycin, alvespimycin, and monorden ranked among the top six hits (see appendix E), that also included the protein synthesis inhibitor emetine. However, the next top hits up to position 29 were a miscellaneous of chemicals most of which cannot clearly be related to the Hsp90 and/or ubiquitin protein degradation inhibition. Known proteasome inhibitors ranked position 29 and 30. Similar results were obtained by querying the cMap classic online tool with the gene signatures of the other two Hsp90 tested inhibitors.
On the contrary, the subnetwork containing the tested compounds (Figure 6.1, A) and all their significant neighbors provides a modular and meaningful view of the DN approach output where drugs are grouped according to their effects. This allows users to easily interpret the obtained output and to make a hypothesis on the MoA of a new drug in a clearer way.

cMap Classic Query System			**DN MES distance**		**DN AES distance**	
#Connections = 68, PPV = 3%, First TP in 29th position.			#Connections = 26, PPV = 20%, First TP in 1st position.		#Connections = 5, PPV = 40%, First TP in 2nd position.	
CS	*p*	*Compound*	MES	*Compound*	AES	*Compound*
0.761	0.00000	resveratrol	0.559	daunorubicin	0.781	mycophenolic acid
0.598	0.00000	thioridazine	0.649	GW-8510	0.793	etoposide
0.572	0.00000	trichostatin A	0.654	hycanthone	0.794	daunorubicin
0.977	0.00004	camptothecin	0.655	ellipticine	0.810	hycanthone
0.927	0.00004	trifluridine	0.669	irinotecan	0.823	MG-262
0.574	0.00004	trifluoperazine	0.690	camptothecin		
0.572	0.00006	15-delta prostaglandin J2	0.692	etoposide		
0.266	0.00006	LY-294002	0.693	mycophenolic acid		
0.959	0.00010	mycophenolic acid	0.700	phenoxybenzamine		
0.959	0.00010	proscillaridin	0.718	doxorubicin		
0.885	0.00018	digitoxigenin	0.726	0175029-0000		
0.476	0.00026	fluphenazine	0.734	mepacrine		
0.879	0.00030	bufexamac	0.743	5151277		
0.866	0.00044	thiostrepton	0.744	apigenin		
0.864	0.00046	phenoxybenzamine	0.752	5109870		
0.918	0.00114	irinotecan	0.758	vorinostat		
0.712	0.00159	cloperastine	0.760	scriptaid		
0.813	0.00235	digoxin	0.763	alsterpaullone		
0.503	0.00239	vorinostat	0.764	resveratrol		
0.789	0.00394	norcyclobenzaprine	0.772	cytochalasin B		
0.869	0.00413	scriptaid	0.782	piperlongumine		
0.716	0.00433	antimycin A	0.786	tyrphostin AG-825		
0.775	0.00487	hycanthone	0.790	HC toxin		
0.771	0.00525	monobenzone	0.793	trifluridine		
0.759	0.00635	withaferin A	0.800	MG-262		
0.623	0.00886	helveticoside	0.806	ciclopirox		
0.737	0.00939	pinacidil				
0.927	0.01012	quinostatin				
0.732	0.01017	daunorubicin				
0.927	0.01022	MS-275				
0.820	0.01166	0297417-0002B				
0.718	0.01307	alimemazine				
0.596	0.01412	quercetin				
0.709	0.01476	pimozide				
0.701	0.01643	zalcitabine				
0.795	0.01747	cefotetan				
0.637	0.01788	cinchocaine				
0.688	0.01975	etoposide				
0.687	0.02007	methyldopate				
0.686	0.02085	zuclopenthixol				
0.680	0.02246	strophanthidin				
0.776	0.02269	fenoterol				

Color Legend: Topo II inhibitors, Topo I inhibitors, CDK2

Table 6.7: The NMS-doxorubicin classification case.

6.6 Rank Merging Impact on the performances

As shown in Chapter 3, some recent approaches attempted to use cMap data to build a drug similarity network by selectively comparing pairs of individual GEPs (62) rather than pairs of drug PRLs, as done in our approach. However the use of individual GEPs tend to group together profiles coming from the same cMap batch experiment, or the same cell line, rather than grouping drugs with similar MoA. To avoid this problem, it is necessary to merge together all of the differential expression profiles obtained with the same drug, on different cell lines and at different dosages, prior to computing distances. As introduced in Section 3.3.2, to show the effect of using individual GEPs, for each GEP we considered the K closest GEPs in the cMap dataset, according to the distance. We then computed the percentage of these closest GEPs (i.e. PPV, see Figure 6.6) that were obtained by treating cells with the same drug (green line in Figure 6.6) as the GEP under consideration, or in the same cell lines, regardless of the drug (blue line in Figure 6.6), or in the same batch experiment, regardless of the drug (red line in Figure 6.6).

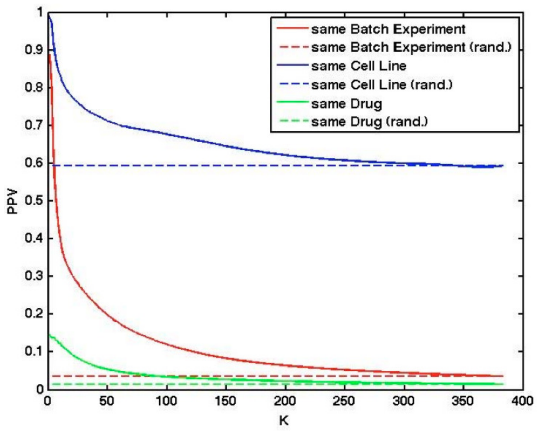

Figure 6.6: Individual GEPs distance assessment - PPV curves considering as positive predicted values neighboring GEPs obtained in the same batch experiment (red), by treating the same cell line (blue), or by treating with the same drug (green) of the GEP under consideration. Average values have been computed across the whole dataset. Dashed lines denotes random performances.

We therefore concluded from Figure 6.6 that using individual GEPs to compute the similarity distance between drugs is not able to catch similarities in MoAs because

of the inability to discriminate treatments obtained with different drugs in the same experimental setting.

In order to additionally assess the impact of the PRL merging procedure on the classification performance of our tool, we specifically produced additional microarray data by treating U251 cells with PHA-848125 at 3 μM, a dose equal to 5$\times$ the IC_{50} for 6 hours. We then merged the set of GEPs by using different combinations of them, and we evaluated the ability of our tool to classify the resulting different PRLs. Results of this assessment are summarized in table 6.8.

Performances are measured by means of ROC analysis, considering the neighborhoods as sets of predictions and the sets of drugs in table 6.4 as correct predictions. The Area under the curve (AUC) (i.e. ROC curve) has been measured as well.

	Treated Cell Line	PPV							AUC
		$n = 1$	$n = 2$	$n = 3$	$n = 5$	$n = 10$	$n = 50$	$n = 100$	
1	SF539	0	0.5	0.33	0.2	0.1	0.2	0.12	46.81
2	A2780	0	0.5	0.33	0.2	0.4	0.12	0.07	48.73
3	U251	0	0.5	0.33	0.4	0.4	0.24	0.12	54.27
4	MCF7	0	0.5	0.67	0.6	0.4	0.22	0.11	55.7
5	MCF7, SF539	0	0.5	0.67	0.4	0.4	0.24	0.12	55.03
6	A2780, MCF7, U251	1	0.5	0.33	0.4	0.5	0.22	0.12	57.06
7	A2780, MCF7, U251, SF539	1	0.5	0.33	0.4	0.4	0.24	0.12	56.58

Table 6.8: Impact of the rank merging on the classification performances.

As expected the best performance is obtained when the PHA- 848125 PRL derives from treatments on all three cell lines. Specifically, by using the profiles individually the best classified was the one obtained by treating the MCF7 cells. This is quite obvious, first of all because MCF7 is the most recurrent cell line among those treated in the cMap dataset. Moreover for A2780 and U251 there are no treatments at all in the cMap.

However, once we combined the profiles from MCF7 with that from A2780 or U251, classification performances are still good, although the combination with U251 gives a less efficient classification.

U251 cell line is the most diverse cell line among the three, since glioblastoma is a very heterogeneous disease where different pathways are known to be disrupted, which might explain the observed signal dilution.

Nevertheless, when combining the profiles coming from all three cell lines together

(MCF7, A2780, and U251), we obtained the best performances in classification, supporting the hypothesis that a sufficiently large combination of treated cell lines provides a sufficiently general summary of the drug activity, which is well classified in the majority of the cases. We further explored the robustness of our method in classifying drugs by pooling together profiles coming from treatments on cell lines with a very different genetic background, potentially causing a significant signal dilution.

To this aim we collected additional gene expression data by treating SF539 human glioma cell line with PHA-848125 for 6 hours.

The SF539 cell line is genotipically characterized by a mutation in the RB gene (encoding for the retinoblastoma tumor suppressor protein), whereas the other three treated cell lines (A2780, MCF7, and U251) are RB wild type.

DNA replication and the regulation of the G1/S transition is under the control of the RB/E2F pathway. In wild-type cells RB binds the E2F-1 transcription factor, thus inhibiting its regulatory activity. When RB is phosphorylated by CDK2, it releases E2F-1 that mediates the cell cycle progression (107).

In the SF539 cell line, RB is no longer able to block E2F-1, which is constitutively active in this cell line as a result. As a consequence, inhibiting CDK2 with PHA-848125 on SF539 will not have the same effect on the E2F mediated transcription that is elicited in the RB wild-type cell lines.

Following the strategy previously described, we merged the set of gene expression profiles obtained by treating A2780, MCF7, U251, and SF539 with PHA-848125, and we evaluated the ability of our tool to classify the resulting different PRLs.

The whole neighborhoods obtained in this assessment are available in the appendix G.

Results, listed in Table 6.8, show that by using expression profiles individually from a single cell line, the best classification is obtained with the MCF7 cell line. This is to be expected, because MCF7 is the most recurrent cell line among those treated in the cMap dataset. The worst classification was instead obtained with the SF539 cell line, which is coherent with the RB inactivation that mediates the MoA of the PHA-848125 compound.

Nevertheless, once we combined the profile coming from all four cell lines together (MCF7, A2780, U251, and SF539), or even of two cell lines (MCF7 and SF539) only, we improved the classification performance considerably.

These results support the hypothesis that a sufficiently large combination of treated

cell lines provides a sufficiently general summary of the drug activity, which is well classified by our method.

6.7 Discussion

In this chapter, we showed that our method is a general procedure that is able to predict the molecular effects and MoA of new compounds. We were able to exploit information hidden in the gene expression profiles following drug treatment to capture similarity in drug MoA. Previous attempts to use gene expression profiles following compound treatment in mammalian cells did not consider the variability in the transcriptional response to the compound due to cell-line effects, to different dosages, and to different experimental settings. Moreover, information embedded in the global structure of the network of similarities among drugs has not been fully exploited in the past. We removed unspecific effects by capturing the consensus transcriptional response to a compound across multiple cell lines and dosages. We then automatically extracted a gene signature for each compound and computed pairwise similarities between compounds using a gene signature-based approach.

In Chapter 4 we analyzed the resulting network to identify communities of drugs with similar MoA and to determine the biological pathways perturbed by these compounds. We remark that, differently from other methods, whose aim is to identify the specific drug substrates (34, 96), our approach also groups together compounds interacting with distinct members of the same pathway.

In conclusion, the DN can be used to infer the MoA and targeted pathways of anticancer compounds still being studied. We correctly classified both known and previously undescribed Hsp90 inhibitors. Interestingly, in addition to the Hsp90 inhibitors present in the database (alvespimycin, geldanamycin, and monorden), several drugs included in the top 10 closest neighbors for NMS-tanespimycin and NMS-E973 were connected to inhibitors of the proteasome/NF-kB pathway, including disulfiram (29), withaferin A (151), and parthenolide (57).

We also investigated the ability of our method in classifying well-known (flavopiridol) and novel CDK inhibitors (PHA-690509, PHA-793887, and PHA-848125). These drugs were correctly classified as CDK inhibitors, distinct from the other kinase inhibitors in the database, and were also predicted to be very similar to Topo inhibitors.

Although the induction of p21 by DNA damage-inducing agents was previously reported, here we showed that this is clearly detected at the transcriptional level, supporting the concept that gene modulations can be used as a biomarker to monitor the effect of DNA damage-inducing agents.

As described in Chapter 5, the DN can be useful for formulating hypotheses on the MoA of novel compounds by simply measuring multiple transcriptional responses in different cell lines. In addition, by analyzing the PRLs associated to each drug in the network, we may identify the drug communities that consistently up-, or down-regulate a given set of genes, thus hinting to drug classes able to modulate a specific pathway of interest. The major limitation of our approach is in the limited number of compounds in the network. Because our approach is based on comparing how similar two drugs are, if a compound is not similar to any of the drugs in the network, no inference on its MoA or its biological effects can be done. Moreover, for a compound having inconsistent effects on different cell lines (for example, due to a cell line with a mutated substrateprotein targeted by the compound) merging gene expression profiles from distinct cell lines may dilute the biological effects of the compound. Nevertheless, when no information on the drug MoA is available a priori, the best strategy is still to merge profiles from multiple cell lines.

As shown in this chapter, merging profiles coming from a sufficiently large, even if heterogeneous, pool of treated cell lines, provides a summary of the transcriptional response to the drug that can still be well classified by the DN. Considering that we have made our approach publicly available as an online tool, it is clear that the DN can be easily queried with the transcriptional responses of a unique compound, thus providing a valuable tool to the research community.

7

MANTRA predicts candidates for Drug Repositioning

7.1 Introduction

This chapter contains an interesting example of drug reposition proposed by MANTRA. Thanks to our tool we discovered that fasudil, a safe vasodilator, enhances a metabolic process known as cellular autophagy. In Section 7.2 a brief overview of this process, the description of its roles in the aetiology of neurodegenerative disorders, and a discussion about how its enhancement is clinically effective are provided.
In section 7.3 we explain how, starting from a known and safe cellular autophagy enhancer already present in our DN, we obtained a list of similar drug by using MANTRA. This list contained a number of known cellular autophagy enhancer and fasudil, whose ability in enhancing cellular autophagy has never been reported before.
The rest of the chapter contains the description of the experiments we conducted in order to verify this novel MoA of fasudil and a discussion about the possible implications of this discovery on some therapeutic approaches. Conclusions are reported in the final section.

7.2 Overview of the mechanism of cellular autophagy

Autophagy, or autophagocytosis, is a catabolic process involving the degradation of a cell's own components through the lysosomal machinery. It is a tightly-regulated process that plays a normal part in cell growth, development, and homeostasis, helping

to maintain a balance between the synthesis, degradation, and subsequent recycling of cellular products. It is a major mechanism by which a starving cell reallocates nutrients from unnecessary processes to more-essential processes.
A variety of autophagic processes exist, all having in common the degradation of intracellular components via the lysosome. The most well-known mechanism of autophagy involves the formation of a membrane around a targeted region of the cell, separating the contents from the rest of the cytoplasm. The resultant vesicle then fuses with a lysosome and subsequently degrades the contents (Figure 7.1).

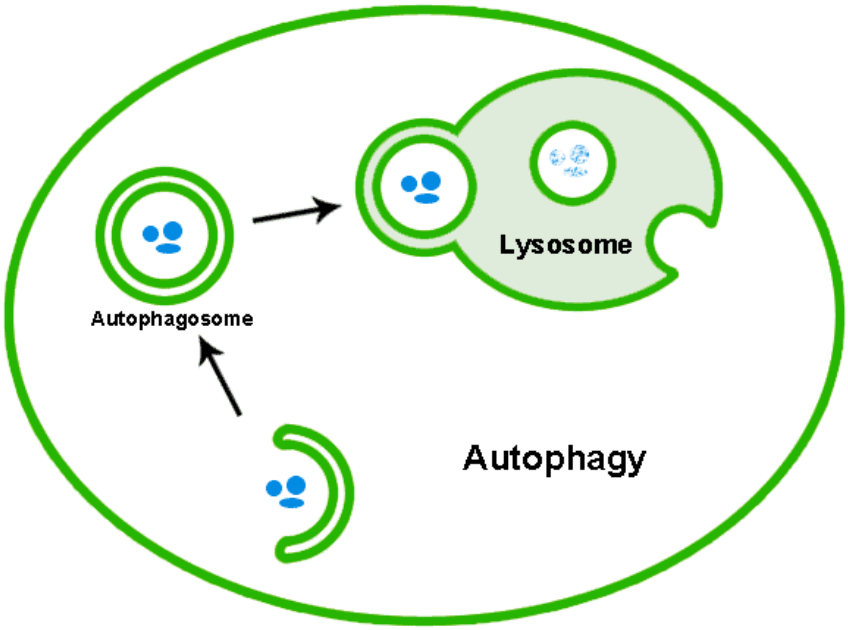

Figure 7.1: Cellular Autophagy - An autophagosome sequesters cytoplasmic constituents, such as mitochondria, endoplasmic reticulum, and ribosomes, by forming a double membrane vesicle. The outer membrane of the autophagosome then fuses with the lysosome in mammalian cells delivering the sequestered content to the lumen of lysosome for degradation

Autophagy is critical for the survival of yeast and mammalian cells under starvation conditions because it functions to recycle intracellular material for macromolecular synthesis and energy production (90). Autophagy occurs in all cells at low basal levels under normal conditions to perform homeostatic functions, but it can be rapidly up-regulated under starvation or stress conditions (90).

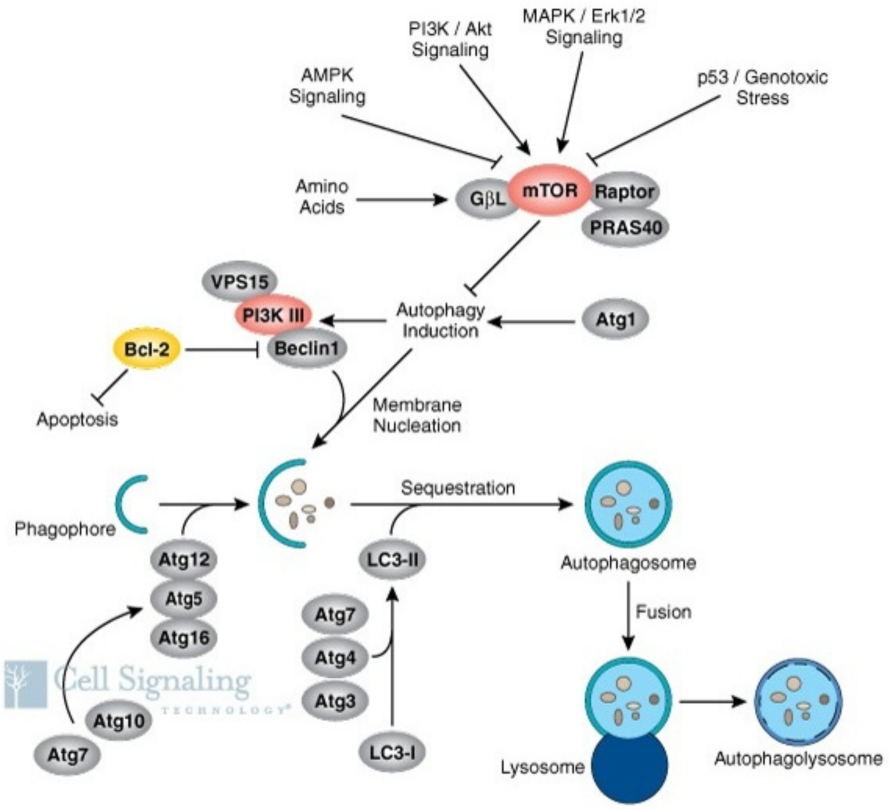

Figure 7.2: Autophagic pathways - Biological pathways involved in cellular autophagy [Image from: http://www.cellsignal.com/].

During autophagy the Light chain 3 (LC3) protein localizes to the autophagosomal membrane (72) and mammalian LC3 has been shown to mark the autophagosome membrane specifically.

Autophagy is a key process involved in the pathogenesis of a wide range of human disorders in which misfolded protein aggregation is the causative pathologic event. This happens in various neurodegenerative diseases such as Parkinsons disease, Alzheimers disease and Huntington disease. As an example, in Huntington disease, the UPS is thought to be impaired, which leads to the formation of insoluble protein aggregates. In this case, autophagy helps maintain cellular homeostasis by clearing damaged organelles and unfolded proteins. Moreover, autophagy is involved in the degradation of various pathogens and its deficiency predisposes to tumorigenesis and aneuploidy.

Increasing autophagy may provide clinical benefit in the treatment of various diseases, and therefore there is a great effort in developing drugs enhancing this function.

As shown in Figure 7.2, in mammalian cells, mTOR kinase, the target of rapamycin, mediates the major inhibitory signal that shuts off autophagy under nutrient-rich conditions (90) but also other pathways are involved in this process. mTOR activity can be inhibited by rapamycin hence causing an mTOR-dependant cellular autophagy. However, the cytotoxicity of rapamycin and its dangerous side effects (such as immune system suppression) has prevented its applications.
Nowaday, there is a great effort in developing safe drugs modulating autophagy, and various approaches have been taken towards this goal.

7.3 Drug repositioning proposals through established-drug neighborhood analysis

Recently, glucose has been proposed as a novel, natural and safe enhancer of cellular autophagy (118). The cMap (and our DN as well) contains gene expression profiles obtained by treating with 2DOG, a glucose molecule which has the 2-hydroxyl group replaced by hydrogen, so that it cannot undergo further glycolysis (i.e. the process that converts glucose into pyruvate, releasing free energy used to form the high-energy compounds ATP).

By using 2DOG to interrogate the drug network we identified a list of drugs that were included in the neighborhood of 2DOG, hence predicted to share a similar mode of action with this drug. This list is provided in table 7.1 and included fasudil, thapsigargin, trifluoperazine, gossypol and niclosamide as closest neighbors (Figure 7.3). Of these, thapsigargin, trifluoperazine, gossypol and niclosamide are previously known inducers of autophagy (25, 83, 110, 119).

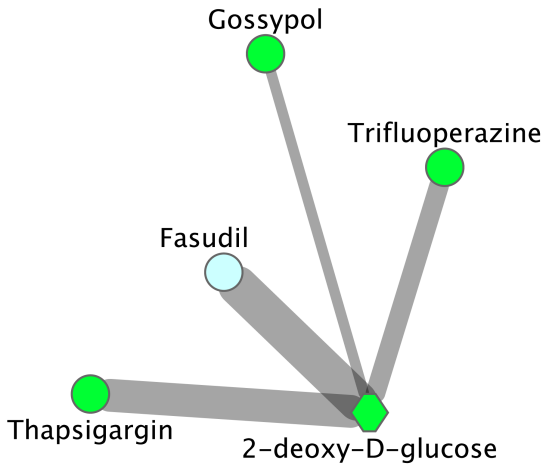

Figure 7.3: 2-deoxy-d-glucose closest neighbors - Each dot corresponds to a drug of the DN and edge thickness is inversely proportional to the MES drug distance. In green are shown drugs previously known to induce autophagy.

Additionally, 2DOG is the exemplar of community n. 1, which contains, in increasing order of distance to 2DOG, fasudil, sodium-phenylbutirate, tamoxifen, arachidonyl-trifluoromethane, and novobiocin (see table 7.2). In this community 2 drugs are known autophagy inducers (2DOG and tamoxifen (16, 31)).
2DOG is also linked to other communities exemplars and is part of the rich-club whose members are listed in table 7.3.

In this rich-club, 3 out of the 4 exemplars connected to 2DOG are known autophagy inducers: trifluoperazine (25), ciclosporin (113) and oligomycin (142).

Neighbor	MES distance from 2DOG
Fasudil	0.5162
Thapsigargin*	0.5644
Trifluoperazine*	0.5770
Gossypol*	0.6330
Niclosamide*	0.6539
Tyrphostin AG-1478	0.6682
Valinomycin	0.6780
Ivermectin	0.6792
Sodium phenylbutyrate	0.6833
BW-B70C	0.6905
Calmidazolium	0.6912
5224221	0.6968
MG-132	0.6971
Desipramine	0.7007
Rottlerin	0.7013
Clotrimazole	0.7054
Mefloquine	0.7066
Ionomycin	0.7087
Tamoxifen*	0.7143
Cytochalasin B	0.7164
Ciclosporin*	0.7201
Puromycin	0.7268
Pyrvinium	0.7283
Astemizole	0.7290
Alexidine	0.7305
Disulfiram	0.7311
Fendiline	0.7329
Prochlorperazine	0.7387
Anisomycin	0.7397
Pararosaniline	0.7417
Chlorprothixene	0.7420
Loperamide	0.7422
Mometasone	0.7439
Iloprost	0.7475
0297417-0002B	0.7480
Thioridazine	0.7488
MG-262	0.7500
Spiperone	0.7556
Arachidonyltrifluoromethane	0.7599
Methylbenzethonium chloride	0.7615
5707885	0.7630
Oligomycin	0.7701
Podophyllotoxin	0.7725
Homochlorcyclizine	0.7736
Perphenazine	0.7742
Celastrol	0.7752
Vanoxerine	0.7760
Idoxuridine	0.7760
5666823	0.7765
Hydroxyzine	0.7766
Nordihydroguaiaretic acid	0.7776
Geldanamycin	0.7776
Metergoline	0.7777
Novobiocin	0.7779
Terfenadine	0.7781
Butoconazole	0.7787
Piroxicam	0.7808

*Known enhancers of cellular autophagy

Table 7.1: Neighbors of 2DOG in the DN

Community n. 1
2-deoxy-D-glucose*
fasudil
tamoxifen*
sodium phenylbutyrate
arachidonyltrifluoromethane
novobiocin

*Known enhancers of cellular autophagy

Table 7.2: Composition of community n. 1

Exemplar	Community
2-deoxy-D-glucose *	1
Trifluoperazine*†	100
Ciclosporin*†	43
Astemizole†	34
Oligomycin*†	78
Gefitinib	60
5114445	4
Esculetin	54
Dimethyloxalylglycine	51
Demecolcine	48
Zardaverine	106
CP-319743	10
Terconazole	92
3-aminobenzamide	2
Mycophenolic acid	75
HC toxin	16

*Known enhancers of cellular autophagy

†Exemplars connected to 2DOG

Table 7.3: 2DOG network rich-club

7.4 MANTRA predicts that Fasudil promotes cellular autophagy

The DN topology surrounding 2DOG suggests that fasudil could be an autophagy enhancer.

Despite being a drug with a well-characterized MoA, fasudil has never been previously linked to autophagy so this could be a very interestingly case of "repositionable" drug for conditions in which cellular autophagy could have a therapeutic efficacy.

Fasudil is a RHO kinase (ROCK) inhibitor used to reverse blood vessel spasm occurring after subarachnoid hemorrhage. Besides subarachnoid hemorrhage, clinical applications of fasudil include various types of cardiovascular diseases, such as acute ischemic stroke, stable angina pectoris, coronary artery spasm, heart failure-associated vascular resistance and constriction, pulmonary arterial hypertension, essential hypertension, atherosclerosis and aortic stiffness.

Interestingly, previous studies have shown that Y-27632, an analog of fasudil not currently approved for clinical use, is effective at reducing the aggregation of several polyglutamine proteins (130), including mutant Huntingtin (Htt) (10), which plays a crucial role in a patology known as Huntington's disease (HD).

HD is a progressive neurodegenerative genetic disorder, which affects muscle coordination and leads to cognitive decline and dementia. It typically becomes noticeable in middle age. HD is the most common genetic cause of abnormal involuntary writhing movements called chorea and is much more common in people of Western European descent than in those from Asia or Africa. The disease is caused by an autosomal dominant mutation on either of an individual's two copies of the gene coding for the Htt protein, which means any child of an affected parent has a 50% risk of inheriting the disease. In rare situations where both parents have an affected gene, or either parent has two affected copies, this risk is greatly increased. Physical symptoms of HD can begin at any age from infancy to old age, but usually begin between 35 and 44 years of age. About 6% of cases start before the age of 21 years with an akinetic-rigid syndrome; they progress faster and vary slightly.

The mutation of the Huntingtin gene codes for a different form of the protein, whose presence results in gradual damage to specific areas of the brain. The exact way this happens is not fully understood. The Htt protein interacts with over 100 other proteins,

and appears to have multiple biological functions. The behavior of mutated Htt protein is not completely understood, but it is toxic to certain types of cells, particularly in the brain.
The discovered effect of Y-27632 on mutant Htt was attributed to enhancement of degradation by macroautophagy and the UPS, mediated by inhibition of ROCK 1 and 2 (130).
Interestingly, in our drug network, in contrast to fasudil, Y-27632 was found in community n. 40, which is enriched for small molecules functioning as UPS modulators (see appendix B.1), raising the hypothesis that the effects of Y-27632 on UPS are stronger than those on autophagy.

7.5 Experimental validation

To verify the efficacy of fasudil in enhencing autophagic activity, as predicted by our approach, we evaluated the levels of the second isoform of the LC3 protein (LC3-II) in wild-type human fibroblasts treated with fasudil, by WB with anti-LC3 antibody, a well-established assay for the activation of autophagy (105). In fact, as introduced in Section 7.2, during autophagy, a cytosolic form of the LC3 protein (LC3-I) is conjugated to phosphatidylethanolamine to form LC3-phosphatidylethanolamine conjugate (LC3-II), which is recruited to autophagosomal membranes, and LC3-II is degraded by lysosomal hydrolases after the fusion of autophagosomes with lysosomes. Therefore, lysosomal turnover of LC3-II reflects starvation-induced autophagic activity, and detection of LC3 by immunoblotting or immunofluorescence has become a reliable method for monitoring autophagy.
We measured a marked increase in LC3-II levels in fibroblasts treated with fasudil and trifluoperazine identified by the DN, as well as, in cells treated with 2DOG and rapamycin, two well-known inducers of autophagy (Figure 7.4).

Immunostaining with LC3 antibody further confirmed the WB analysis, demonstrating a strong activation of autophagic degradation upon treatment with fasudil (Figure 7.5).

The effect of fasudil on autophagy enhancement was further confirmed in HeLa cells (Figure 7.6).

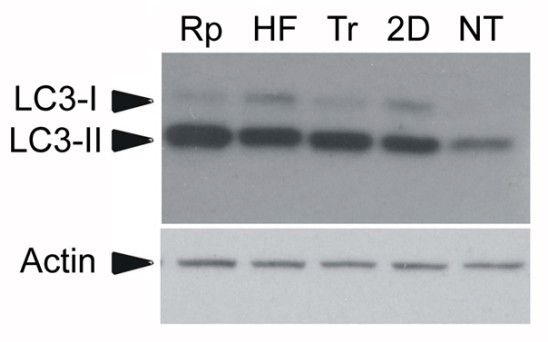

Figure 7.4: Effects of fasudil on autophagy (1) - Evaluation of LC3 levels in human fibroblasts after treatment with drugs: (Rp, rapamycin; HF, fasudil; Tr, trifluoperazine; 2D, 2-deoxy-D-glucose; NT, untreated). The experiments were performed in triplicate, and representative results are shown.

7.6 Hypotheses and consequences

We do not know which is the mechanism resulting in the enhancement of autophagy by fasudil. ROCKs, existing as two isoforms (ROCK1 and ROCK2), are Serine/Threonine (Ser/Thr) protein specific kinases, which are downstream targets of the small Guanosine Triphosphate (GTP)ase Ras homolog gene family (Rho)A, primarily involved in cytoskeletal regulation.

ROCKs regulate a wide range of biological functions including cell growth, migration and apoptosis. Since knockdown of both ROCK1 and 2 results in autophagy activation (10), the effect of fasudil on autophagy is likely mediated by its known inhibitory effect on these proteins. However, whether and how cytoskeletal changes due to Rho/ROCK inhibition results in activation of UPS and autophagy remain unknown.

Activation of ROCKs by GTP-bound Rho results in phosphorylation of various target proteins. One of the main substrates of ROCK is Myosin Light Chain (MLC) that stimulates myosin-actin interactions. Other downstream targets of ROCKs include the Ser/Thr kinases LIM kinase (LIMK) 1 and 2.

Besides the action on MLC, which underlies its therapeutic effect against vasospasm, fasudil appears to have other cellular effects.

Substrates of ROCK, for example, include the glial fibrillary acidic protein, neurofilaments that upon phosphorylation by ROCK undergo depolymerization and Microtubule-Associated Protein 2 (MAP2).

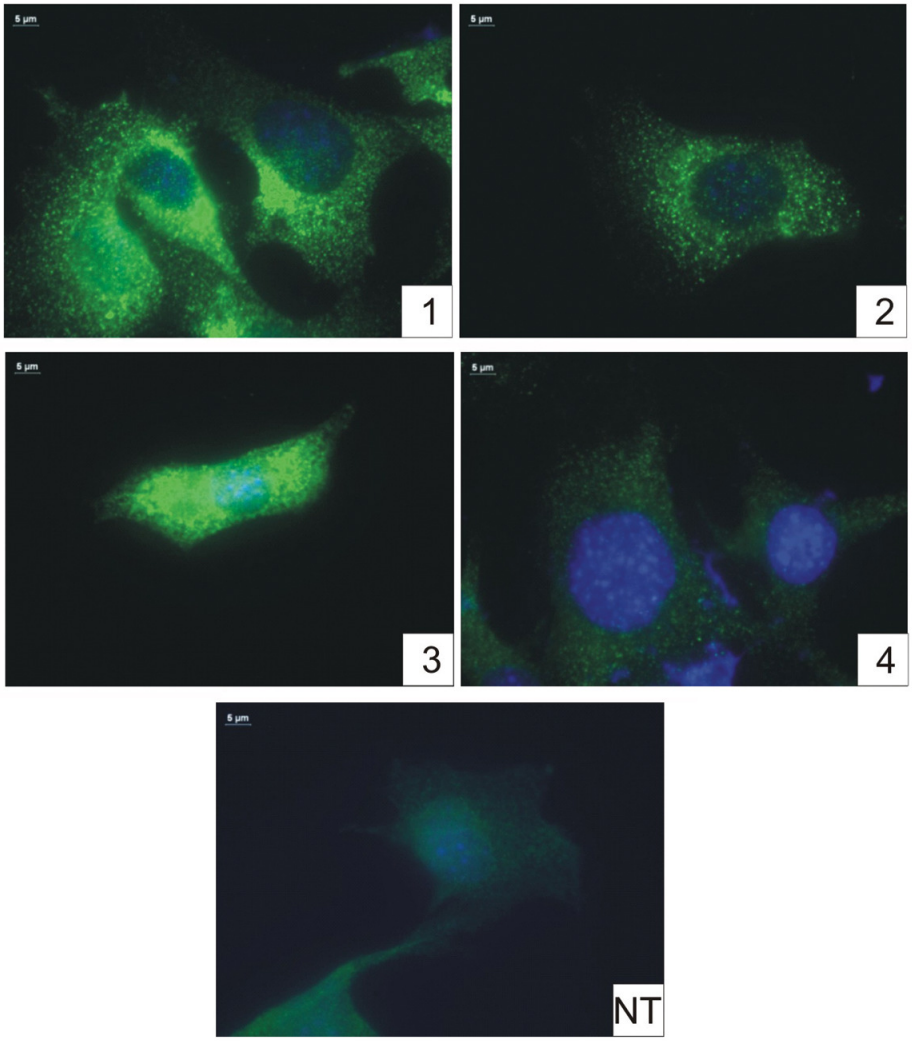

Figure 7.5: Effects of fasudil on autophagy (2) - Immunofluorescence with anti-LC3 antibody in fibroblasts treated with drugs promoting autophagy. Evaluation of LC3 levels in human fibroblasts after treatment with drugs: 1, Rapamycin; 2, Fasudil; 3, Trifluoperazine; 4, 2DOG; NT, untreated. The experiments were performed in triplicate and representative results are shown.

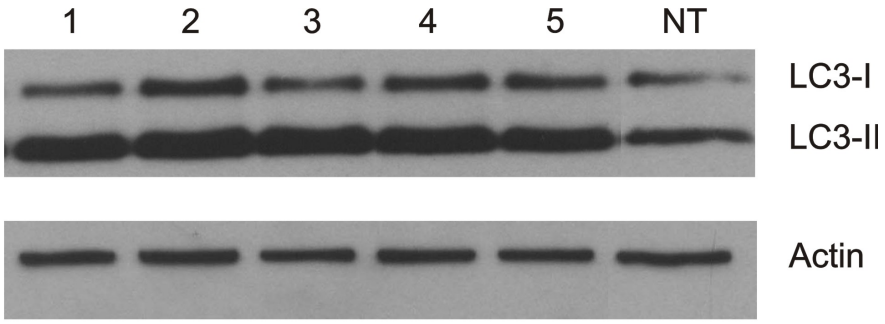

Figure 7.6: Effects of fasudil on autophagy (3) - Effects of Fasudil on HeLa Cells. Western blot with anti-LC3 antibody in drug treated Hela cells treated: 1, rapamycin; 2, dasudil 10 μm; 3, fasudil 30 μM; 4, trifluoperazine 1 μM; 5, 2DOG 100 μM; NT, untreated. LC3-II levels are increased following treatment with fasudil, trifluoperazine, and 2DOG as compared to the untreated control. The experiments were performed in duplicate and representative results are shown.

Based on results obtained in animal models (85), ROCK inhibitors have been proposed to slow down the degenerative process in Alzheimer disease by reducing toxic levels of Aβ42, whose accumulation is thought to be involved in the disease aetiology, and stimulating regenerative growth of neurites (106). Moreover, peripheral delivery of fasudil reduces neuronal death and epilepsy in mice and improves spatial cognition and memory in rats (64). Whether these effects are mediated by enhancing autophagy is an interesting hypothesis which warrants further studies. Nevertheless, these studies suggest that fasudil is able to cross the blood-brain barrier and to reach therapeutic concentrations in the brain, at least in rodents.
In summary, fasudil is a clinically approved drug with potential applications to various human disorders where enhancement of autophagy can provide clinical benefit.

7.7 Discussion

In this chapter we described how we obtained a surprising prediction from MANTRA: fasudil promotes cellular autophagy. Given the excellent safety profile, this newly recognized effect of fasudil could be exploited for disorders due to protein misfolding, including neurodegenerative diseases. This shows that MANTRA is a valid tool for finding previously unrecognized MoAs of well-characterized drugs. Moreover, this can be accomplished in a very quick, easy and cheap way: by simply looking to the topology

surrounding a drugs with the desired MoA and searching in it for safe drugs never linked before to that MoA.
Considering that MANTRA is publicly available and easily usable on-line it is clear that it has an incredible potential (quickly exploitable by users) in finding novel applications for a huge number of approved drugs, hence strikingly speeding up the drug discovery pipe-line.

8

Future directions and Discussion

8.1 Introduction

In this final chapter we discuss about two possible extensions of our approach describing how it could be improved or used for an alternative purpose.
A conclusive summary is provided in the final section and the major strengths and weakness of our method are finally listed.

8.2 Cross platform/species compatibility

The currently implemented version of MANTRA can be used with data coming from Affymetrix®HG-U133a gene chip hybridizations only. This because the used reference dataset (i.e. the cMap) has been created by using this platform exclusively and the DN is obtained by comparing lists of MPI rather than lists of genes in order to retain as more information as possible. In our classification tests (see Chapter 6) we used a more recent platform (Affymetrix®HG-U133 Plus 2.0) including all the MPI of the Affymetrix®HG-U133a platform. Hence we processed this data by simply filtering out the entries corresponding to probe sets not belonging to the platform we used.
Generally, this approach is unsuitable for the integration of data coming from other microarray platforms. The probe/gene mapping of microarray from different brand and for species different from human are even more heterogeneous and hence unusable.
As an example, there are no MPI in common between the Affymetrix®HG-U133a platform and the equivalently popular Affymetrix®Mo430 mouse platform and according to the Affymetrix®conversion table only an amount of 120 sequences match well among

the two platforms. Therefore, reducing the analysis on the MPI mapping for this sequence is unconceivable. In fact our platform contains 22,283 MPI and the mouse one 45,101 (see Figure 8.1). By considering only 120 MPI a massive quantity of informative data would be filtered out wasting the advantages of a large-scale approach such those realized by the DNA microarrays.
On the other hand, if we focus on the "gene-symbol domain" we can see that the MPI set of our reference platform contains 14,467 distinct gene sequences or sub-sequences while the mouse platform MPI is mapped into a set of 21,970 gene sequences or sub-sequences. The intersection of these two set of genes contains 9,783 elements corresponding to genes that are conserved among the two species (see Figure 8.1).
Reducing the analysis of data coming from this mouse platform to these 9,783 Cross-platform conserved genes (CPCGs) affords to take into account of the corresponding 16,063 MPI of the human platform (70%) and 20,528 MPI of the mouse platform (45%).

There will be indeed a consistent loss of information regarding the mouse data, but most of the information contained in our reference dataset and the DN is conserved.
In order to quantify how this filtering influences the classification performances, we sought to test our classification algorithm by using publicly available gene expression data from the ArrayExpress repository (`http://www.ebi.ac.uk/arrayexpress/`) (15) (114). The data, which we downloaded, corresponded to the following experiments:

- Transcription profiling of mouse embryonic stem cell line CGR8 grown in presence of Leukemia Inhibitory Factor and treated with trichostatin A (ArrayExpress ID. = E-TABM-670)

- Transcription profiling of mouse embryonic stem cells cultured with PI3-K signalling inhibitor LY-294002 to identify PI3K-target genes (ArrayExpress ID. = E-TABM-673)

These data have been created and are described in the studies (73) and (139) respectively.
Trichostatin A is an organic compound that serves as an anti-fungal antibiotic and selectively inhibits the class I and II mammalian HDAC families of enzymes. Our DN contains a community (n. 16) that is highly enriched for this MoA and, generally, we observed that this class of compounds elicit a well defined transcriptional response. Hence we chose this example because it is relatively simple.

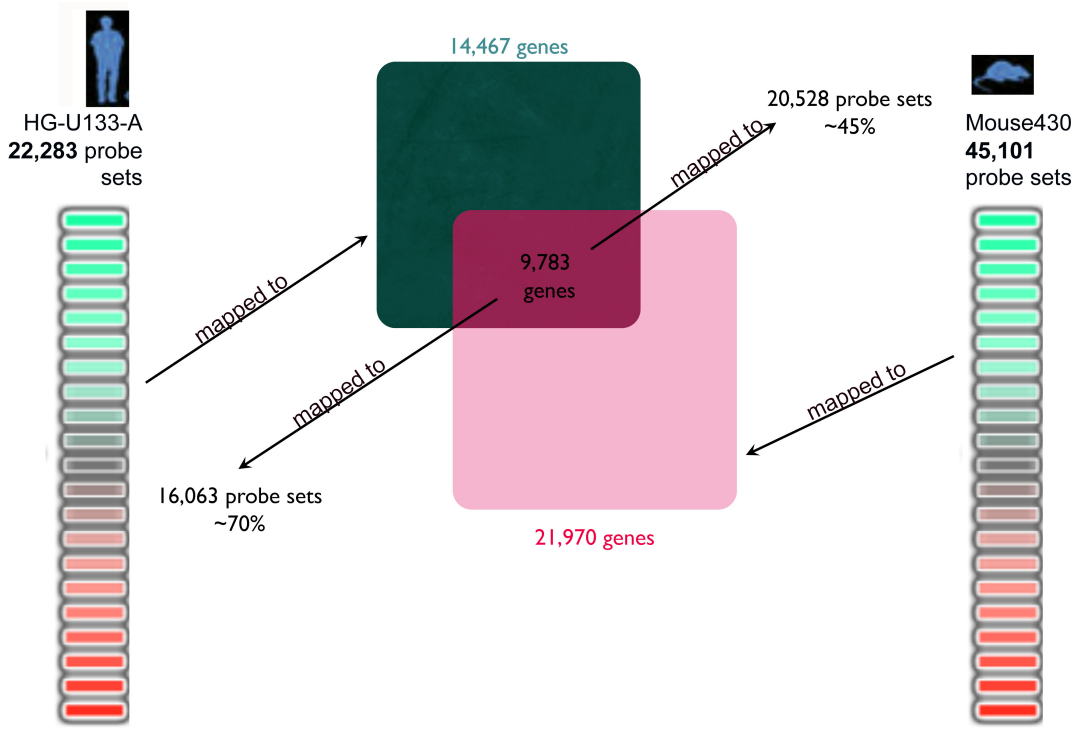

Figure 8.1: Cross-platform conserved genes - Affymetrix human and mouse platforms: 22,283 MPI in the human platform and 45,101 in the mouse platform. 14,467 distinct genes are mapped by the MPI of the human platform and 21,970 genes are mapped by the MPI of the mouse platform. 9,783 gene are contained in the intersection of these two sets and they are mapped by 16,063 MPI of the human platform (70%) and 20,528 MPI of the mouse platform (45%).

LY-294002 is a potent inhibitor of Phosphoinositide 3-kinasess (PI3Ks). The effect of LY-294002 on the transcription is very wide and actually this drug is an outlier in its community in the DN. Hence this could be considered a more difficult example of classification test.
After computing differential expression values for these experiments we applied a filter on the MPI by keeping only those mapping for sequences in the set of the CPCGs and we computed a PRL by sorting the surviving MPI according to their differential expression values. Then we assigned to each of the survived MPI the corresponding gene-symbol as a label. At this stage we removed probe-set mapping for more than one gene and we kept for genes mapped by several MPI only the "farthest-from-the-centre" probe-set (i.e. the probe set whose rank position has the greatest difference from the average value).
Obviously, before computing drug distances we had to apply this pre-processing also to our reference collection of drug PRLs and recomputing distance statistical threshold levels as detailed in the Section 4.5.2.
After this preliminary step we classified the two drugs with our classification algorithm (see Section 5.3 and pseudocode 5). Results of this classification test are shown in Figure 8.2 and Table 8.1.
The trichostatin A four closest neighbors, once it has been integrated in the DN, are known HDAC inhibitors while LY-294002 was connected to sirolimus (rapamycin) and quinostatin two drugs modulating the PI3K-Akt-mTOR pathway. This hints that even if in a broader sense also in the more difficult case the MoA of the tested compound could have been deduced by looking to our classification result.
The outcome of this pilot study suggests that it is more than reasonable to plan an extension of MANTRA with the described preprocessing step in order to enabling cross-platform/cross-specie compatibility and that good classification performances could be kept also in the "gene symbol domain".

8.3 Classification of diseases

Can a generic biological state of interest be represented by a gene signature and a pattern of expression only? A positive response to this question is the leading concept of the cMap and other computational approaches.

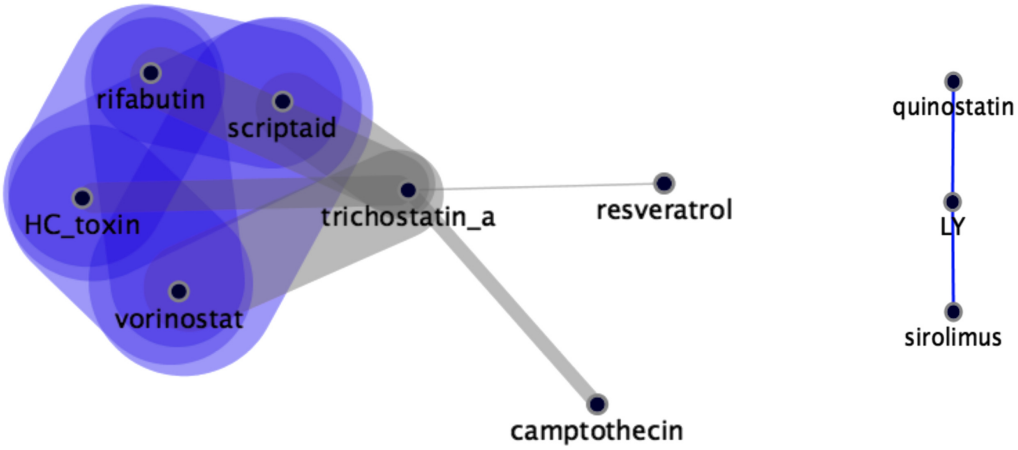

Figure 8.2: Pilot study results on mouse data - Result of MANTRA on the gene symbol domain by using microarray data from Mouse.

Trichostatin A		LY-294002	
MES Distance	**Compound**	**MES Distance**	**Compound**
0.4470	vorinostat*	0.4890	sirolimus†
0.5114	scriptaid*	0.6788	quinostatin†
0.5457	rifabutin*		
0.6502	HC-toxin*		
0.7064	camptothecin		
0.7477	resveratrol		
*HDAC inhibitors		†PI3K-Akt-mTOR pathway modulators	

Table 8.1: Classification results on mouse data: Significant neighbors of the tested compounds

We exploited the cMap database and its querying system by improving its performance on well defined sub-class of biological states: those eliciting a response to a drug treatment. We further investigated a possible use of our approach in classifying also PRLs describing other biological states such as, for example, a disease with transcriptional-influencing phenotype.

To this end, we downloaded from the GEO database, a set of gene expression profiles from hybridizations of the substantia nigra (a sub-structure of the brain, located in the mesencephalon, o midbrain, that plays an important role in reward, addiction, and movement) of postmortem human brain in humans affected by Parkinson's disease (PD) (GEO accession number: GSE7621). This dataset was created in the study (84) where the authors investigated common gene variations that predispose to complex diseases. Some of the obtained results were validated by using generated in-house gene expression data so the cited dataset was created.

To do this, substantia nigra tissue from postmortem brain of normal and PD patients was used for RNA extraction and hybridization on Affymetrix microarrays: 9 replicates for the controls and 16 replicates for the PD patients were used.

We computed profiles of differential expression from this data (PD versus normal patient), we generated a PRL for PD by using the KRUBOR algorithm (subsection 4.2.4, pseudocode 1) and we classified this disease as we classify a compound with our algorithm (see Section 5.3 and pseudocode 5). Results of this classification are reported in Figure 8.3.

According to our classification result, the most "similar to PD" drug (by the induced transcriptional response point of view) is 1,5-isoquinolinediol. Differently from other isoquinoline derivatives that find many therapeutic applications, the pharmacological and toxicological properties of 1,5-isoquinolinediol have not been fully investigated and it is not used in humans. This compound is reported as Poly (ADP-ribose) polymerase (PARP) inhibitor in several studies but more interestingly (and consistently with our classification result) it belongs to a family of endogenous neurotoxins thought to be involved in the aetiology of PD (103). Particularly a neurotoxin called MPTP (1[N]-methyl-4-phenyl-1,2,3,6-tetrahydropyridine) was found and linked to PD in the 1980s. The active neurotoxins destroy dopaminergic neurons, leading to parkinsonism and PD. Several tetrahydroisoquinoline derivatives have been found to have the same neurochemical properties as MPTP. These derivatives may act as neurotoxin precursors

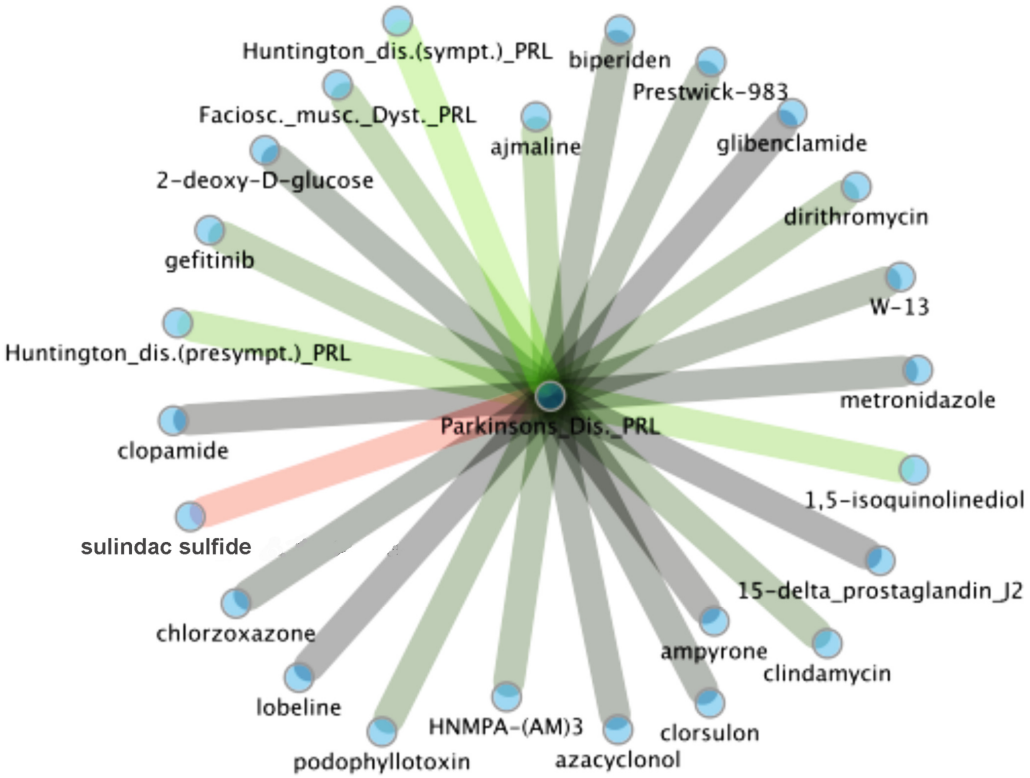

Figure 8.3: Classification of Parkinson's disease - Neighbors (connected by green edges) and an "anti-neighbors" (connected by a red edge) of Parkinson's Disease, once it has been integrated in our DN. Color intensities are proportional to the correspondig similarity relationships.

to active neurotoxins.
In order to check wether the cMap contains drug whose transcriptional response is "anti-correlated" to the transcriptional activity in PD we computed also distances by considering the PD PRL in a reversed order (i.e. down-regulated genes at the top of the list and up-regulated genes at the bottom). The closest drug to this "reversed" PRL was sulindac sulfide. This drug is a dopamine level reducer (30) and dopaminergic agonist are used in the treatment of PD. Combined together, these results provide a very strong rationale to our classification results and hints that gene expression could be used to monitor the activity of the cited neurotoxin (in order to better understand their role in the aetiology of PD) as well as to find novel drug candidates to treat this condition.
In conclusion, this example additionally confirm that genomic signatures together with patterns of expression can be used to summarize biological state of interests and that MANTRA could be extended in order to study complex diseases and other conditions by using gene expression data only.

8.4 Conclusions

In this PhD thesis we presented a three year project in which we designed and implemented a general procedure for the prediction of the molecular effects and the mode of action of new compounds, and to find previously unrecognized applications of well-known drugs.
We were able to exploit the information hidden in a public available collection of gene expression profiles following drug treatment to capture similarity in drug effects. Previous attempts to use gene expression profiles following compound treatment in mammalian cells did not consider the variability in the transcriptional response to the compound due to cell-line effects, to different dosages, and to different experimental settings. Moreover, information embedded in the global structure of the network of similarities among drugs has never been fully exploited in the past.
We removed unspecific effects by capturing the consensus transcriptional response to a compound across multiple cell lines and dosages. We then automatically extracted a gene signature for each compound and computed pairwise similarities between compounds using a gene signature-based approach. We analyzed the resulting network to

identify communities of drugs with similar effects and to determine the biological pathways perturbed by these compounds. We remark that, differently from other methods, whose aim is to identify the specific drug substrates, our approach also groups together compounds interacting with distinct members of the same pathway. Our drug network can be used to infer the mode of action and the targeted pathways of anticancer compounds still being studied and to find candidates for drug repositioning (i.e., to suggest novel clinical application for well-known and approved drugs).
We correctly classified both known and novel drugs. By doing this we discovered a previously unreported similarity in the effect elicited by two different classes of compounds.
In addition, we experimentally verified a surprising prediction by discovering an unreported effect of a well known and approved drug. Given the excellent safety profile of this drug this could have a significant impact on the treatment of several neurodegenerative disorder.

Our drug network can be useful for formulating hypotheses on the mode of action of previously undescribed compounds by simply measuring multiple transcriptional responses in different cell lines. In addition, drug repositioning is the easiest way to find previously undescribed drug therapies for different conditions. We have shown that it is possible to find previously unrecognized mode of action of well-characterized drugs by simply looking for the drugs neighboring a drug of interest in the network. In addition, by analyzing the prototype ranked lists of genes associated to each drug in the network, we may identify the drug communities that consistently up-, or down-regulate a given set of genes, thus hinting to drug classes able to modulate a specific pathway of interest.
The major limitation of our approach is in the limited number of compounds in the network. Because our approach is based on comparing how similar two drugs are, if a compound is not similar to any of the drugs in the network, no inference on its mode of action or its biological effects can be done.
Moreover, for a compound having inconsistent effects on different cell lines (for example, due to a cell line with a mutated substrate-protein targeted by the compound) merging gene expression profiles from distinct cell lines may dilute the biological effects of the compound. Nevertheless, when no information on the drug mode of action is available

a priori, the best strategy is still to merge profiles from multiple cell lines. We have evidences, that merging profiles coming from a sufficiently large, even if heterogeneous, pool of treated cell lines, provides a summary of the transcriptional response to the drug that can still be well classified by the DN.

We have made our approach publicly available as an online tool that has been enthusiastically welcomed by the international scientific community as proven by the increasing number of MANTRA user account requests and the number of citations in reviews and papers appeared on peer reviewed journals in the fields of computational biology, statistics and network theory (3, 12, 21, 46, 69, 77, 99, 101, 137).

The drug network can be easily searched for a compound of interest, or queried with the transcriptional responses of a unique compound, thus providing a valuable tool for computational drug discovery and repositioning.

References

[1] Abal, M., Bras-Goncalves, R., Judde, J. G. et al. 2004. *Enhanced sensitivity to irinotecan by Cdk1 inhibition in the p53-deficient HT29 human colon cancer cell line*. **Oncogene**. 23, 1737 - 1744.
Cited in page(s): 136

[2] Akaike, H. 1969. *Fitting autoregressive models for prediction*. **Annals of the institute of statistical mathematics**. 21, 243 - 247.
Cited in page(s): 41

[3] Amato, R., Pinelli, M., Monticelli, A., et al. 2009. *Genome-Wide Scan for Signatures of Human Population Differentiation and Their Relationship with Natural Selection, Functional Pathways and Diseases*. **PLoS ONE**. 11, e7927.
Cited in page(s): 172

[4] Amente, S., Gargano, B., Napolitano, G. et al. 2009. *Camptothecin releases P-TEFb from the inactive 7SK snRNP complex*. **Cell Cycle**. 8, 1249 - 1255.
Cited in page(s): 136

[5] Arany, Z., Wagner, B.K., Ma, Y., et al. 2008. *Gene expression-based screening identifies microtubule inhibitors as inducers of PGC-1 and oxidative phosphorylation*. **Proc Natl Acad Sci USA**. 105, 4721 - 4726.
Cited in page(s): 74

[6] Arcamone, F., Cassinelli, G., Fantini, G. et al. 2000. *Adriamycin, 14-hydroxydaunomycin, a new antitumor anti- biotic from S. peucetius var. caesius.*. **Biotechnol Bioeng**. 67, 704 - 713.
Cited in page(s): 124

[7] Ashburner, M., Ball, C. A., Blake, J. A., et al. 2000 *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium*. **Nat Genet**. 25, 25 - 9.
Cited in page(s): 97, 104

[8] Bach, S., Knockaert, M., Reinhardt, J. et al. 2005 *Roscovitine targets, protein kinases and pyridoxal kinase*. **J Biol Chem**. 280, 31208 - 31219.
Cited in page(s): 128

[9] Bansal, M., Belcastro, V., Ambesi-Impiombato, A. et al. 2007 *How to infer gene networks from expression profiles*. **Mol Syst Biol**. 3, 122.
Cited in page(s): 39

[10] Bauer, P. O., Wong, H. K., Oyama, F. et al. 2009 *Inhibition of Rho kinases enhances the degradation of mutant huntingtin*. **J Biol Chem**. 280, 31208 - 31219.
Cited in page(s): 156, 158

[11] Ben-Hur, A., Elisseeff, A., and Guyon, I. 2002 *A stability based method for discovering structure in clustered data*. **Pac Symp Biocomput**. 6 - 17.
Cited in page(s): 32

[12] Berger S. I. and Iyengar, R., I. 2009 *Network analyses in systems pharmacology*. **Bioinformatics**. 25, 2466 - 2472.
Cited in page(s): 35, 37, 172

[13] Brasca, M. G., Albanese, C., Alzani, R. et al. 2010 *Optimization of 6,6-dimethyl pyrrolo[3,4-c]pyrazoles: Identification of PHA-793887, a potent CDK inhibitor suitable for intravenous dosing*. **Bioorg Med Chem**. 18, 1844 - 1853.
Cited in page(s): 124, 129

[14] Brasca, M. G., Amboldi, N., Ballinari, D. et al. 2009 *Identification of N,1,4,4-Tetramethyl-8-{[4-(4-methylpiperazin-1-yl-phenylamino}-4,5-dihydro-1H-pyrazolo[4,3-h]quinazoline-3-carboxamide (PHA- 848125), a Potent, Orally Available Cyclin Dependent Kinase Inhibitor*. **J Med Chem**. 52, 5152 - 5163.
Cited in page(s): 124, 129

[15] Brazma, A., Parkinson, H., Sarkans, U. et al. 2003 *ArrayExpress–a public repository for microarray gene expression data at the EBI*. **Nucleic Acids Res**. 31, 68 - 71.
Cited in page(s): 164

[16] Bursch, W., Ellinger, A., Kienzl, H. et al. 1996 *Active cell death induced by the anti-estrogens tamoxifen and ICI 164 384 in human mammary carcinoma cells (MCF-7) in culture: the role of autophagy*. **Carcinogenesis**. 17, 1595.
Cited in page(s): 153

[17] Califano, A., Stolovitzky, G., and Tu, Y. 2000 *Analysis of gene expression microarrays for phenotype classification*. **Proc Int Conf Intell Syst Mol Biol**. 8, 75 - 85.
Cited in page(s): 48

[18] Campillos, M., Kuhn, M., Gavin, A. C. et al. 2008 *Drug target identification using side-effect similarity*. **Science**. 321, 263 - 6.
Cited in page(s): 35

[19] Cantone, I., Marucci, L., Iorio, F. et al. 2009 *A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches*. **Cell**. 137, 172 - 81.
Cited in page(s): 42

[20] Chang, H. Y., Nuyten, D. S., Sneddon, J. B. et al. 2005 *Robustness, scalability, and integration of a wound-response gene expression signature in pre-dicting breast cancer survival*. **Proc Natl Acad Sci U S A**. 102, 3738 - 3743.
Cited in page(s): 56

[21] Chesler, E. J. and Baker, E. J. 2010 *The importance of open-source integrative genomics to drug discovery*. **Curr Opin Drug Discov Devel**. 13, 310 - 316.
Cited in page(s): 172

[22] Cormen, T.H., Leiserson, C.E., and Rivest, R. 1990. *Mininimum Spanning Trees*, in: **Introduction to algorithms**. MIT Press, Cambridge, MA.
Cited in page(s): 60, 61

[23] Cox, T. and Cox, M. 1984 **Multidimensional Scaling**. Chapman and Hall, London.
Cited in page(s): 30

REFERENCES

[24] Crevel, G., Bates, H., Huikeshoven, H. et al. 2001 *The Drosophila Dpit47 protein is a nuclear Hsp90 co-chaperone that interacts with DNA polymerase alpha*. **J Cell Sci**. 114, 2015 - 25.
Cited in page(s):

[25] Criollo, A., Maiuri, M. C., Tasdemir, E. et al. 2007 *Regulation of autophagy by the inositol trisphosphate receptor*. **Cell Death Differ**. 14, 1029 - 1039.
Cited in page(s): 153

[26] Croons, V., Martinet, W., Herman, A. G., et al. 2009 *The Protein Synthesis Inhibitor Anisomycin Induces Macrophage Apoptosis in Rabbit Atherosclerotic Plaques through p38 Mitogen-Activated Protein Kinase*. **J Pharmacol Exp Ther**. 329, 856 - 864.
Cited in page(s): 76

[27] Csermely, P., Schnaider, T., Soti, C. et al. 1998 *The 90-kDa molecular chaperone family: structure, function, and clinical applications. A comprehensive review*. **Pharmacol Ther**. 79, 129 - 68.
Cited in page(s): 126

[28] Cullen B. R. 2005 *RNAi the natural way*. **Nat Genet**. 37, 1163 - 5.
Cited in page(s): 17

[29] Cvek, B. and Dvorak, Z. 2008 *The value of proteasome inhibition in cancer. Can the old drug, disulfiram, have a bright new future as a novel proteasome inhibitor?*. **Drug Discov Today**. 13, 716 - 722.
Cited in page(s): 147

[30] Dairam, A., Antunesa, E. M., Saravanan, K.S., et al. 2006 *Non-steroidal anti-inflammatory agents, tolmetin and sulindac, inhibit liver tryptophan 2,3-dioxygenase activity and alter brain neurotransmitter levels*. **Life Sciences**. 79, 2269 - 2274.
Cited in page(s): 170

[31] de Medina, P., Silvente-Poirot, S. and Poirot, M. 2009 *Tamoxifen and AEBS ligands induced apoptosis and autophagy in breast cancer cells through the stimulation of sterol accumulation*. **Autophagy**. 5, 1066 - 1067.
Cited in page(s): 153

[32] De Maio, A. 1999 *Heat shock proteins: facts, thoughts, and dreams*. **Shock**. 11, 1 - 12.
Cited in page(s): 126

[33] Diaconis, P., and Graham, R. 1977. *Spearmans foot-rule as a measure of disarray*. **J R Statist Soc**. 39, 262 - 268.
Cited in page(s): 60

[34] di Bernardo, D., Thompson, M. J., Gardner, T. S. et al. 2005. *Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks*. **Nat Biotechnol**. 23, 377 - 383.
Cited in page(s): 39, 147

[35] Eccles, S. A., Massey, A., Raynaud, F. I. et al. 2008 *NVP-AUY922: A novel heat shock protein 90 inhibitor active against xenograft tumor growth, angiogenesis, and metastasis*. **Cancer Res**. 68, 2850 - 2860.
Cited in page(s): 124

[36] Edgar, R., Domrachev, M., and Lash, A. E. 2002 *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository*. **Nucleic Acids Res**. 30, 207 - 10.
Cited in page(s):

[37] Eisen, M., P.T. Spellman, P.O. Brown et al. 1998 *Genetics cluster analysis and display of genome-wide expression patterns*. **Proc Natl Acad Sci USA**. 95, 14863 - 14868.
Cited in page(s): 39

[38] Ennis, H. L., and Lubin, M. 1964. *Cycloheximide: Aspects of Inhibition of Protein Synthesis in Mammalian Cells*. **Science**. 146, 1474 - 1476.
Cited in page(s): 74

[39] Farmer, P., Bonnefoi, H., Anderle, P. et al. 2009 *A stroma-related gene signature predicts resistance to neoadjuvant chemotherapy in breast cancer*. **Nat Med**. 15, 68 - 74.
Cited in page(s): 56

[40] Fire, A., Xu, S., Montgomery, M. K. et al. 1998 *Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans*. **Nature**. 391, 806 - 811.
Cited in page(s): 17

[41] Fisher, L. M. and Pan, X. S. 2008 *Methods to assay inhibitors of DNA gyrase and topoisomerase IV activities*. **Methods Mol Med**. 142, 11 - 23.
Cited in page(s): 133

[42] Fogliatto, G. et al. 2009 *Identification of a potent and specific inhibitor of Hsp90 showing in vivo efficacy*. **American Association for Cancer Reasearch - Annual Meeting**. Poster 37 (Abstract #4685).
Cited in page(s): 124

[43] Fortunato, S. 2009 *Community detection in graphs*. **Phys Rep**. 486, 75-174.
Cited in page(s): 30

[44] Frey, B. J., and Dueck D. 2007 *Clustering by passing messages between data points*. **Science**. 315, 972 - 976.
Cited in page(s): 80, 87

[45] Gardner, T. S., di Bernardo, D., Lorenz, D. 2003 *Inferring genetic networks and identifying compound mode of action via expression profiling*. **Science**. 301, 102 - 105.
Cited in page(s): 39, 47

[46] Genest, C., Neslehova, J., and Ben Ghorbal, N. 2010 *Spearman's footrule and Gini's gamma: a review with complements*. **J Nonparametr Stat**. 22, 937 - 954.
Cited in page(s): 172

[47] Gentleman, R. C., Carey, V. J., Bates, D. M. et al. 2004 *Bioconductor: Open software development for computational biology and bioinformatics*. **Genome Biol**. 5, R80.
Cited in page(s): 139

[48] Girvan, M., and Newman, M.E. 2002 *Community structure in social and biological networks*. **Proc Natl Acad Sci USA**. 99, 7821 - 7826.
Cited in page(s): 30, 79, 80

[49] Goetz, M. P., Toft, D. O., Ames, M. M. et al. 2003 *The Hsp90 chaperone complex as a novel target for cancer therapy*. **Ann onc**. 14, 1169.
Cited in page(s): 127

[50] Greenfield, J. P., Leung, L. W. , Cai, D. et al. 2002 *Estrogen lowers Alzheimer beta-amyloid generation by stimulating trans-Golgi network vesicle biogenesis*. **J Biol Chem**. 277, 12128 - 36.
Cited in page(s): 111

[51] Grollman, A. P. 1967. *Inhibitors of protein biosynthesis. II. Mode of action of anisomycin.* **J Biol Chem**. 242, 3226 - 33.
Cited in page(s): 76

[52] Hauptman, P. J., Garg, R., and Kelly, R. A. 1999. *Cardiac glycosides in the next millennium.* **Prog Cardiovasc Dis**. 41: 247 - 254.
Cited in page(s): 74

[53] Hall, L. H., and Kier, L. B. 2000 *The E-state as the basis for molecular structure space definition and structure similarity.* **J Chem Inf Comp Sci**. 31, 784 - 791.
Cited in page(s): 102

[54] Hall, L. H., Mohney, B., and Kier, L. B. 1991 *The electrotopological state: structure information at the atomic level for molecular graphs.* **J Chem Inf Comp Sci**. 31, 76 - 82.
Cited in page(s): 102

[55] Hansen, N. T., Brunak, S., and Altman, R. B. 2009 *Generating genome-scale candidate gene lists for pharmacogenomics.* **Clin Pharmacol Ther**. 86, 183 - 9.
Cited in page(s): 35

[56] Hassane, D. C., Guzman, M. L., Corbett, C. et al. 2008 *Discovery of Agents that Eradicate Leukemia Stem Cells Using an In Silico Screen of Public Gene Expression Data.* **Blood**. 111, 5654 - 5662.
Cited in page(s): 50

[57] Hehner, S. P., Hofmann, T. G., and Droge, W. 1999 *The antiinflammatory sesquiterpene lactone parthenolide inhibits NF-kappa B by targeting the I kappa B kinase complex.* **J Immunol**. 163, 5617 - 5623.
Cited in page(s): 147

[58] Hieronymus, H., Lamb, J., Ross., K. N. et al. 2006 *Gene expression signature-based chemical genomic prediction identifies novel class of HSP90 pathway modulators.* **Cancer Cell**. 10, 321 - 330.
Cited in page(s): 50

[59] Hollander, M. and Wolfe, D. A. 1999 **Nonparametric Statistical Methods**. Wiley, New York.
Cited in page(s): 49

[60] Hooper, S.D. and Bork, P. 2005 *Medusa: a simple tool for interaction graph analysis.* **Bioinformatics**. 21, 4432 - 4433.
Cited in page(s): 119

[61] Hopkins, A. L. 2008 *Network pharmacology: The next paradigm in drug discovery.* **Nat Chem Biol**. 4, 682 - 690.
Cited in page(s): 37

[62] Hu, G. and Agarwal, P. 2009 *Human disease-drug network based on genomic expression profiles.* **PLoS One**. 4, e6536.
Cited in page(s): 47, 144

[63] Hu, Y., Lounkine, and E., Bajorath, J. 2009 *Improving the search performance of extended connectivity fingerprints through activity-oriented feature filtering and application of a bit-density-dependent similarity function.* **ChemMedChem**. 4, 540 - 548.
Cited in page(s): 102

[64] Huentelman, M. J., Stephan, D. A., Talboom, J. et al. 2009 *Peripheral delivery of a ROCK inhibitor improves learning and working memory.* **Behav neurosci**. 123, 218
Cited in page(s): 160

[65] Hughes, T. R., Marton, M. J., Jones, A. R. et al. 2000 *Functional discovery via a compendium of expression profiles.* **Cell**. 102, 109 - 126
Cited in page(s): 47

[66] Iorio, F., Bosotti, R., Schacheri, E., et al. 2010 *Discovery of drug mode of action and drug repositioning from transcriptional responses.* **Proc Natl Acad Sci USA**. 107, 14621 - 14626.
Cited in page(s): 119, 121, 122

[67] Iorio, F., Isacchi, A., di Bernardo, D., et al. 2010 *Identification of small molecules enhancing autophagic function from drug network analysis.* **Autophagy**. 6, (in press).
Cited in page(s): 121

[68] Irizarry, R. A., Hobbs, B., Collin, F., et al. 2003 *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* **Biostatistics**. 4, 249 - 264.
Cited in page(s): 139

[69] Iskar, M., Campillos, M., Kuhn, M., et al. 2010 *Drug-Induced Regulation of Target Expression.* **PLoS Comput Biol**. 6, e1000925
Cited in page(s): 172

[70] Jensen, B. F., Vind, C., Padkjaer, S. B., et al. 2007 *In silico prediction of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors.* **J Med Chem**. 50, 501 - 511.
Cited in page(s): 102

[71] Jolliffe, I. T. 1986 **Principal Component Analysis**. Springer-Verlag.
Cited in page(s): 30

[72] Kabeya, Y., Mizushima, N., Ueno, T. et al. 2000 *LC3, a mammalian homologue of yeast Apg8p, is localized in autophagosome membranes after processing.* **EMBO J**. 19, 5720 - 5728.
Cited in page(s): 152

[73] Karantzali, E., Schulz, H., Hummel, O. et al. 2008 *Histone deacetylase inhibition accelerates the early events of stem cell differentiation: transcriptomic and epigenetic analysis.* **Genome Biol**. 9, 15.
Cited in page(s): 164

[74] Kasner, S. E., and Ganz, M. B. 1992. *Regulation of intracellular potassium in mesangial cells: a fluorescence analysis using the dye.* **Am J Physiol**. 262, F462 - F467.
Cited in page(s): 74

[75] Kawato, Y., Aonuma, M., Hirota, Y. et al. 1991 *Intracellular roles of SN-38, a metabolite of the camptothecin derivative CPT-11, in the antitumor effect of CPT-11.* **Cancer Res**. 51, 4187 - 4191.
Cited in page(s): 124

[76] Keiser, M. J., Roth, B. L., Armbruster, B. N. et al. 2007 *Relating protein pharmacology by ligand chemistry.* **Nat Biotechnol**. 25, 197 - 206.
Cited in page(s): 35

[77] Keiser, M. J., Irwin, J. J., and Shoichet, B. K. 2010 *The Chemical Basis of Pharmacology.* **Biochemistry**. 49, 10267 - 10276.
Cited in page(s): 172

[78] Khatri, P., and Draghici, S. 2005 *Ontological analysis of gene expression data: current tools, limitations, and open problems.* **Bioinformatics**. 21, 3587 - 95.
Cited in page(s): 105

REFERENCES

[79] Lamb, J., Crawford, E. D., Peck, D., et al. 2006. *The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease.* **Science**. 313, 1929 - 1935.
Cited in page(s): 47, 113, 137

[80] Lamb, J. 2007. *The Connectivity Map: A new tool for biomedical research.* **Nat Rev Cancer**. 7, 54 - 60.
Cited in page(s): 47, 113, 137

[81] Lander, E. S. 1999 *Array of hope.* **Nat Genet**. 21, 3 - 4.
Cited in page(s): 4, 58

[82] Lauria, M., Iorio, F., and di Bernardo, D. 2009 *NIRest: A Tool for Gene Network and Mode of Action Inference.* **Annals of the New York Academy of Sciences**. 1158, 257 - 264.
Cited in page(s): 42

[83] Lei, X., Chen, Y., Duet, G. et al. 2006 *Gossypol induces Bax/Bak-independent activation of apoptosis and cytochrome c release via a conformational change in Bcl-2.* **FASEB J**. 20, 2147 - 2149.
Cited in page(s): 153

[84] Lesnick, T. G., Papapetropoulos, S., Mash, D. C. et al. 2007 *A genomic pathway approach to a complex disease: axon guidance and Parkinson disease.* **PLoS Genet**. 3, e98.
Cited in page(s): 168

[85] Li, M., Huang, Y., Ma, A. A. K. et al. 2009 *Y-27632 improves rotarod performance and reduces huntingtin levels in R6/2 mice.* **Neurobiol Dis**. 36, 413 - 20.
Cited in page(s): 160

[86] Lin, S. 2010. *Space oriented rank-based data integration.* **Stat Appl Genet Mol Biol**. 9, Article20.
Cited in page(s): 60

[87] Liu, R., Wang, X., Chen, G. Y., et al. 2007 *The prognostic role of a gene signature from tumorigenic breast-cancer cells.* **N Engl J Med**. 356, 217 - 226.
Cited in page(s): 56

[88] Liu, W. and Zhang, R. 1998 *Upregulation of p21WAF1/CIP1 in human breast cancer cell lines MCF-7 and MDA-MB-468 undergoing apoptosis induced by natural product anticancer drugs 10-hydroxycamptothecin and camptothecin through p53-dependent and independent pathways.* **Int J Oncol**. 12, 793 - 804.
Cited in page(s): 136

[89] Low, J., Chakravartty, A., Blosser, W. et al. 2009 *Phenotypic Fingerprinting of Small Molecule Cell Cycle Kinase Inhibitors for Drug Discovery.* **Curr Chem Genomics**. 3, 13 - 21.
Cited in page(s): 44

[90] Lum, J. J., DeBerardinis, R. J., and Thompson, C. B. 2005 *Autophagy in metazoans: cell survival in the land of plenty.* **Nat Rev Mol Cell Biol**. 6, 439 - 448.
Cited in page(s): 150, 152

[91] Macpherson, J. I., Pinney, J. W., and Robertson, D. L. 2009 *JNets: exploring networks by integrating annotation.* **BMC Bioinformatics**. 26, 95
Cited in page(s): 35

[92] Malumbres, M. 2005 *Revisiting the Cdk-centric view of the mam- malian cell cycle.* **Cell Cycle**. 4, 206 - 210.
Cited in page(s): 129

[93] Malumbres, M., and Barbacid, M. 2006 *Is Cyclin D1-CDK4 kinase a bona fide cancer target?.* **Cancer Cell**. 9, 2 - 4.
Cited in page(s): 129

[94] Malumbres, M., and Barbacid, M. 2005 *Mammalian cyclin-dependent kinases.* **Trends Biochem Sci**. 30, 630 - 41.
Cited in page(s): 110, 128

[95] Malumbres, M., and Barbacid, M. 2001 *To cycle or not to cycle: a critical decision in cancer.* **Nat Rev Cancer**. 1, 222 - 231.
Cited in page(s): 128

[96] Mani, K. M., Lefebvre, C., Wang, K. et al. 2008 *A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas.* **Mol Syst Biol**. 4, 169.
Cited in page(s): 37, 147

[97] Margolin, A., Nemenman, I., Basso, K. et al. et al. 2006 *Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* **BMC Bioinformatics**. 7 Suppl 1:S7.
Cited in page(s): 39

[98] Marinetti, G. V., Temple, K., and Stotz, E. 1961 *The in vivo effect of digitoxin on rat heart phosphatides.* **J Lipid Res**. 2, 188 - 90.
Cited in page(s): 111

[99] Mazza, T., Romanel, A., and Jordan, F. 2010 *Estimating the divisibility of complex biological networks by sparseness indices.* **Brief Bioinf**. 11, 364 - 374.
Cited in page(s): 172

[100] McAuley, J. J., da Fontoura Costa, L., and Caetano, Tibrio S. 2007 *Rich-club phenomenon across complex network hierarchies.* **Appl Phys Lett**. 91, id. 084103.
Cited in page(s): 30

[101] Mitsos, A., Melas, I. N., Siminelakis, P., et al. 2009 *Identifying Drug Effects via Pathway Alterations using an Integer Linear Programming Optimization Formulation on Phosphoproteomic Data.* **PLoS Comput Biol**. 5, e1000591.
Cited in page(s): 172

[102] McIntyre, T. A., Han, C., and Davis, C. B. 2009 *Prediction of animal clearance using naive Baye- sian classification and extended connectivity fingerprints.* **Xenobiotica**. 39, 487 - 494.
Cited in page(s): 102

[103] McNaught, K. S., Carrupt, P. A., Altomare, C. et al. 1998 *Isoquinoline derivatives as endogenous neurotoxins in the aetiology of Parkinson's disease.* **Biochem Pharmacol**. 56, 921 - 33.
Cited in page(s): 168

[104] Mitscher, L. A. 2005 *Bacterial topoisomerase inhibitors: quinolone and pyridone antibacterial agents.* **Chem Rev**. 105, 559 - 92
Cited in page(s): 133

[105] Mizushima, N. 2004 *Methods for monitoring autophagy.* **Int J Biochem Cell Biol**. 36, 2491 - 2502.
Cited in page(s): 157

[106] Mueller, B. K., Mack, H., and Teusch, N. 2005 *Rho kinase, a promising drug target for neurological disorders.* **Nat Rev Drug Discov**. 4, 387 - 98.
Cited in page(s): 160

[107] Nevins, J. R. 2001 *The Rb/E2F pathway and cancer*. **Hum Mol Genet**. 10, 699 - 703.
Cited in page(s): 146

[108] Nevins, J. R. and Potti, A. 2007 *Mining gene expression profiles: expression signatures as cancer phenotypes*. **Nat Rev Genet**. 8, 601 - 609.
Cited in page(s): 56

[109] Newman, M. E. J. 2004 *Detecting community structure in networks*. **Eur Phys J**. 38, 321 - 330.
Cited in page(s): 30

[110] Ogata, M., Hino, S., Saito, A. et al. 2006 *Autophagy is activated for cell survival after endoplasmic reticulum stress*. **Mol Cell Biol**. 26, 9220 - 9231.
Cited in page(s): 153

[111] Okada, M., Itoh, H., Hatakeyama, T. et al. 2003 *Hsp90 is a direct target of the anti-allergic drugs disodium cromoglycate and amlexanox*. **Biochem J**. 374, 433.
Cited in page(s): 127

[112] Pahor, M., Chrischilles, E. A., Guralnik, J. M., et al. 1994. *Drug data coding and analysis in epidemiologic studies*. **Eur J Epidemiol**. 10, 405 - 411.
Cited in page(s): 66, 97

[113] Pallet, N., Bouvier, N., Legendre, C., et al. 2008. *Autophagy protects renal tubular cells against cyclosporine toxicity*. **Autophagy**. 4, 783 - 791.
Cited in page(s): 153

[114] Parkinson, H., Sarkans, U., Kolesnikov, N. et al. 2010. *ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments*. **Nucleic Acids Res**. Nov 10. [Epub ahead of print]
Cited in page(s): 164

[115] Piccioni, F., Roman, B. R., Fischbeck, K. H., et al. 2004 *A screen for drugs that protect against the cytotoxicity of polyglutamine-expanded androgen receptor*. **Hum Mol Genet**. 13, 437 - 446.
Cited in page(s): 76

[116] Poole, M. C., Easley, C. S., and Hodson, C. A. 1991 *Alteration of the mammotroph Golgi complex by the dopamine agonist 2 Br-alpha- ergocryptine (CB-154) in ovariectomized estrogen primed rats*. **Anat Rec**. 231, 339 - 46.
Cited in page(s): 111

[117] Qu, X. A., Gudivada, R. C., Jegga, A. G. et al. 2009 *Inferring novel disease indications for known drugs by semantically linking drug action and disease mechanism relationships*. **BMC Bioinformatics**. 6;10 Suppl 5:S4.
Cited in page(s): 35

[118] Ravikumar, B., Stewart, A., Kita, H. et al. 2003 *Raised intracellular glucose concentrations reduce aggregation and cell death caused by mutant huntingtin exon 1 by decreasing mTOR phosphorylation and inducing autophagy*. **Hum Mol Genet**. 12, 985 - 994.
Cited in page(s): 152

[119] Renna, M., Jimenez-Sanchez, M., Sarkar, S. et al. 2010 *Alteration of the mammotroph Golgi complex by the dopamine agonist 2 Br-alpha- ergocryptine (CB-154) in ovariectomized estrogen primed rats*. **J Biol Chem**. 285, 11061 - 11067.
Cited in page(s): 153

[120] Rice, J. J., Tu, Y., and Stolovitzky, G. 2005 *Reconstructing biological networks using conditional correlation analysis*. **Bioinformatics**. 21, 765 - 73
Cited in page(s): 39

[121] Rodgers, J. L., and Nicewander, W. A. 1988 *Thirteen ways to look at the correlation coefficient*. **The American Statistician**. 42, 59 - 66.
Cited in page(s): 102

[122] Rogers, D., Brown, R. D., and Hahn, M. 2005 *Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up*. **J Biomol Screen**. 10, 682 - 686.
Cited in page(s): 102

[123] Rose, W. C., Balke, C. W., Wier, W. G. 1992 *Macroscopic and unitary properties of physiological ion flux through L-type Ca^{2+} channels in guinea-pig heart cells*. **J Physiol**. 456, 277 - 284.
Cited in page(s): 74

[124] Rosenbluth, J. M., Mays, D. J., Pino, M. F et al. 2008 *A Gene Signature-Based Approach Identifies mTOR as a Regulator of p73*. **Mol Cell Biol**. 28, 5951 - 5964
Cited in page(s): 50

[125] Scheibel, T. and Buchner, J. 1998. *The Hsp90 complex–a super-chaperone machine as a novel drug target*. **Biochem pharmacol**. 56, 675 - 682.
Cited in page(s): 127

[126] Schwabe, U. 1995. *ATC-Code*. Wissenschaftliches Institut der AOK, Bonn, Germany.
Cited in page(s): 66, 97

[127] Seiler, K. P., George, G. A., Happ, M. P., et al. 2008 *ChemBank: A small-molecule screening and cheminformatics resource database*. **Nucleic Acids Res**. 36 (Database issue), D351 - 359.
Cited in page(s): 94, 97, 119, 140

[128] Senderowicz, A. M. 1999 *Flavopiridol: the first cyclin-dependent kinase inhibitor in human clinical trials*. **Invest New Drugs**. 17, 313 - 320.
Cited in page(s): 124, 128

[129] Senderowicz, A. M. 2005 *Inhibitors of cyclin-dependent kinase modu- lators for cancer therapy*. **Prog Drug Res**. 63, 183 - 206.
Cited in page(s): 128

[130] Shao, J., Welch, W. J., and Diamond, M. I. 2008 *ROCK and PRK-2 mediate the inhibitory effect of Y-27632 on polyglutamine aggregation*. **FEBS Lett**. 582, 1637 - 42
Cited in page(s): 156, 157

[131] Shapiro, G. I. 2006 *Cyclin-dependent kinase pathways as targets for cancer treatment*. **J Clin Oncol**. 24, 1770 - 1783.
Cited in page(s): 128

[132] Schulte, T. W. and Neckers, L. M. 1998 *The benzoquinone ansamycin 17-allylamino- 17-demethoxygeldanamycin binds to HSP90 and shares important biologic activities with geldanamycin*. **Cancer Chemother Pharmacol**. 42, 273 - 279.
Cited in page(s): 124

[133] Schwartz, G. K., Shah, M. A. 2005 *Targeting the cell cycle: a new approach to cancer therapy*. **J Clin Oncol**. 23, 9408 - 9421.
Cited in page(s): 128

REFERENCES

[134] Smolkin, M., and Gosh, D. 2003 *Cluster stability scores for microarray data in cancer studies.* **BMC Bioinformatics**. 4, 36.
Cited in page(s): 32

[135] Smyth, G. K. 2004 *Linear models and empirical Bayes methods for assessing differential expression in microarray experiments.* **Stat Appl Genet Mol Biol**. 3, Article3.
Cited in page(s): 139

[136] Solit, D. B. and Chiosis, G. 2008 *Development and application of Hsp90 inhibitors.* **Drug Discov Today**. 13, 38 - 43.
Cited in page(s): 127

[137] Stajdohar, M., Mramor, M., Zupan, B., et al. 2010 *FragViz: visualization of fragmented networks.* **BMC Bioinformatics**. 11, 475.
Cited in page(s): 172

[138] Stolovitzky, G., Prill, R. J. and Califano, A. 2009 *Lessons from the DREAM2 Challenges: A Community Effort to Assess Biological Network Inference.* **Annals of the New York Academy of Sciences**. 1158.
Cited in page(s): 42

[139] Storm, M. P., Kumpfmueller, B., Thompson, B., et al. 2009 *Characterization of the phosphoinositide 3-kinase-dependent transcriptome in murine embryonic stem cells: identification of novel regulators of pluripotency.* **Stem Cells**. 27, 764 - 75.
Cited in page(s): 164

[140] Subramanian, A., Tamayo, P., Mootha, V.K., et al. 2005. *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* **Proc Natl Acad Sci USA**. 102, 15545 - 15550.
Cited in page(s): 49, 64

[141] Taldone, T., Sun, W., and Chiosis, G. 2009 *Discovery and development of heat shock protein 90 inhibitors.* **Bioorg Med Chem**. 17, 2225 - 2235.
Cited in page(s): 109

[142] Tettamanti, G., Malagoli, D., Marchesini, E. et al. 2006 *Oligomycin A induces autophagy in the IPLB-LdFB insect cell line.* **Cell Tissue Res**. 326, 179 - 186
Cited in page(s): 153

[143] Tetsu, O., and McCormick, F. 2003 *Proliferation of cancer cells despite CDK2 inhibition.* **Cancer Cell**. 3, 233 - 245.
Cited in page(s): 129

[144] von Mering, C., Jensen, L. J., Snel, B. et al. 2005 *STRING: known and predicted protein-protein associations, integrated and transferred across organisms.* **Nucleic Acids Res**. 33 (Database issue), D433 - 7.
Cited in page(s): 119

[145] Vulpetti, A., Pevarello, P. 2005 *An analysis of the binding modes of ATP-competitive CDK2 inhibitors as revealed by X-ray struc- tures of protein-inhibitor complexes.* **Curr Med Chem Anticancer Agents**. 5, 561 - 573.
Cited in page(s): 128

[146] Weininger, D. 1988 *SMILES, a chemical language and information system.* **J Chem Inf Comp Sci**. 28, 31 - 36.
Cited in page(s): 101

[147] Westfall, P. H. and Young, S. S. 1993 *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment.* (New York, Wiley).
Cited in page(s): 139

[148] Wishart, D. S. 2008 *DrugBank and its relevance to pharmacogenomics.* **Pharmacogenomics**. 9, 1155 - 1162.
Cited in page(s): 94, 97, 102, 119, 140

[149] Wu, Z., Irizarry, R. A., Gentlemant, R. et al. 2004 *A model based background adjustment for oligonucleotide epression arrays.* Johns Hopkins University, Dept. of Biostatistics (Working Paper 1).
Cited in page(s): 139

[150] Yang, Q., Huang, W., Jozwik, C. et al. 2005 *Cardiac glycosides inhibit TNF-alpha/NF-kappaB signaling by blocking recruitment of TNF receptor-associated death domain to the TNF receptor.* **Proc Natl Acad Sci USA**. 102, 9631 - 9636
Cited in page(s): 130

[151] Yang, H., Shi, G., and Dou Q. P. 2007 *The tumor proteasome is a primary target for the natural anticancer compound Withaferin A isolated from Indian winter cherry.* **Mol Pharmacol**. 71, 426 - 437.
Cited in page(s): 147

[152] Yang, K., Bai, H., Ouyang, Q. et al. 2008 *Finding multiple target optimal intervention in disease-related molecular network.* **Mol Syst Biol**. 4, 228.
Cited in page(s): 47

[153] Yu, J., Anne Smith, V. P., Wang, P. et al. 2004. *Advances to bayesian network inference for generating causal networks from observational biological data.* **Bioinformatics**. 20, 3594 - 3603.
Cited in page(s): 39

[154] Zadeh, L.A. 1965 *Fuzzy sets.* **Information and Control**. 8, 338 - 353.
Cited in page(s): 105

[155] Zadeh, L. A., Klir, G. J., and Yuanet B. 1996 *Fuzzy Sets, Fuzzy Logic, Fuzzy Systems: Selected Papers.* **Advances in Fuzzy Systems - Applications and Theory**. World Scientific Press 6.
Cited in page(s): 104

Appendix A

Abbreviations

2DOG 2-deoxy-D-glucose

A2780 Human ovarian cancer cell line

AES Average Enrichment-Score

APC Affinity Propagation Clustering

ATC Anatomical Therapeutic Chemical

ATP Adenosine-5'-triphosphate

AUC Area under the curve

bp base pairs

Ca^{2+} Calcium ions

CDK Cyclin-Dependent kinase

cDNA complementary DNA

cMap Connectivity Map

CPCG Cross-platform conserved gene

CS Connectivity Score

diB-LAB Systems, Synthetic and Computational Biology Laboratory

DN Drug Network

DNA Deoxyribonucleic acid

DREAM Dialogue for Reverse Engineering Assessment and Methods

EMEA European Medicine Agency

ER Endoplasmatic Reticulum

ERK Extracellular signal-regulated kinases

ES Enrichment Score

ESF Electrotopological States

FDA U.S. Food and Drug Administration

FDR False Discovery Rate

FP False Positive

GEO Gene Expression Omnibus database

GEP Gene Expression Profile

GO Gene Ontology

GSEA Gene Set Enrichment Analysis

GSK Glycogen synthase kinase

GTP Guanosine Triphosphate

HDAC Histone deacetylases

Hsp90 Heat Shock Protein 90

HD Huntington's disease

HTS High-Throughput Screening

Htt Huntingtin

IC_{50} Half maximal inhibitory concentration

IRMA In-vivo Reverse-engineering and Modelling Assessment

K^+ Potassium ion

A. ABBREVIATIONS

KRUBOR Kruskal-Borda

KS Kolmogorov-Smirnov

ITES Inverse Total Enrichment Score

LC3 Light chain 3

LIMK LIM kinase

mRNA messenger RNA

MANTRA Mode of Action by Network Analysis

MAP2 Microtubule-Associated Protein 2

MAPK Mitogen-activated protein kinase

MCF7 Human breast cancer cell line

MCL Myeloid Cell Leukemia sequence

MDS MultiDimensional Scaling

MES Maximum Enrichment-Score

MLC Myosin Light Chain

MoA Mode of Action

MPI Microarray Probe-set Identifiers

N-TRAP NeTwork by Recursive Affinity Propagation

Na^+ Sodium ion

Na^+/K^+**-ATPase** Sodium-Potassium pump

NIR Network Inference by multiple Regression

NIRest NIR with perturbation Estimates

NMS Nerviano Medical Science

ODE Ordinary Differential Equation

PARP Poly (ADP-ribose) polymerase

PCA Principal Components Analysis

PCR Polymerase Chain Reaction

PD Parkinson's disease

pdf Probability Density Function

PI3K Phosphoinositide 3-kinases

PPV Positive Predicted Value

PPAR-gamma Peroxisome proliferator-activate receptor gamma

PRL Prototype Ranked List

PVDF Polyvinylidene Fluoride

QC Quality control

RB Retinoblastoma

RNA Ribonucleic acid

ROCK RHO kinase

RMA Robust Multiarray Algorithm

ROC Receiver Operating Characteristic

Rho Ras homolog gene family

RSS Residual Sum of Square

SDD Supplementary Data Disc

Ser/Thr Serine/Threonine

SF539 Human glioblastoma cell line

SMILES Simplified Molecular Input Line Entry Specification

SNP Single Nucleotide Polymorphism

Topo Topoisomerase

TIGEM TeleThon Institute of Genetics and Medicine

tRNA Transfer RNA

TP True Positive

U251 Human glioma cell line

UPS Ubiquitin Proteasome System

WB Western Blot

WHOCC World Health Organization Collaborating Centre

Appendix B

Community enrichments

B.1 Literature based evidences

Community n. 3

Drug name	Class
5109870	
5182598	
resveratrol	G1/S Cell Cycle Blockers
5248896	
ciclopirox	G1/S Cell Cycle Blockers
etoposide	G1/S Cell Cycle Blockers
blebbistatin	
5255229	
trifluridine	
5279552	
5211181	
triamterene	
deferoxamine	G1/S Cell Cycle Blockers
kaempferol	G1/S Cell Cycle Blockers
apomorphine	
colforsin	G1/S Cell Cycle Blockers
quercetin	G1/S Cell Cycle Blockers
guaifenesin	
hycanthone	

Community n. 9

Drug name	Class
BCB000038	
16-phenyltetranorprostaglandin-E2	fatty acids and prostaglandin derivatives
CP-944629	
IC-86621	
11-deoxy-16,16-dimethylprostaglandin-E2	fatty acids and prostaglandin derivatives

Community n. 10

Drug name	Class
CP-319743	
15(S)-15-methylprostaglandin-E2	fatty acids and prostaglandin derivatives
BCB000040	
BCB000039	

Community n. 13

Drug name	Class
DL-thiorphan	
atropine	Anticholinergics

methacholine-chloride	
papaverine	Anticholinergics
protoveratrine-A	
labetalol	
scopolamine	Anticholinergics
flufenamic-acid	
tremorine	Cholinergic Receptor Agonists
bupropion	
demeclocycline	
tomatidine	
hydrastine-hydrochloride	
clonidine	Anticholinergics
diclofenac	
cefuroxime	
sulfadiazine	
physostigmine	Anticholinergic antagonists
gramine	
hydroxyachillin	
ramipril	
testosterone	
lidoflazine	
alpha-yohimbine	
ifosfamide	

Community n. 14

Drug name	Class
GW-8510	CDK2 inhibitors
doxorubicin	TopoII inhibitors
alsterpaullone	CDK2 inhibitors
H-7	CDK2 inhibitors
tyrphostin-AG-825	CDK2 inhibitors
camptothecin	TopoI inhibitors
daunorubicin	TopoII inhibitors
mitoxantrone	TopoII inhibitors
ellipticine	CDK2/TopoII inhibitors
azacitidine	TopoII inhibitors
fisetin	CDK2 inhibitors
staurosporine	CDK2 inhibitors
MS-275	
bromopride	
gallamine-triethiodide	

Community n. 15

Drug name	Class
Gly-His-Lys	carbonic anhydrase inhibitors
pilocarpine	carbonic anhydrase inhibitors
STOCK1N-35874	
diclofenamide	carbonic anhydrase inhibitors

Community n. 16

Drug name	Class
HC-toxin	HDAC inhibitors
vorinostat	HDAC inhibitors
rifabutin	HDAC inhibitors
scriptaid	HDAC inhibitors
trichostatin-A	HDAC inhibitors
valproic-acid	HDAC inhibitors
idoxuridine	
mepacrine	
LY-294002	PI3K inhibitors
bufexamac	
spironolactone	
pergolide	
fusidic-acid	

trimethobenzamide	
rosiglitazone	PI3K inhibitors
diprophylline	

Community n. 22

Drug name	Class
Prestwick-1080	
Prestwick-1100	
nialamide	Antidepressants
levcycloserine	Antidepressants
santonin	
quinethazone	
chlorambucil	
homosalate	
ciclacillin	
lithocholic-acid	
phenelzine	Antidepressants
Prestwick-984	

Community n. 23

Drug name	Class
SB-203580	PGE2 antagonists
NS-398	PGE2 antagonists
pioglitazone	PGE2 antagonists
SC-19220	PGE2 antagonists

Community n. 26

Drug name	Class
aciclovir	
guanabenz	adrenoreceptor and histamine receptor antagonists
lisuride	adrenoreceptor and histamine receptor antagonists
timolol	adrenoreceptor and histamine receptor antagonists
thioperamide	adrenoreceptor and histamine receptor antagonists
finasteride	

Community n. 28

Drug name	Class
alvespimycin	HSP90 inhibitors
geldanamycin	HSP90 inhibitors
monorden	HSP90 inhibitors
tanespimycin	HSP90 inhibitors
fulvestrant	

Community n. 32

Drug name	Class
apigenin	CDK2/TopoII inhibitors
luteolin	CDK2/TopoI inhibitors
chrysin	CDK2 inhibitors
thioguanosine	DNA precursors/antimetabolites (S phase)
harmine	CDK2 inhibitors
skimmianine	TopoI inhibitors
0175029-0000	
rimexolone	
sulfametoxydiazine	antibacterials
trioxysalen	
flunixin	
metyrapone	
cefalotin	antibacterials
irinotecan	TopoI inhibitors
sulfaphenazole	antibacterials
acetylsalicylic-acid	
metacycline	
pancuronium-bromide	
dextromethorphan	
amoxicillin	antibacterials
lymecycline	antibacterials

2,6-dimethylpiperidine	
etidronic-acid	
ethotoin	
lysergol	
vidarabine	antibacterials
Prestwick-664	
Prestwick-665	
heliotrine	
todralazine	
proxymetacaine	
calcium-pantothenate	
tiabendazole	
paroxetine	
tinidazole	antibacterials
harman	CDK2 inhibitors
acacetin	
harmol	CDK2 inhibitors
nifedipine	

Community n. 34

Drug name	Class
astemizole	antihistamines and anticholinergics
terfenadine	antihistamines and anticholinergics
mefloquine	antihistamines and anticholinergics
prenylamine	
suloctidil	
isoconazole	
spiperone	antihistamines and anticholinergics
rescinnamine	
trimipramine	antihistamines and anticholinergics
dihydroergotamine	antihistamines and anticholinergics
nicergoline	
nordihydroguaiaretic-acid	
dosulepin	
chlorphenamine	antihistamines and anticholinergics
proadifen	antihistamines and anticholinergics
aminophenazone	
atropine-oxide	antihistamines and anticholinergics
penbutolol	antihistamines and anticholinergics

Community n. 39

Drug name	Class
carbinoxamine	Histamine receptor H1 antagonists
drofenine	Histamine receptor H1 antagonists
isopropamide-iodide	
ribostamycin	Aminoglicosidic Antibiotics
butirosin	Aminoglicosidic Antibiotics

Community n. 40

Drug name	Class
celastrol	Proteasome inhibitors and UPS modulators
MG-132	Proteasome inhibitors and UPS modulators
5224221	
MG-262	Proteasome inhibitors and UPS modulators
thapsigargin	Proteasome inhibitors and UPS modulators
puromycin	protein synthesis inhibitors (elongation inhibitors)
ionomycin	calcium signal modulators
piperlongumine	
disulfiram	Proteasome inhibitors and UPS modulators
5253409	
1,4-chrysenequinone	
mometasone	Proteasome inhibitors and UPS modulators
fendiline	calcium signal modulators
tyrphostin-AG-1478	

econazole	calcium signal modulators
cyproheptadine	calcium signal modulators
isotretinoin	
(+)-chelidonine	
lynestrenol	
primaquine	protein synthesis inhibitors (elongation inhibitors)
dienestrol	
pimethixene	
loxapine	
PF-00875133-00	
5230742	
dicycloverine	
mianserin	
dyclonine	
nalbuphine	
Y-27632	
betazole	
vinburnine	
propidium-iodide	
prilocaine	
PHA-00846566E	

Community n. 42

Drug name	Class
chlorzoxazone	
glibenclamide	
clindamycin	antibiotics and bactericidals
dirithromycin	antibiotics and bactericidals
lobeline	
chlortetracycline	antibiotics and bactericidals
danazol	
clopamide	
ajmaline	
ampyrone	
betaxolol	
chlorhexidine	antibiotics and bactericidals
methazolamide	
hydrastinine	
Prestwick-689	

Community n. 43

Drug name	Class
ciclosporin	
estrone	Estrogens
diethylstilbestrol	Estrogens
pizotifen	
equilin	Estrogens
naringenin	Estrogen inhibitors
betulinic-acid	
saquinavir	
MK-886	

Community n. 44

Drug name	Class
indometacin	Dopaminergic agents
dopamine	Dopaminergic agents
quinpirole	Dopaminergic agents
cobalt-chloride	

Community n. 48

Drug name	Class
demecolcine	plant alkaloids
(-)-catechin	plant alkaloids
12,13-EODE	plant alkaloids
phenanthridinone	plant alkaloids

DL-PPMP	
3-hydroxy-DL-kynurenine	plant alkaloids
5252917	
paclitaxel	plant alkaloids
chloroquine	plant alkaloids

Community n. 49

Drug name	Class
dexverapamil	L-type calcium channel blockers
exemestane	Aromatase inhibitors
4,5-dianilinophthalimide	
mesalazine	COX2 inhibitors
rofecoxib	COX2 inhibitors
verapamil	L-type calcium channel blockers
kanamycin	
midecamycin	Estrogens (PGE2 increasers)
mepyramine	
alpha-estradiol	Estrogens (PGE2 increasers)
troleandomycin	Estrogens (PGE2 increasers)
racecadotril	

Community n. 50

Drug name	Class
dicoumarol	antibacterials
benfotiamine	
spiramycin	antibacterials
sulfadimidine	antibacterials
ethoxyquin	antibacterials
flecainide	
piperacillin	antibacterials

Community n. 52

Drug name	Class
doxazosin	Alpha-adrenoreceptor modulators
carbachol	Alpha-adrenoreceptor modulators
rolitetracycline	Antibacterials for systemic use
ethaverine	
xylazine	Alpha-adrenoreceptor modulators
Prestwick-860	
ioversol	
betahistine	
cinchocaine	
natamycin	Antibacterials for systemic use
meropenem	Antibacterials for systemic use
piromidic-acid	Antibacterials for systemic use
progesterone	Progestogen Hormons
levamisole	
pivampicillin	Antibacterials for systemic use
esculin	
naringin	
fluorocurarine	
crotamiton	Alpha-adrenoreceptor modulators
cinoxacin	Antibacterials for systemic use
cefsulodin	Antibacterials for systemic use
dimethadione	
bephenium-hydroxynaphthoate	
dydrogesterone	Progestogen Hormons
edrophonium-chloride	
lomefloxacin	Antibacterials for systemic use
bemegride	
palmatine	
bendroflumethiazide	
cisapride	

Community n. 53

Drug name	Class
emetine	protein synthesis inhibitors
cephaeline	protein synthesis inhibitors
anisomycin	protein synthesis inhibitors
cicloheximide	protein synthesis inhibitors
vanoxerine	
propofol	

Community n. 59

Drug name	Class
flunisolide	corticosteroids
citiolone	
mepenzolate-bromide	
eldeline	
solasodine	corticosteroids
pempidine	
halcinonide	corticosteroids
fludroxycortide	corticosteroids
alcuronium-chloride	

Community n. 60

Drug name	Class
gefitinib	PI3Ks inhibitors
1,5-isoquinolinediol	PARP inhibitors
clomipramine	
tolfenamic-acid	PARP inhibitors
famotidine	
chlorphenesin	
guanethidine	
vinpocetine	
wortmannin	PI3Ks inhibitors
SC-58125	PI3Ks inhibitors
bambuterol	
nimesulide	COX2 modulators
amodiaquine	
hexamethonium-bromide	
tetracycline	COX2 modulators
acebutolol	
josamycin	

Community n. 62

Drug name	Class
gossypol	Sodium/Calcium Decreasers and calcium channel blockers
pararosaniline	
niclosamide	Antiinfectives, Antiseptics, Antiparasitics
pyrvinium	Antiinfectives, Antiseptics, Antiparasitics
valinomycin	
rottlerin	
clotrimazole	Antiinfectives, Antiseptics, Antiparasitics
5707885	
dequalinium-chloride	Antiinfectives, Antiseptics, Antiparasitics
miconazole	Antiinfectives, Antiseptics, Antiparasitics
butoconazole	Antiinfectives, Antiseptics, Antiparasitics
benzamil	Sodium/Calcium Decreasers and calcium channel blockers
antimycin-A	
azacyclonol	
clioquinol	Antiinfectives, Antiseptics, Antiparasitics
felodipine	Sodium/Calcium Decreasers and calcium channel blockers
ticlopidine	Antiplatelets and vasoprotectives
tribenoside	Antiplatelets and vasoprotectives
abamectin	
erastin	
mifepristone	
clofazimine	

profenamine	
LM-1685	
rotenone	Antiinfectives, Antiseptics, Antiparasitics
0179445-0000	
isradipine	Sodium/Calcium Decreasers and calcium channel blockers
naftifine	Antiinfectives, Antiseptics, Antiparasitics
mercaptopurine	
foliosidine	

Community n. 63

Drug name	Class
helveticoside	Na+/K+ - ATPase (sodium Potassium) membrane pump inhibitors
lanatoside-C	Na+/K+ - ATPase (sodium Potassium) membrane pump inhibitors
digoxin	Na+/K+ - ATPase (sodium Potassium) membrane pump inhibitors
ouabain	Na+/K+ - ATPase (sodium Potassium) membrane pump inhibitors
digoxigenin	Na+/K+ - ATPase (sodium Potassium) membrane pump inhibitors
proscillaridin	Na+/K+ - ATPase (sodium Potassium) membrane pump inhibitors
digitoxigenin	Na+/K+ - ATPase (sodium Potassium) membrane pump inhibitors
strophanthidin	Na+/K+ - ATPase (sodium Potassium) membrane pump inhibitors
bisacodyl	Na+/K+ - ATPase (sodium Potassium) membrane pump inhibitors
lycorine	Na+/K+ - ATPase (sodium Potassium) membrane pump inhibitors
hydroquinine	Na+/K+ - ATPase (sodium Potassium) membrane pump inhibitors

Community n. 65

Drug name	Class
imatinib	COX2 modulators
pirinixic-acid	COX2 modulators
celecoxib	COX2 modulators

Community n. 69

Drug name	Class
mebendazole	Microtubule inhibitors
nocodazole	Microtubule inhibitors
colchicine	Microtubule inhibitors
fenbendazole	Microtubule inhibitors
dobutamine	Alpha and Beta Adrenergic receptor modulators
scoulerine	Alpha and Beta Adrenergic receptor modulators
nifuroxazide	
(+)-isoprenaline	Alpha and Beta Adrenergic receptor modulators
amiloride	
oxedrine	Alpha and Beta Adrenergic receptor modulators

Community n. 73

Drug name	Class
methylergometrine	dopamine receptors agonists
bromocriptine	dopamine receptors agonists
diltiazem	dopamine receptors agonists
fenoterol	Alfa and Beta adrenergic modulators
alfuzosin	Alfa and Beta adrenergic modulators
hesperidin	Alfa and Beta adrenergic modulators
orciprenaline	Alfa and Beta adrenergic modulators
(-)-isoprenaline	Alfa and Beta adrenergic modulators
dexibuprofen	Alfa and Beta adrenergic modulators

Community n. 75

Drug name	Class
mycophenolic-acid	hepatic enzymes inducers
methotrexate	hepatic enzymes inducers
ribavirin	hepatic enzymes inducers
rifampicin	hepatic enzymes inducers

Community n. 77

Drug name	Class
nitrendipine	calcium channel blockers
paracetamol	
nimodipine	calcium channel blockers
oxybenzone	

Community n. 81

Drug name	Class
pirenperone	serotonin receptors modulators / antiparkinsonians
serotonin	serotonin receptors modulators / antiparkinsonians
epiandrosterone	
biperiden	serotonin receptors modulators / antiparkinsonians
propranolol	serotonin receptors modulators / antiparkinsonians

Community n. 88

Drug name	Class
ritodrine	
androsterone	Steroids hormons
ketoconazole	Steroid hormon synthesis inhibitors, cythochrome P450 blockers, domperidone metabolism blocker
domperidone	P450 substrates
procainamide	P450 substrates
clenbuterol	P450 substrates

Community n. 89

Drug name	Class
sanguinarine	Hemostatic agents
cantharidin	
8-azaguanine	
verteporfin	
ginkgolide-A	Hemostatic agents
talampicillin	
menadione	Hemostatic agents
ipratropium-bromide	
hydrocotarnine	
dacarbazine	
etamsylate	Hemostatic agents
solanine	
dipyridamole	Hemostatic agents
N-acetyl-L-leucine	
desoxycortone	
epivincamine	
zimeldine	
tracazolate	
pargyline	Antihypertensive agents
sitosterol	Antihypertensive agents
picrotoxinin	
6-benzylaminopurine	
altizide	Antihypertensive agents
terbutaline	
ketoprofen	Hemostatic agents
laudanosine	
phentolamine	Antihypertensive agents
tolbutamide	
flumequine	
oxytetracycline	
sotalol	Antihypertensive agents
methyldopa	Antihypertensive agents
R-atenolol	Antihypertensive agents

Community n. 90

Drug name	Class
sulconazole	Imidazoles and Sulfonamides
phthalylsulfathiazole	Imidazoles and Sulfonamides
cetirizine	
medrysone	Corticosteroids
tyloxapol	
Prestwick-1084	
omeprazole	Imidazoles and Sulfonamides
promethazine	
famprofazone	

lorglumide	
metamizole-sodium	Imidazoles and Sulfonamides
eucatropine	
pentoxyverine	
cefalexin	
liothyronine	
piperidolate	
amiodarone	
tropicamide	
riboflavin	
acepromazine	
latamoxef	
hexestrol	
meclozine	
monobenzone	
cyproterone	
(-)-MK-801	
ketanserin	
etofenamate	
trazodone	
pyrazinamide	
sulfafurazole	Imidazoles and Sulfonamides
ronidazole	Imidazoles and Sulfonamides
dexpropranolol	
capsaicin	
apramycin	
denatonium-benzoate	
nitrofurantoin	
naloxone	
tranylcypromine	
norethisterone	
fluocinonide	Corticosteroids
ifenprodil	
oxybuprocaine	
sulfamerazine	Imidazoles and Sulfonamides
cycloserine	
beta-escin	
clobetasol	Corticosteroids
xylometazoline	
methapyrilene	
scopoletin	
Prestwick-674	
bezafibrate	
sertaconazole	Imidazoles and Sulfonamides
difenidol	
anabasine	
ganciclovir	
sulfamethoxazole	Imidazoles and Sulfonamides
adrenosterone	
fluorometholone	Corticosteroids
cortisone	Corticosteroids
mevalolactone	
flunarizine	
pentamidine	
aminophylline	
zoxazolamine	
ranitidine	
parbendazole	Imidazoles and Sulfonamides
colecalciferol	
ethosuximide	
clorgiline	

furaltadone	
etanidazole	Imidazoles and Sulfonamides
sulfinpyrazone	
pipenzolate-bromide	
nifurtimox	
methylprednisolone	Corticosteroids
alverine	
prasterone	
bretylium-tosilate	

Community n. 91

Drug name	Class
sulindac-sulfide	sulindac metabolites
sulindac	sulindac metabolites
exisulind	sulindac metabolites
tacrolimus	
phenyl-biguanide	

Community n. 93

Drug name	Class
tiapride	
ofloxacin	TopoII inhibitors
nalidixic-acid	TopoII inhibitors
griseofulvin	
metaraminol	
flutamide	

Community n. 96

Drug name	Class
tolazoline	
sulfaguanidine	sulfonamides
succinylsulfathiazole	sulfonamides
minoxidil	
cyclopenthiazide	sulfonamides

Community n. 99

Drug name	Class
trichlormethiazide	non-selective phosphodiesterase inhibitors and adenosine receptor antagonists
reserpine	
mestranol	
galantamine	
fosfosal	non-selective phosphodiesterase inhibitors and adenosine receptor antagonists
atovaquone	
spectinomycin	
simvastatin	non-selective phosphodiesterase inhibitors and adenosine receptor antagonists
trimethylcolchicinic-acid	
canavanine	
theophylline	non-selective phosphodiesterase inhibitors and adenosine receptor antagonists
theobromine	non-selective phosphodiesterase inhibitors and adenosine receptor antagonists
arcaine	
colistin	

Community n. 100

Drug name	Class
trifluoperazine	Antipsychotics (Phenothiazines)
metergoline	Antipsychotics (5-HT and Dopamine Receptors modulators)
perphenazine	Antipsychotics (Phenothiazines)
loperamide	
methylbenzethonium-chloride	
alexidine	
calmidazolium	
chlorprothixene	Antipsychotics (Thioxanthene derivatives)

B. COMMUNITY ENRICHMENTS

tetrandrine	calcium channel blockers and Ca2+ level increaser
protriptyline	Antidepressants (Non-selective monoamine reuptake inhibitors)
fluphenazine	Antipsychotics (Phenothiazines)
0297417-0002B	
cloperastine	Antipsychotics (Thioxanthene derivatives)
benzethonium-chloride	
nortriptyline	Antidepressants (Non-selective monoamine reuptake inhibitors)
fluspirilene	Antipsychotics (Diphenylbutylpiperidine derivatives)
homochlorcyclizine	
prochlorperazine	Antipsychotics (Phenothiazines)
maprotiline	Antidepressants (Non-selective monoamine reuptake inhibitors)
BW-B70C	
levomepromazine	Antipsychotics (Phenothiazines)
desipramine	Antidepressants (Non-selective monoamine reuptake inhibitors)
metitepine	Antipsychotics (5-HT and Dopamine Receptors modulators)
bepridil	calcium channel blockers and Ca2+ level increaser
chlorcyclizine	Antipsychotics (Thioxanthene derivatives)
cytochalasin-B	
piperacetazine	Antipsychotics (Phenothiazines)
norcyclobenzaprine	Antidepressants (Non-selective monoamine reuptake inhibitors)
ivermectin	
tonzonium-bromide	
pimozide	Antipsychotics (Diphenylbutylpiperidine derivatives)
metixene	Antipsychotics (Thioxanthene derivatives)
perhexiline	calcium channel blockers and Ca2+ level increaser
thioproperazine	Antipsychotics (Phenothiazines)
chlorpromazine	Antipsychotics (Phenothiazines)
thioridazine	Antipsychotics (Phenothiazines)
flupentixol	Antipsychotics (Thioxanthene derivatives)
monensin	Antidepressants (Non-selective monoamine reuptake inhibitors)
clomifene	calcium channel blockers and Ca2+ level increaser
raloxifene	calcium channel blockers and Ca2+ level increaser
hexetidine	
CP-645525-01	
imipramine	Antidepressants (Non-selective monoamine reuptake inhibitors)
amoxapine	Antidepressants (Non-selective monoamine reuptake inhibitors)
quinisocaaine	Antihistamines
clofilium-tosylate	calcium channel blockers and Ca2+ level increaser
zuclopenthixol	Antipsychotics (Thioxanthene derivatives)
amitriptyline	
phenazopyridine	
co-dergocrine-mesilate	
clemastine	Antihistamines
fluvoxamine	Antidepressants (Non-selective monoamine reuptake inhibitors)
ursolic-acid	
clozapine	Antipsychotics (others)
promazine	
podophyllotoxin	
cyclobenzaprine	Antidepressants (Non-selective monoamine reuptake inhibitors)
dihydroergocristine	
5666823	
orphenadrine	Antipsychotics (Thioxanthene derivatives)
S-propranolol	
sirolimus	
albendazole	
troglitazone	
diperodon	
lasalocid	
doxepin	Antidepressants (Non-selective monoamine reuptake inhibitors)
alimemazine	Antihistamines
thiethylperazine	Antihistamines

procarbazine	
fluoxetine	Antidepressants (Non-selective monoamine reuptake inhibitors)
lovastatin	
amphotericin-B	
proguanil	
raubasine	
diphenylpyraline	Antihistamines

Community n. 104

Drug name	Class
withaferin-A	
lomustine	nitrosourea compounds used as alkylating agents in chemotherapy.
F0447-0125	
parthenolide	
phenoxybenzamine	
semustine	nitrosourea compounds used as alkylating agents in chemotherapy.
5155877	
thiostrepton	
STOCK1N-35215	
15-delta-prostaglandin-J2	
securinine	
oxyphenbutazone	
5194442	
etacrynic-acid	
carmustine	nitrosourea compounds used as alkylating agents in chemotherapy.
C-75	
myricetin	

Community n. 105

Drug name	Class
yohimbic-acid	natural occuring alkaloids and derivatives
diflorasone	natural occuring alkaloids and derivatives
Prestwick-691	
gelsemine	natural occuring alkaloids and derivatives
conessine	natural occuring alkaloids and derivatives

B.2 ATC-Codes

Community n. 14

Enriched ATC code	Definition	p-value	Odds-ratio
L01D	CYTOTOXIC ANTIBIOTICS AND RELATED SUBSTANCES	0.0000	192.00
L01DB	Anthracyclines and related substances	0.0000	192.00
L01	ANTINEOPLASTIC AGENTS	0.0001	24.00
L	ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS	0.0005	14.77

Community n. 63

Enriched ATC code	Definition	p-value	Odds-ratio
C01A	CARDIAC GLYCOSIDES	0.0000	153.60
C01AA	Digitalis glycosides	0.0000	153.60
C01	CARDIAC THERAPY	0.0010	12.45
C	CARDIOVASCULAR SYSTEM	0.0491	3.18

Community n. 48

Enriched ATC code	Definition	p-value	Odds-ratio
L01C	PLANT ALKALOIDS AND OTHER NATURAL PRODUCTS	0.0001	128.00
L01	ANTINEOPLASTIC AGENTS	0.0028	21.33
L	ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS	0.0073	13.13

Community n. 65

Enriched ATC code	Definition	p-value	Odds-ratio
L01X	OTHER ANTINEOPLASTIC AGENTS	0.0001	109.71
L01	ANTINEOPLASTIC AGENTS	0.0009	32.00
L	ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS	0.0025	19.69

Community n. 96

Enriched ATC code	Definition	p-value	Odds-ratio
A07AB	Sulfonamides	0.0001	102.40
A07A	INTESTINAL ANTIINFECTIVES	0.0026	23.63
A07	ANTIDIARRHEALS, INTESTINAL ANTIINFLAMMATORY/ANTIINFECTIVE AGENTS	0.0088	12.80
A	ALIMENTARY TRACT AND METABOLISM	0.1137	3.30
C	CARDIOVASCULAR SYSTEM	0.0491	3.18

Community n. 104

Enriched ATC code	Definition	p-value	Odds-ratio
L01AD	Nitrosoureas	0.0000	96.00
L01A	ALKYLATING AGENTS	0.0000	54.86
L01	ANTINEOPLASTIC AGENTS	0.0005	16.00
L	ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS	0.0022	9.85
C	CARDIOVASCULAR SYSTEM	0.3170	1.77

Community n. 77

Enriched ATC code	Definition	p-value	Odds-ratio
C08C	SELECTIVE CALCIUM CHANNEL BLOCKERS WITH MAINLY VASCULAR EFFECTS	0.0002	85.33
C08CA	Dihydropyridine derivatives	0.0002	85.33
C08	CALCIUM CHANNEL BLOCKERS	0.0007	42.67
C	CARDIOVASCULAR SYSTEM	0.0931	3.53

Community n. 22

Enriched ATC code	Definition	p-value	Odds-ratio

N06AF	Monoamine oxidase inhibitors, non-selective	0.0002	76.80
N06A	ANTIDEPRESSANTS	0.0059	15.36
N06	PSYCHOANALEPTICS	0.0096	12.00
N	NERVOUS SYSTEM	0.1297	3.02

Community n. 43

Enriched ATC code	Definition	p-value	Odds-ratio
G03CC	Estrogens, combinations with other drugs	0.0002	76.80
G03C	ESTROGENS	0.0003	61.44
G03	SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	0.0062	15.36
L	ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS	0.0228	7.88
G	GENITO URINARY SYSTEM AND SEX HORMONES	0.0499	5.21

Community n. 69

Enriched ATC code	Definition	p-value	Odds-ratio
P02CA	Benzimidazole derivatives	0.0004	54.86
C01CA	Adrenergic and dopaminergic agents	0.0025	24.38
P02C	ANTINEMATODAL AGENTS	0.0031	21.94
C01C	CARDIAC STIMULANTS EXCL. CARDIAC GLYCOSIDES	0.0053	16.88
P02	ANTHELMINTICS	0.0071	14.63
P	ANTIPARASITIC PRODUCTS, INSECTICIDES AND REPELLENTS	0.0367	6.27
C01	CARDIAC THERAPY	0.0407	5.93
C	CARDIOVASCULAR SYSTEM	0.1284	2.27

Community n. 15

Enriched ATC code	Definition	p-value	Odds-ratio
S01E	ANTIGLAUCOMA PREPARATIONS AND MIOTICS	0.0004	48.00
S01	OPHTHALMOLOGICALS	0.0121	9.04
S	SENSORY ORGANS	0.0127	8.83

Community n. 75

Enriched ATC code	Definition	p-value	Odds-ratio
L04	IMMUNOSUPPRESSANTS	0.0006	48.00
L04A	IMMUNOSUPPRESSANTS	0.0006	48.00
L	ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS	0.0141	9.85
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.1225	3.12

Community n. 89

Enriched ATC code	Definition	p-value	Odds-ratio
B02B	VITAMIN K AND OTHER HEMOSTATICS	0.0005	45.18
B02	ANTIHEMORRHAGICS	0.0027	22.59
B	BLOOD AND BLOOD FORMING ORGANS	0.0070	7.13
R03	DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	0.0822	4.11
G	GENITO URINARY SYSTEM AND SEX HORMONES	0.3805	1.53
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.3978	1.33
R	RESPIRATORY SYSTEM	0.5219	1.17
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.5290	1.10
S01	OPHTHALMOLOGICALS	0.5784	1.06
S	SENSORY ORGANS	0.5919	1.04

Community n. 73

Enriched ATC code	Definition	p-value	Odds-ratio
R03A	ADRENERGICS, INHALANTS	0.0011	36.57
G02	OTHER GYNECOLOGICALS	0.0001	32.91
G02C	OTHER GYNECOLOGICALS	0.0015	31.35
R03C	ADRENERGICS FOR SYSTEMIC USE	0.0015	31.35

B. COMMUNITY ENRICHMENTS

R03	DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	0.0151	9.97
G	GENITO URINARY SYSTEM AND SEX HORMONES	0.0009	7.44
R	RESPIRATORY SYSTEM	0.5219	1.17

Community n. 99

Enriched ATC code	Definition	p-value	Odds-ratio
R03DA	Xanthines	0.0012	34.13
R03D	OTHER SYSTEMIC DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	0.0018	28.44
J01X	OTHER ANTIBACTERIALS	0.0033	21.33
C03	DIURETICS	0.0208	8.53
R03	DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	0.0250	7.76
C	CARDIOVASCULAR SYSTEM	0.0704	2.35
R	RESPIRATORY SYSTEM	0.2256	2.22
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.3405	1.67
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.4360	1.39

Community n. 61

Enriched ATC code	Definition	p-value	Odds-ratio
N06B	PSYCHOSTIMULANTS, AGENTS USED FOR ADHD AND NOOTROPICS	0.0021	25.60
N06BX	Other psychostimulants and nootropics	0.0021	25.60
B01AC	Platelet aggregation inhibitors excl. heparin	0.0051	17.07
B01	ANTITHROMBOTIC AGENTS	0.0119	11.38
B01A	ANTITHROMBOTIC AGENTS	0.0119	11.38
R01A	DECONGESTANTS AND OTHER NASAL PREPARATIONS FOR TOPICAL USE	0.0455	5.69
R01	NASAL PREPARATIONS	0.0503	5.39
B	BLOOD AND BLOOD FORMING ORGANS	0.0503	5.39
N06	PSYCHOANALEPTICS	0.1260	3.20
C01	CARDIAC THERAPY	0.1600	2.77
R	RESPIRATORY SYSTEM	0.0542	2.66
N	NERVOUS SYSTEM	0.4631	1.21
A	ALIMENTARY TRACT AND METABOLISM	0.5598	1.10
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.6136	1.01

Community n. 64

Enriched ATC code	Definition	p-value	Odds-ratio
R03D	OTHER SYSTEMIC DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	0.0027	23.27
R03	DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES	0.0369	6.35
M01	ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS	0.0857	3.99
M01A	ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS, NON-STEROIDS	0.0857	3.99
M	MUSCULO-SKELETAL SYSTEM	0.1936	2.45
R	RESPIRATORY SYSTEM	0.3038	1.81
S01	OPHTHALMOLOGICALS	0.3483	1.64
S	SENSORY ORGANS	0.3594	1.61
C	CARDIOVASCULAR SYSTEM	0.3452	1.44
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.4409	1.37
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.5473	1.14

Community n. 61

Enriched ATC code	Definition	p-value	Odds-ratio
J01M	QUINOLONE ANTIBACTERIALS	0.0027	23.27
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.1834	2.51
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.2473	2.08

Community n. 49

Enriched ATC code	Definition	p-value	Odds-ratio
J01FA	Macrolides	0.0033	21.33
J01F	MACROLIDES, LINCOSAMIDES AND STREPTOGRAMINS	0.0052	17.07

A07	ANTIDIARRHEALS, INTESTINAL ANTIINFLAMMATORY/ANTIINFECTIVE AGENTS	0.0020	10.67
A	ALIMENTARY TRACT AND METABOLISM	0.0840	2.75
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.1050	2.51
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.1622	2.08

Community n. 44

Enriched ATC code	Definition	p-value	Odds-ratio
C01	CARDIAC THERAPY	0.0023	20.76
C	CARDIOVASCULAR SYSTEM	0.0354	5.30

Community n. 76

Enriched ATC code	Definition	p-value	Odds-ratio
B	BLOOD AND BLOOD FORMING ORGANS	0.0034	20.21
A	ALIMENTARY TRACT AND METABOLISM	0.0740	4.13

Community n. 13

Enriched ATC code	Definition	p-value	Odds-ratio
S01F	MYDRIATICS AND CYCLOPLEGICS	0.0039	19.20
S01FA	Anticholinergics	0.0039	19.20
A03	DRUGS FOR FUNCTIONAL GASTROINTESTINAL DISORDERS	0.0412	6.00
S01E	ANTIGLAUCOMA PREPARATIONS AND MIOTICS	0.0412	6.00
S01	OPHTHALMOLOGICALS	0.0241	2.82
S	SENSORY ORGANS	0.0264	2.76
M01	ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS	0.1625	2.74
M01A	ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS, NON-STEROIDS	0.1625	2.74
M	MUSCULO-SKELETAL SYSTEM	0.3358	1.68
G	GENITO URINARY SYSTEM AND SEX HORMONES	0.3517	1.63
A	ALIMENTARY TRACT AND METABOLISM	0.3043	1.55
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.3591	1.41
C	CARDIOVASCULAR SYSTEM	0.3567	1.32
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.4857	1.17
N	NERVOUS SYSTEM	0.5089	1.13
D	DERMATOLOGICALS	0.5443	1.13

Community n. 62

Enriched ATC code	Definition	p-value	Odds-ratio
G01AF	Imidazole derivatives	0.0003	19.20
C08C	SELECTIVE CALCIUM CHANNEL BLOCKERS WITH MAINLY VASCULAR EFFECTS	0.0051	17.07
C08CA	Dihydropyridine derivatives	0.0051	17.07
D01AC	Imidazole and triazole derivatives	0.0093	12.80
G01	GYNECOLOGICAL ANTIINFECTIVES AND ANTISEPTICS	0.0002	11.38
G01A	ANTIINFECTIVES AND ANTISEPTICS, EXCL. COMBINATIONS WITH CORTICOSTEROIDS	0.0002	11.38
S02A	ANTIINFECTIVES	0.0119	11.38
S02AA	Antiinfectives	0.0119	11.38
D01	ANTIFUNGALS FOR DERMATOLOGICAL USE	0.0029	9.60
D01A	ANTIFUNGALS FOR TOPICAL USE	0.0029	9.60
A01AB	Antiinfectives and antiseptics for local oral treatment	0.0177	9.31
C08	CALCIUM CHANNEL BLOCKERS	0.0210	8.53
S02	OTOLOGICALS	0.0283	7.31
P02	ANTHELMINTICS	0.0323	6.83
A01	STOMATOLOGICAL PREPARATIONS	0.0365	6.40
A01A	STOMATOLOGICAL PREPARATIONS	0.0365	6.40
P	ANTIPARASITIC PRODUCTS, INSECTICIDES AND REPELLENTS	0.0271	4.39
G	GENITO URINARY SYSTEM AND SEX HORMONES	0.0037	4.34
D	DERMATOLOGICALS	0.0736	2.41

B. COMMUNITY ENRICHMENTS

S	SENSORY ORGANS	0.5215	1.18
A	ALIMENTARY TRACT AND METABOLISM	0.5598	1.10
C	CARDIOVASCULAR SYSTEM	0.5605	1.06

Community n. 59

Enriched ATC code	Definition	p-value	Odds-ratio
D07	CORTICOSTEROIDS, DERMATOLOGICAL PREPARATIONS	0.0056	16.17
D07A	CORTICOSTEROIDS, PLAIN	0.0056	16.17
D	DERMATOLOGICALS	0.0970	3.61

Community n. 6

Enriched ATC code	Definition	p-value	Odds-ratio
M01AB	Acetic acid derivatives and related substances	0.0065	15.36
M01	ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS	0.0722	4.39
M01A	ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS, NON-STEROIDS	0.0722	4.39
M	MUSCULO-SKELETAL SYSTEM	0.1660	2.69
A	ALIMENTARY TRACT AND METABOLISM	0.1097	2.48
N	NERVOUS SYSTEM	0.2202	1.81

Community n. 100

Enriched ATC code	Definition	p-value	Odds-ratio
N05AB	Phenothiazines with piperazine structure	0.0000	15.06
N05AF	Monoamine oxidase inhibitors, non-selective	0.0003	15.06
N05AG	Diphenylbutylpiperidine derivatives	0.0043	15.06
N06AA	Non-selective monoamine reuptake inhibitors	0.0000	10.95
R06AD	Phenothiazine derivatives	0.0124	10.04
N05AA	Phenothiazines with aliphatic side-chain	0.0025	9.04
N05A	ANTIPSYCHOTICS	0.0000	8.07
C08E	NON-SELECTIVE CALCIUM CHANNEL BLOCKERS	0.0238	7.53
N05	PSYCHOLEPTICS	0.0000	6.84
N06A	ANTIDEPRESSANTS	0.0000	6.02
N06AB	Selective serotonin reuptake inhibitors	0.0380	6.02
R06AA	Aminoalkyl ethers	0.0380	6.02
N06	PSYCHOANALEPTICS	0.0000	4.71
N	NERVOUS SYSTEM	0.0000	3.08
P02C	ANTINEMATODAL AGENTS	0.1383	3.01
R06	ANTIHISTAMINES FOR SYSTEMIC USE	0.0280	2.79
R06A	ANTIHISTAMINES FOR SYSTEMIC USE	0.0280	2.79
C08	CALCIUM CHANNEL BLOCKERS	0.1866	2.51
D04	ANTIPRURITICS, INCL. ANTIHISTAMINES, ANESTHETICS, ETC.	0.1866	2.51
D04A	ANTIPRURITICS, INCL. ANTIHISTAMINES, ANESTHETICS, ETC.	0.1866	2.51
P02	ANTHELMINTICS	0.2623	2.01
G03	SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	0.3883	1.51
P	ANTIPARASITIC PRODUCTS, INSECTICIDES AND REPELLENTS	0.4149	1.29
R	RESPIRATORY SYSTEM	0.4049	1.17
G	GENITO URINARY SYSTEM AND SEX HORMONES	0.5637	1.02

Community n. 32

Enriched ATC code	Definition	p-value	Odds-ratio
J01ED	Long-acting sulfonamides	0.0068	14.63
J01A	TETRACYCLINES	0.0228	8.13
J01AA	Tetracyclines	0.0228	8.13
J01E	SULFONAMIDES AND TRIMETHOPRIM	0.0532	5.22
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.0141	2.51
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.0112	2.38
P	ANTIPARASITIC PRODUCTS, INSECTICIDES AND REPELLENTS	0.2475	2.09
D	DERMATOLOGICALS	0.4161	1.29
S01	OPHTHALMOLOGICALS	0.4161	1.29

S	SENSORY ORGANS	0.4317	1.26

Community n. 90

Enriched ATC code	Definition	p-value	Odds-ratio
C01BD	Antiarrhythmics, class III	0.0049	14.22
D10AA	Corticosteroids, combinations for treatment of acne	0.0139	9.48
R06AE	Piperazine derivatives	0.0266	7.11
A03A	DRUGS FOR FUNCTIONAL BOWEL DISORDERS	0.0204	4.74
C05AA	Corticosteroids	0.0606	4.74
S01BA	Corticosteroids, plain	0.0277	4.27
J04A	DRUGS FOR TREATMENT OF TUBERCULOSIS	0.0811	4.06
D01AC	Imidazole and triazole derivatives	0.1034	3.56
A02	DRUGS FOR ACID RELATED DISORDERS	0.1271	3.16
A02B	DRUGS FOR PEPTIC ULCER AND GASTRO-OESOPHAGEAL REFLUX DISEASE (GORD)	0.1271	3.16
D10	ANTI-ACNE PREPARATIONS	0.1271	3.16
D10A	ANTI-ACNE PREPARATIONS FOR TOPICAL USE	0.1271	3.16
H02AB	Glucocorticoids	0.1271	3.16
J04	ANTIMYCOBACTERIALS	0.1271	3.16
J01E	SULFONAMIDES AND TRIMETHOPRIM	0.0688	3.05
D07	CORTICOSTEROIDS, DERMATOLOGICAL PREPARATIONS	0.0384	2.99
D07A	CORTICOSTEROIDS, PLAIN	0.0384	2.99
R05	COUGH AND COLD PREPARATIONS	0.1520	2.84
A03	DRUGS FOR FUNCTIONAL GASTROINTESTINAL DISORDERS	0.0959	2.67
S01B	ANTIINFLAMMATORY AGENTS	0.0959	2.67
H02	CORTICOSTEROIDS FOR SYSTEMIC USE	0.1777	2.59
H02A	CORTICOSTEROIDS FOR SYSTEMIC USE, PLAIN	0.1777	2.59
D04	ANTIPRURITICS, INCL. ANTIHISTAMINES, ANESTHETICS, ETC.	0.2040	2.37
D04A	ANTIPRURITICS, INCL. ANTIHISTAMINES, ANESTHETICS, ETC.	0.2040	2.37
C01B	ANTIARRHYTHMICS, CLASS I AND III	0.2308	2.19
R06	ANTIHISTAMINES FOR SYSTEMIC USE	0.1144	2.11
R06A	ANTIHISTAMINES FOR SYSTEMIC USE	0.1144	2.11
C05	VASOPROTECTIVES	0.2578	2.03
C05A	AGENTS FOR TREATMENT OF HEMORRHOIDS AND ANAL FISSURES FOR TOPICAL USE	0.2578	2.03
D01	ANTIFUNGALS FOR DERMATOLOGICAL USE	0.3119	1.78
D01A	ANTIFUNGALS FOR TOPICAL USE	0.3119	1.78
H	SYSTEMIC HORMONAL PREPARATIONS, EXCL. SEX HORMONES AND INSULINS	0.3119	1.78
D	DERMATOLOGICALS	0.0631	1.67
R	RESPIRATORY SYSTEM	0.1620	1.48
G03	SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	0.4170	1.42
S01	OPHTHALMOLOGICALS	0.2381	1.34
S	SENSORY ORGANS	0.2590	1.31
J01D	OTHER BETA-LACTAM ANTIBACTERIALS	0.4906	1.24
A	ALIMENTARY TRACT AND METABOLISM	0.3248	1.22
N06A	ANTIDEPRESSANTS	0.5365	1.14
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.5076	1.04

Community n. 102

Enriched ATC code	Definition	p-value	Odds-ratio
C10	LIPID MODIFYING AGENTS	0.0079	13.96
C10A	LIPID MODIFYING AGENTS, PLAIN	0.0079	13.96
C04	PERIPHERAL VASODILATORS	0.0166	9.60
C04A	PERIPHERAL VASODILATORS	0.0166	9.60
S01B	ANTIINFLAMMATORY AGENTS	0.0166	9.60
C	CARDIOVASCULAR SYSTEM	0.0251	2.65
D	DERMATOLOGICALS	0.3054	1.81
S01	OPHTHALMOLOGICALS	0.3054	1.81
S	SENSORY ORGANS	0.3156	1.77

B. COMMUNITY ENRICHMENTS

A	ALIMENTARY TRACT AND METABOLISM	0.3462	1.65

Community n. 42

Enriched ATC code	Definition	p-value	Odds-ratio
J01F	MACROLIDES, LINCOSAMIDES AND STREPTOGRAMINS	0.0079	13.96
A01AB	Antiinfectives and antiseptics for local oral treatment	0.0096	12.69
A01	STOMATOLOGICAL PREPARATIONS	0.0201	8.73
A01A	STOMATOLOGICAL PREPARATIONS	0.0201	8.73
S01E	ANTIGLAUCOMA PREPARATIONS AND MIOTICS	0.0201	8.73
S01A	ANTIINFECTIVES	0.0652	4.65
S01	OPHTHALMOLOGICALS	0.0251	3.29
S	SENSORY ORGANS	0.0272	3.21
D	DERMATOLOGICALS	0.1122	2.46
G	GENITO URINARY SYSTEM AND SEX HORMONES	0.2043	2.37
A	ALIMENTARY TRACT AND METABOLISM	0.1381	2.25
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.1698	2.05
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.2518	1.70
C	CARDIOVASCULAR SYSTEM	0.3452	1.44

Community n. 82

Enriched ATC code	Definition	p-value	Odds-ratio
N01	ANESTHETICS	0.0081	13.71
N	NERVOUS SYSTEM	0.0042	3.78

Community n. 52

Enriched ATC code	Definition	p-value	Odds-ratio
J01MB	Other quinolones	0.0082	13.47
J01M	QUINOLONE ANTIBACTERIALS	0.0019	11.02
G03D	PROGESTOGENS	0.0149	10.11
N07	OTHER NERVOUS SYSTEM DRUGS	0.0279	7.35
G03	SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	0.0845	4.04
J01D	OTHER BETA-LACTAM ANTIBACTERIALS	0.1077	3.51
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.0077	2.77
S01A	ANTIINFECTIVES	0.1673	2.69
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.0215	2.30
G	GENITO URINARY SYSTEM AND SEX HORMONES	0.1735	2.06
S01	OPHTHALMOLOGICALS	0.1490	1.90
S	SENSORY ORGANS	0.1586	1.86

Community n. 58

Enriched ATC code	Definition	p-value	Odds-ratio
M01AE	Propionic acid derivatives	0.0094	12.80
M01	ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS	0.1000	3.66
M01A	ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS, NON-STEROIDS	0.1000	3.66
M	MUSCULO-SKELETAL SYSTEM	0.2218	2.25
S01	OPHTHALMOLOGICALS	0.3904	1.51
S	SENSORY ORGANS	0.4022	1.47
N	NERVOUS SYSTEM	0.6159	1.01

Community n. 34

Enriched ATC code	Definition	p-value	Odds-ratio
R06AX	Other antihistamines for systemic use	0.0103	12.19
N06AA	Non-selective monoamine reuptake inhibitors	0.0155	9.97
N02	ANALGESICS	0.0283	7.31
C04	PERIPHERAL VASODILATORS	0.0320	6.86
C04A	PERIPHERAL VASODILATORS	0.0320	6.86
R06	ANTIHISTAMINES FOR SYSTEMIC USE	0.0109	6.10

R06A	ANTIHISTAMINES FOR SYSTEMIC USE	0.0109	6.10
N06A	ANTIDEPRESSANTS	0.0729	4.39
N06	PSYCHOANALEPTICS	0.1121	3.43
R	RESPIRATORY SYSTEM	0.1577	2.14
C	CARDIOVASCULAR SYSTEM	0.1050	1.89
N	NERVOUS SYSTEM	0.1882	1.73

Community n. 5

Enriched ATC code	Definition	p-value	Odds-ratio
S03	OPHTHALMOLOGICAL AND OTOLOGICAL PREPARATIONS	0.0135	10.67
J01CA	Penicillins with extended spectrum	0.0166	9.60
S02	OTOLOGICALS	0.0320	6.86
H	SYSTEMIC HORMONAL PREPARATIONS, EXCL. SEX HORMONES AND INSULINS	0.0412	6.00
J01C	BETA-LACTAM ANTIBACTERIALS, PENICILLINS	0.0412	6.00
A07	ANTIDIARRHEALS, INTESTINAL ANTIINFLAMMATORY/ANTIINFECTIVE AGENTS	0.0860	4.00
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.0124	2.82
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.0302	2.34
A	ALIMENTARY TRACT AND METABOLISM	0.3043	1.55
N	NERVOUS SYSTEM	0.5089	1.13
S01	OPHTHALMOLOGICALS	0.5443	1.13
S	SENSORY ORGANS	0.5577	1.10

Community n. 29

Enriched ATC code	Definition	p-value	Odds-ratio
D06B	CHEMOTHERAPEUTICS FOR TOPICAL USE	0.0152	10.04
J01DC	Second-generation cephalosporins	0.0152	10.04
N03	ANTIEPILEPTICS	0.0226	8.21
N03A	ANTIEPILEPTICS	0.0226	8.21
D06	ANTIBIOTICS AND CHEMOTHERAPEUTICS FOR DERMATOLOGICAL USE	0.0070	7.13
B	BLOOD AND BLOOD FORMING ORGANS	0.0632	4.76
P01	ANTIPROTOZOALS	0.0632	4.76
J01D	OTHER BETA-LACTAM ANTIBACTERIALS	0.0888	3.93
S01A	ANTIINFECTIVES	0.1396	3.01
P	ANTIPARASITIC PRODUCTS, INSECTICIDES AND REPELLENTS	0.1791	2.58
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.0110	2.57
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.0624	2.21
D	DERMATOLOGICALS	0.1083	2.13
N	NERVOUS SYSTEM	0.5526	1.07
S01	OPHTHALMOLOGICALS	0.5784	1.06
S	SENSORY ORGANS	0.5919	1.04

Community n. 40

Enriched ATC code	Definition	p-value	Odds-ratio
R06AX	Other antihistamines for systemic use	0.0170	9.48
N01B	ANESTHETICS, LOCAL	0.0209	8.53
N01	ANESTHETICS	0.0400	6.10
G03	SEX HORMONES AND MODULATORS OF THE GENITAL SYSTEM	0.0768	4.27
R06	ANTIHISTAMINES FOR SYSTEMIC USE	0.1289	3.16
R06A	ANTIHISTAMINES FOR SYSTEMIC USE	0.1289	3.16
P	ANTIPARASITIC PRODUCTS, INSECTICIDES AND REPELLENTS	0.1959	2.44
R	RESPIRATORY SYSTEM	0.0965	2.22
G	GENITO URINARY SYSTEM AND SEX HORMONES	0.1541	2.17
N	NERVOUS SYSTEM	0.0609	2.02
D	DERMATOLOGICALS	0.3204	1.51

Community n. 67

Enriched ATC code	Definition	p-value	Odds-ratio

B. COMMUNITY ENRICHMENTS

C01	CARDIAC THERAPY	0.0206	8.30
C	CARDIOVASCULAR SYSTEM	0.2396	2.12

Community n. 106

Enriched ATC code	Definition	p-value	Odds-ratio
C01	CARDIAC THERAPY	0.0206	8.30
G	GENITO URINARY SYSTEM AND SEX HORMONES	0.0499	5.21
C	CARDIOVASCULAR SYSTEM	0.2396	2.12

Community n. 16

Enriched ATC code	Definition	p-value	Odds-ratio
D06	ANTIBIOTICS AND CHEMOTHERAPEUTICS FOR DERMATOLOGICAL USE	0.0232	8.08
S01A	ANTIINFECTIVES	0.0546	5.12
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.2058	1.87
D	DERMATOLOGICALS	0.3054	1.81
S01	OPHTHALMOLOGICALS	0.3054	1.81
S	SENSORY ORGANS	0.3156	1.77
N	NERVOUS SYSTEM	0.5121	1.21

Community n. 97

Enriched ATC code	Definition	p-value	Odds-ratio
M01	ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS	0.0270	7.31
M01A	ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS, NON-STEROIDS	0.0270	7.31

Community n. 46

Enriched ATC code	Definition	p-value	Odds-ratio
M	MUSCULO-SKELETAL SYSTEM	0.0669	4.49
A	ALIMENTARY TRACT AND METABOLISM	0.1574	2.75
M	MUSCULO-SKELETAL SYSTEM	0.0295	6.74

Community n. 26

Enriched ATC code	Definition	p-value	Odds-ratio
G	GENITO URINARY SYSTEM AND SEX HORMONES	0.0315	6.51

Community n. 36

Enriched ATC code	Definition	p-value	Odds-ratio
D	DERMATOLOGICALS	0.0626	4.52
S01	OPHTHALMOLOGICALS	0.0626	4.52
S	SENSORY ORGANS	0.0654	4.41
D	DERMATOLOGICALS	0.0337	6.02
N	NERVOUS SYSTEM	0.0726	4.03

Community n. 2

Enriched ATC code	Definition	p-value	Odds-ratio
C	CARDIOVASCULAR SYSTEM	0.0354	5.30

Community n. 88

Enriched ATC code	Definition	p-value	Odds-ratio
G	GENITO URINARY SYSTEM AND SEX HORMONES	0.0499	5.21

Community n. 7

Enriched ATC code	Definition	p-value	Odds-ratio

A01	STOMATOLOGICAL PREPARATIONS	0.0513	5.33
A01A	STOMATOLOGICAL PREPARATIONS	0.0513	5.33
H	SYSTEMIC HORMONAL PREPARATIONS, EXCL. SEX HORMONES AND INSULINS	0.0513	5.33
S01B	ANTIINFLAMMATORY AGENTS	0.0513	5.33
M02AA	Antiinflammatory preparations, non-steroids for topical use	0.0573	5.02
C02	ANTIHYPERTENSIVES	0.0636	4.74
M02	TOPICAL PRODUCTS FOR JOINT AND MUSCULAR PAIN	0.0636	4.74
M02A	TOPICAL PRODUCTS FOR JOINT AND MUSCULAR PAIN	0.0636	4.74
N05	PSYCHOLEPTICS	0.0380	3.88
J01D	OTHER BETA-LACTAM ANTIBACTERIALS	0.0982	3.71
M01	ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS	0.0442	3.66
M01A	ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS, NON-STEROIDS	0.0442	3.66
M	MUSCULO-SKELETAL SYSTEM	0.1429	2.25
N	NERVOUS SYSTEM	0.1618	1.68
R	RESPIRATORY SYSTEM	0.2665	1.66
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.5702	1.04
D	DERMATOLOGICALS	0.6107	1.01
S01	OPHTHALMOLOGICALS	0.6107	1.01

Community n. 50

Enriched ATC code	Definition	p-value	Odds-ratio
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.0337	3.76
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.0555	3.12

Community n. 26

Enriched ATC code	Definition	p-value	Odds-ratio
S01	OPHTHALMOLOGICALS	0.0486	3.39
S	SENSORY ORGANS	0.0516	3.31

Community n. 37

Enriched ATC code	Definition	p-value	Odds-ratio
C07	BETA BLOCKING AGENTS	0.0552	5.12
C07A	BETA BLOCKING AGENTS	0.0552	5.12
G	GENITO URINARY SYSTEM AND SEX HORMONES	0.1935	1.95
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.2688	1.51
C	CARDIOVASCULAR SYSTEM	0.3208	1.32
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.4019	1.25
A	ALIMENTARY TRACT AND METABOLISM	0.4441	1.24

Community n. 29

Enriched ATC code	Definition	p-value	Odds-ratio
M01	ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS	0.0722	4.39
M01A	ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS, NON-STEROIDS	0.0722	4.39
M	MUSCULO-SKELETAL SYSTEM	0.1660	2.69
C	CARDIOVASCULAR SYSTEM	0.1006	2.12
R	RESPIRATORY SYSTEM	0.2647	1.99
A	ALIMENTARY TRACT AND METABOLISM	0.3462	1.65

Community n. 12

Enriched ATC code	Definition	p-value	Odds-ratio
R	RESPIRATORY SYSTEM	0.0813	3.99

Community n. 4

Enriched ATC code	Definition	p-value	Odds-ratio

B. COMMUNITY ENRICHMENTS

G	GENITO URINARY SYSTEM AND SEX HORMONES	0.0949	3.72

Community n. 60

Enriched ATC code	Definition	p-value	Odds-ratio
N06	PSYCHOANALEPTICS	0.0985	3.69
M01	ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS	0.1149	3.38
M01A	ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS, NON-STEROIDS	0.1149	3.38
M	MUSCULO-SKELETAL SYSTEM	0.2502	2.07
D	DERMATOLOGICALS	0.4312	1.39
S01	OPHTHALMOLOGICALS	0.4312	1.39
S	SENSORY ORGANS	0.4436	1.36
A	ALIMENTARY TRACT AND METABOLISM	0.4801	1.27
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.5326	1.16

Community n. 20

Enriched ATC code	Definition	p-value	Odds-ratio
C	CARDIOVASCULAR SYSTEM	0.0931	3.53

Community n. 47

Enriched ATC code	Definition	p-value	Odds-ratio
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.1335	3.01
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.1833	2.50

Community n. 64

Enriched ATC code	Definition	p-value	Odds-ratio
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.1335	3.01
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.1833	2.50

Community n. 83

Enriched ATC code	Definition	p-value	Odds-ratio
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.1335	3.01
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.1833	2.50

Community n. 33

Enriched ATC code	Definition	p-value	Odds-ratio
C	CARDIOVASCULAR SYSTEM	0.1634	2.65

Community n. 41

Enriched ATC code	Definition	p-value	Odds-ratio
C	CARDIOVASCULAR SYSTEM	0.1634	2.65

Community n. 72

Enriched ATC code	Definition	p-value	Odds-ratio
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.1834	2.51
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.2473	2.08

Community n. 85

Enriched ATC code	Definition	p-value	Odds-ratio
A	ALIMENTARY TRACT AND METABOLISM	0.2037	2.36
C	CARDIOVASCULAR SYSTEM	0.3925	1.51

Community n. 4

Enriched ATC code	Definition	p-value	Odds-ratio

C	CARDIOVASCULAR SYSTEM	0.1284	2.27
N	NERVOUS SYSTEM	0.3265	1.73

Community n. 95

Enriched ATC code	Definition	p-value	Odds-ratio
N	NERVOUS SYSTEM	0.1314	2.27
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.2882	1.88
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.3751	1.56

Community n. 8

Enriched ATC code	Definition	p-value	Odds-ratio
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.2473	2.08

Community n. 27

Enriched ATC code	Definition	p-value	Odds-ratio
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.1622	2.08
S01	OPHTHALMOLOGICALS	0.2620	2.01
S	SENSORY ORGANS	0.2712	1.96
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.3405	1.67

Community n. 87

Enriched ATC code	Definition	p-value	Odds-ratio
N	NERVOUS SYSTEM	0.3265	1.73

Community n. 31

Enriched ATC code	Definition	p-value	Odds-ratio
R	RESPIRATORY SYSTEM	0.3425	1.66
C	CARDIOVASCULAR SYSTEM	0.4019	1.32
J01	ANTIBACTERIALS FOR SYSTEMIC USE	0.4880	1.25
J	ANTIINFECTIVES FOR SYSTEMIC USE	0.5969	1.04
N	NERVOUS SYSTEM	0.6159	1.01

B.3 Molecular direct targets

Community n. 77

Enriched direct Target	p-value	Odds-ratio
CACNG1	0.0000	167.33

Community n. 75

Enriched direct Target	p-value	Odds-ratio
IMPDH1	0.0001	83.33

Community n. 14

Enriched direct Target	p-value	Odds-ratio
GSK3B	0.0002	62.50
TOP2A	0.0076	13.89

Community n. 63

Enriched direct Target	p-value	Odds-ratio
ATP1A1	0.0002	62.50

Community n. 29

Enriched direct Target	p-value	Odds-ratio
ABAT	0.0016	27.78
pbpA	0.1246	3.21

Community n. 62

Enriched direct Target	p-value	Odds-ratio
CACNA2D1	0.0021	25.00
ERG11	0.0003	18.75
CALM1	0.0120	11.11

Community n. 52

Enriched direct Target	p-value	Odds-ratio
PGR	0.0043	16.73

Community n. 58

Enriched direct Target	p-value	Odds-ratio
CA2	0.0063	15.21

Community n. 89

Enriched direct Target	p-value	Odds-ratio
PDE4A	0.0063	15.15
ADRA2A	0.0259	7.58

Community n. 34

Enriched direct Target	p-value	Odds-ratio
KCNH2	0.0077	13.89
HRH1	0.0825	4.03

Community n. 47

Enriched direct Target	p-value	Odds-ratio
pbpA	0.0076	12.87

Community n. 100

Enriched direct Target	p-value	Odds-ratio
GRIN2D	0.0063	12.50
CACNA1A	0.0178	8.33
DRD1IP	0.0178	8.33
DRD4	0.0178	8.33
CHRM5	0.0043	7.50
HTR2C	0.0043	7.50
SLC6A2	0.0000	7.14
SLC6A4	0.0000	7.03
CALM1	0.0003	6.94
DRD1	0.0000	5.92
HTR2A	0.0000	5.36
CHRM4	0.0133	5.36
OPRD1	0.0535	5.00
CHRM3	0.0201	4.69
DRD2	0.0000	4.41
HRH1	0.0000	4.03
CHRM2	0.0317	3.13
HTR1A	0.1285	3.13

ADRA1A	0.1158	2.08
CHRM1	0.2994	1.56
ESR1	0.4012	1.47

Community n. 90

Enriched direct Target	p-value	Odds-ratio
PLA2G4A	0.0008	12.10
SERPINA6	0.0098	6.05
CACNA1G	0.0331	6.45
CHRM4	0.0643	4.61
HRH1	0.1166	2.08
HTR2A	0.1342	2.30
ANXA1	0.1667	2.69
ADRA1A	0.1802	2.02

Community n. 43

Enriched direct Target	p-value	Odds-ratio
ESR1	0.0103	11.81
ALB	0.0293	6.92

Community n. 13

Enriched direct Target	p-value	Odds-ratio
ACHE	0.0118	11.11
TTR	0.0118	11.11
CHRM1	0.1584	2.78

Community n. 88

Enriched direct Target	p-value	Odds-ratio
ADRB2	0.0128	10.57

Community n. 73

Enriched direct Target	p-value	Odds-ratio
ADRB2	0.0188	8.77

Community n. 97

Enriched direct Target	p-value	Odds-ratio
PTGS1	0.0186	8.70
PTGS2	0.0219	8.00

Community n. 49

Enriched direct Target	p-value	Odds-ratio
PTGS1	0.0273	7.28
PTGS2	0.0320	6.69

Community n. 60

Enriched direct Target	p-value	Odds-ratio
SLC6A2	0.0290	7.14

Community n. 3

Enriched direct Target	p-value	Odds-ratio
PTGS1	0.0371	6.24
PTGS2	0.0434	5.74

Community n. 32

Enriched direct Target	p-value	Odds-ratio
SLC6A4	0.0448	5.23
pbpA	0.1246	3.21

Community n. 74

Enriched direct Target	p-value	Odds-ratio
PTGS1	0.0599	4.83
PTGS2	0.0697	4.44

Community n. 31

Enriched direct Target	p-value	Odds-ratio
HRH1	0.0643	4.61

Community n. 5

Enriched direct Target	p-value	Odds-ratio
pbpA	0.0310	4.12
ALB	0.1920	2.46

Community n. 40

Enriched direct Target	p-value	Odds-ratio
HTR2A	0.0992	3.66

HRH1	0.1897	2.48

Community n. 7

Enriched direct Target	p-value	Odds-ratio
DRD2	0.0414	3.39
PTGS1	0.1160	3.34
PTGS2	0.1336	3.08
HRH1	0.1897	2.48

Community n. 19

Enriched direct Target	p-value	Odds-ratio
ADRB1	0.1168	3.33
DRD2	0.2714	1.96i

Appendix C

Mode of Action enrichments

C.1 ATC codes

[ATC-codes associated to a set of 482 distinct drugs]

L01DB Anthracyclines and related substances

Community	*p*-**value**	**Odds-ratio**
14 - Rich Club n. 4	0.0000	247.25

L01D CYTOTOXIC ANTIBIOTICS AND RELATED SUBSTANCES

Community	*p*-**value**	**Odds-ratio**
14 - Rich Club n. 4	0.0000	247.25

C01AA Digitalis glycoside

Community	*p*-**value**	**Odds-ratio**
63 - Rich Club n. 3	0.0000	164.83

C01A CARDIAC GLYCOSIDES

Community	*p*-**value**	**Odds-ratio**
63 - Rich Club n. 3	0.0000	164.83

L01C PLANT ALKALOIDS AND OTHER NATURAL PRODUCTS

Community	*p*-**value**	**Odds-ratio**
48 - Rich Club n. 1	0.0000	164.83

C08CA Dihydropyridine derivatives

Community	*p*-**value**	**Odds-ratio**
77 - Rich Club n. 4	0.0001	131.87
62 - Rich Club n. 3	0.0068	14.65

C08C SELECTIVE CALCIUM CHANNEL BLOCKERS WITH MAINLY VASCULAR EFFECTS

Community	*p*-**value**	**Odds-ratio**
77 - Rich Club n. 4	0.0001	131.87
62 - Rich Club n. 3	0.0068	14.65

L01AD Nitrosoureas

Community	*p*-**value**	**Odds-ratio**
104 - Rich Club n. 3	0.0000	123.63

N06AF Monoamine oxidase inhibitors, non-selective

Community	*p*-**value**	**Odds-ratio**
22 - Rich Club n. 7	0.0001	98.90

A07AB Sulfonamides

Community	*p*-**value**	**Odds-ratio**
96	0.0001	94.19

L01X OTHER ANTINEOPLASTIC AGENTS

Community	*p*-**value**	**Odds-ratio**
65 - Rich Club n. 6	0.0001	94.19

G03CC Estrogens, combinations with other drugs

Community	*p*-**value**	**Odds-ratio**

C. MODE OF ACTION ENRICHMENTS

43 - Rich Club n. 1	0.0003	61.81

L01A ALKYLATING AGENTS

Community	***p*-value**	**Odds-ratio**
104 - Rich Club n. 3	0.0000	61.81

P02CA Benzimidazole derivatives

Community	***p*-value**	**Odds-ratio**
69 - Rich Club n. 3	0.0003	61.81

L04A IMMUNOSUPPRESSANTS

Community	***p*-value**	**Odds-ratio**
75 - Rich Club n. 6	0.0006	49.45

S01E ANTIGLAUCOMA PREPARATIONS AND MIOTICS

Community	***p*-value**	**Odds-ratio**
15 - Rich Club n. 6	0.0006	43.96
42 - Rich Club n. 3	0.0489	5.49
13 - Rich Club n. 2	0.0606	4.88

B02B VITAMIN K AND OTHER HEMOSTATICS

Community	***p*-value**	**Odds-ratio**
89 - Rich Club n. 2	0.0006	39.56

R03DA Xanthines

Community	***p*-value**	**Odds-ratio**
99 - Rich Club n. 9	0.0011	35.96

C01CA Adrenergic and dopaminergic agents

Community	***p*-value**	**Odds-ratio**
69 - Rich Club n. 3	0.0016	30.91

R03A ADRENERGICS, INHALANTS

Community	***p*-value**	**Odds-ratio**
73 - Rich Club n. 2	0.0016	29.97

R03D OTHER SYSTEMIC DRUGS FOR OBSTRUCTIVE AIRWAY DISEASES

Community	***p*-value**	**Odds-ratio**
99 - Rich Club n. 9	0.0016	29.97
74 - Rich Club n. 2	0.0035	20.60

G02C OTHER GYNECOLOGICALS

Community	***p*-value**	**Odds-ratio**
73 - Rich Club n. 2	0.0023	25.69

R03C ADRENERGICS FOR SYSTEMIC USE

Community	***p*-value**	**Odds-ratio**
73 - Rich Club n. 2	0.0023	25.69

N06BX Other psychostimulants and nootropics

Community	***p*-value**	**Odds-ratio**
61 - Rich Club n. 2	0.0023	24.73

N06B PSYCHOSTIMULANTS, AGENTS USED FOR ADHD AND NOOTROPICS

Community	***p*-value**	**Odds-ratio**
61 - Rich Club n. 2	0.0023	24.73

P02C ANTINEMATODAL AGENTS

Community	***p*-value**	**Odds-ratio**
69 - Rich Club n. 3	0.0025	24.73

J01M QUINOLONE ANTIBACTERIALS

Community	***p*-value**	**Odds-ratio**
93 - Rich Club n. 7	0.0030	22.48
52 - Rich Club n. 2	0.0029	9.63

J01X OTHER ANTIBACTERIALS

Community	***p*-value**	**Odds-ratio**
99 - Rich Club n. 9	0.0030	22.48

A07A INTESTINAL ANTIINFECTIVES

Community	***p*-value**	**Odds-ratio**
96	0.0032	21.74

J01FA Macrolides

Community	***p*-value**	**Odds-ratio**
49 - Rich Club n. 2	0.0036	20.60

C01C CARDIAC STIMULANTS EXCL. CARDIAC GLYCOSIDES

Community	***p*-value**	**Odds-ratio**

69 - Rich Club n. 3	0.0036	20.60

N06A ANTIDEPRESSANTS

Community	***p*-value**	**Odds-ratio**
22 - Rich Club n. 7	0.0036	19.78
100 - Rich Club n. 1	0.0000	7.60
34 - Rich Club n. 1	0.0527	5.27
90 - Rich Club n. 2	0.5473	1.11

D07A CORTICOSTEOROIDS, PLAIN

Community	***p*-value**	**Odds-ratio**
59 - Rich Club n. 2	0.0040	19.39
90 - Rich Club n. 2	0.0284	3.28

N05AB Phenothiazines with piperazine structure

Community	***p*-value**	**Odds-ratio**
100 - Rich Club n. 1	0.0000	19.02

N05AF Monoamine oxidase inhibitors, non-selective

Community	***p*-value**	**Odds-ratio**
100 - Rich Club n. 1	0.0001	19.02

N05AG Diphenylbutylpiperidine derivatives

Community	***p*-value**	**Odds-ratio**
100 - Rich Club n. 1	0.0027	19.02

S01FA Anticholinergics

Community	***p*-value**	**Odds-ratio**
13 - Rich Club n. 2	0.0042	18.31

S01F MYDRIATICS AND CYCLOPLEGICS

Community	***p*-value**	**Odds-ratio**
13 - Rich Club n. 2	0.0042	18.31

J01F MACROLIDES, LINCOSAMIDES AND STREPTOGRAMINS

Community	***p*-value**	**Odds-ratio**
49 - Rich Club n. 2	0.0046	18.31
42 - Rich Club n. 3	0.0183	9.16

B01AC Platelet aggregation inhibitors excl. heparin

Community	***p*-value**	**Odds-ratio**
61 - Rich Club n. 2	0.0056	16.48

M01AB Acetic acid derivatives and related substances

Community	***p*-value**	**Odds-ratio**
6 - Rich Club n. 6	0.0058	16.48

R06AX Other antihistamines for systemic use

Community	***p*-value**	**Odds-ratio**
34 - Rich Club n. 1	0.0057	16.48
40 - Rich Club n. 3	0.0182	9.16

C01BD Antiarrhythmics, class III

Community	***p*-value**	**Odds-ratio**
90 - Rich Club n. 2	0.0051	13.93

D10AA Corticosteroids, combinations for treatment of acne

Community	***p*-value**	**Odds-ratio**
90 - Rich Club n. 2	0.0051	13.93

N06AA Non-selective monoamine reuptake inhibitors

Community	***p*-value**	**Odds-ratio**
100 - Rich Club n. 1	0.0000	13.83
34 - Rich Club n. 1	0.0109	11.99

G01AF Imidazole derivatives

Community	***p*-value**	**Odds-ratio**
62 - Rich Club n. 3	0.0009	13.74

J01ED Long-acting sulfonamides

Community	***p*-value**	**Odds-ratio**
32 - Rich Club n. 4	0.0079	13.64

M01AE Propionic acid derivatives

Community	***p*-value**	**Odds-ratio**
58 - Rich Club n. 2	0.0090	13.19

R06AD Phenothiazine derivatives

Community	***p*-value**	**Odds-ratio**

C. MODE OF ACTION ENRICHMENTS

100 - Rich Club n. 1	0.0079	12.68

J01MB Other quinolones

Community	***p*-value**	**Odds-ratio**
52 - Rich Club n. 2	0.0108	11.77

G03D PROGESTOGENS

Community	***p*-value**	**Odds-ratio**
52 - Rich Club n. 2	0.0108	11.77

N05AA Phenothiazines with aliphatic side-chain

Community	***p*-value**	**Odds-ratio**
100 - Rich Club n. 1	0.0013	11.41

B01A ANTITHROMBOTIC AGENTS

Community	***p*-value**	**Odds-ratio**
61 - Rich Club n. 2	0.0129	10.99

N05A ANTIPSYCHOTICS

Community	***p*-value**	**Odds-ratio**
100 - Rich Club n. 1	0.0000	10.19

D06B CHEMOTHERAPEUTICS FOR TOPICAL USE

Community	***p*-value**	**Odds-ratio**
29 - Rich Club n. 7	0.0149	10.09

C08E NON SELECTIVE CALCIUM CHANNEL BLOCKERS

Community	***p*-value**	**Odds-ratio**
100 - Rich Club n. 1	0.0152	9.51

D01AC Imidazole and triazole derivatives

Community	***p*-value**	**Odds-ratio**
62 - Rich Club n. 3	0.0182	9.16
90 - Rich Club n. 2	0.1074	3.48

S02AA Antiinfectives

Community	***p*-value**	**Odds-ratio**
62 - Rich Club n. 3	0.0182	9.16

N01B ANESTHETICS, LOCAL

Community	***p*-value**	**Odds-ratio**
40 - Rich Club n. 3	0.0182	9.16

S02A ANTIINFECTIVES

Community	***p*-value**	**Odds-ratio**
62 - Rich Club n. 3	0.0182	9.16

J01DC Second-generation cephalosporins

Community	***p*-value**	**Odds-ratio**
29 - Rich Club n. 7	0.0195	8.83

G01A ANTIINFECTIVES AND ANTISEPTICS, EXCL. COMBINATIONS WITH CORTICOSTEROIDS

Community	***p*-value**	**Odds-ratio**
62 - Rich Club n. 3	0.0008	8.62

C04A PERIPHERAL VASODILATORS

Community	***p*-value**	**Odds-ratio**
34 - Rich Club n. 1	0.0228	8.24
102 - Rich Club n. 4	0.0470	5.62

R06A ANTIHISTAMINES FOR SYSTEMIC USE

Community	***p*-value**	**Odds-ratio**
34 - Rich Club n. 1	0.0047	8.24
100 - Rich Club n. 1	0.0066	3.96

C10A LIPID MODIFYING AGENTS, PLAIN

Community	***p*-value**	**Odds-ratio**
102 - Rich Club n. 4	0.0230	8.17

M01A ANTIINFLAMMATORY AND ANTIRHEUMATIC PRODUCTS, NON-STEROIDS

Community	***p*-value**	**Odds-ratio**
97	0.0229	8.07
6 - Rich Club n. 6	0.0643	8.07
58 - Rich Club n. 2	0.0958	4.71
74 - Rich Club n. 2	0.1071	3.77
60 - Rich Club n. 1	0.1429	3.53
37 - Rich Club n. 6	0.1554	2.97
7 - Rich Club n. 6	0.0998	2.83

13 - Rich Club n. 2	0.2471	2.65

J01CA Penicillins with extended spectrum

Community	*p*-value	Odds-ratio
5 - Rich Club n. 6	0.0237	2.09

N06AB Selective serotonin reuptake inhibitors

Community	*p*-value	Odds-ratio
100 - Rich Club n. 1	0.0245	7.61

J01A TETRACYCLINES

Community	*p*-value	Odds-ratio
32 - Rich Club n. 4	0.0263	7.58

J01AA Tetracyclines

Community	*p*-value	Odds-ratio
32 - Rich Club n. 4	0.0263	7.58

A01AB Antiinfectives and antiseptics for local oral treatment

Community	*p*-value	Odds-ratio
42 - Rich Club n. 3	0.0272	7.49
62 - Rich Club n. 3	0.0339	6.66

R06AE Piperazine derivatives

Community	*p*-value	Odds-ratio
90 - Rich Club n. 2	0.0277	6.96

D01A ANTIFUNGALS FOR TOPICAL USE

Community	*p*-value	Odds-ratio
62 - Rich Club n. 3	0.0080	6.87
90 - Rich Club n. 2	0.3209	1.74

R06AA Aminoalkyl ethers

Community	*p*-value	Odds-ratio
100 - Rich Club n. 1	0.0245	6.50

N03A ANTIEPILEPTICS

Community	*p*-value	Odds-ratio
29 - Rich Club n. 7	0.0363	6.42

S01B ANTIINFLAMMATORY AGENTS

Community	*p*-value	Odds-ratio
102 - Rich Club n. 4	0.0366	6.42
7 - Rich Club n. 6	0.0724	4.42
90 - Rich Club n. 2	0.0728	2.98

C07A BETA BLOCKING AGENTS

Community	*p*-value	Odds-ratio
19	0.0430	5.89

J01E SULFONAMIDES AND TRIMETHOPRIN

Community	*p*-value	Odds-ratio
32 - Rich Club n. 4	0.0457	5.68
90 - Rich Club n. 2	0.0488	3.48

C05AA Corticosteroids

Community	*p*-value	Odds-ratio
90 - Rich Club n. 2	0.0441	5.57

A01A STOMATOLOGICAL PREPARATIONS

Community	*p*-value	Odds-ratio
42 - Rich Club n. 3	0.0489	5.49
62 - Rich Club n. 3	0.0606	4.88
7 - Rich Club n. 6	0.0819	4.12

S01BA Corticosteroids, plain

Community	*p*-value	Odds-ratio
90 - Rich Club n. 2	0.0153	5.22

A03A DRUGS FOR FUNCTIONAL BOWEL DISORDERS

Community	*p*-value	Odds-ratio
90 - Rich Club n. 2	0.0153	5.22

S01A ANTIINFECTIVES

Community	*p*-value	Odds-ratio
16 - Rich Club n. 1	0.0593	4.95
29 - Rich Club n. 7	0.1555	2.83
42 - Rich Club n. 3	0.1207	3.30

C. MODE OF ACTION ENRICHMENTS

52 - Rich Club n. 2	0.1555	2.83
90 - Rich Club n. 2	0.5473	1.11

J01C BETA-LACTAM ANTIBACTERIALS, PENICILLINS

Community	*p*-value	Odds-ratio
5 - Rich Club n. 6	0.0685	4.56

D04A ANTIPRURITICS, INCL. ANTIHISTAMINES, ANESTHETICS, ETC.

Community	*p*-value	Odds-ratio
100 - Rich Club n. 1	0.0770	4.23
90 - Rich Club n. 2	0.1318	3.10

J04A DRUGS FOR THE TREATMENT OF TUBERCOLOSIS

Community	*p*-value	Odds-ratio
90 - Rich Club n. 2	0.0844	3.98

M02AA Antiinflammatory preparations, non-steroids for topical use

Community	*p*-value	Odds-ratio
7 - Rich Club n. 6	0.0918	3.86

M02A TOPICAL PRODUCTS FOR JOINT AND MUSCULAR PAIN

Community	*p*-value	Odds-ratio
7 - Rich Club n. 6	0.1020	3.64

J01D OTHER BETA-LACTAM ANTIBACTERIALS

Community	*p*-value	Odds-ratio
29 - Rich Club n. 7	0.1073	3.53
52 - Rich Club n. 2	0.1073	3.53
7 - Rich Club n. 6	0.1343	3.09
90 - Rich Club n. 2	0.4273	1.39

H02AB Glucocorticoids

Community	*p*-value	Odds-ratio
90 - Rich Club n. 2	0.1074	3.48

A02B DRUGS FOR PEPTIC ULCER AND GASTRO-OESOPHAGEAL REFLUX DISEASE (GORD)

Community	*p*-value	Odds-ratio
90 - Rich Club n. 2	0.1074	3.48

D10A ANTI-ACNE PREPARATIONS FOR TOPICAL USE

Community	*p*-value	Odds-ratio
90 - Rich Club n. 2	0.1074	3.48

H02A CORTICOSTEROIDS FOR SYSTEMIC USE, PLAIN

Community	*p*-value	Odds-ratio
90 - Rich Club n. 2	0.1574	2.79

C05A AGENTS FOR TREATMENT OF HEMORRHOIDS AND ANAL FISSURES FOR TOPICAL USE

Community	*p*-value	Odds-ratio
90 - Rich Club n. 2	0.1838	2.53

C01B ANTIARRHYTHMICS, CLASS I AND III

Community	*p*-value	Odds-ratio
90 - Rich Club n. 2	0.2109	2.32

C.2 Molecular direct targets

[Molecular direct targets associated to a set of distinct drugs]

CACNG1

Community	p-value	Odds-ratio
77 - Rich Club n. 4	0.0000	276.67

ATP1A1

Community	p-value	Odds-ratio
63 - Rich Club n. 3	0.0004	62.25

IMPDH1

Community	p-value	Odds-ratio
75 - Rich Club n. 1	0.0004	56.59

GSK3B

Community	p-value	Odds-ratio
14 - Rich Club n. 4	0.0003	54.13

ABAT

Community	p-value	Odds-ratio
29 - Rich Club n. 7	0.0007	43.68

CACNA2D1

Community	p-value	Odds-ratio
62 - Rich Club n. 3	0.0021	25.94

PTGS1

Community	p-value	Odds-ratio
97	0.0032	21.65
74 - Rich Club n. 2	0.0375	6.37
49 - Rich Club n. 2	0.0375	6.37
3 - Rich Club n. 3	0.0867	4.01
7 - Rich Club n. 3	0.1220	3.28

PTGS2

Community	p-value	Odds-ratio
97	0.0037	19.92
74 - Rich Club n. 2	0.0438	5.86
49 - Rich Club n. 2	0.0438	5.86
3 - Rich Club n. 3	0.1001	3.69
7 - Rich Club n. 3	0.1401	3.02

ERG11

Community	p-value	Odds-ratio
62 - Rich Club n. 3	0.0003	19.45

KCNH2

Community	p-value	Odds-ratio
34 - Rich Club n. 1	0.0053	17.29

PGR

Community	p-value	Odds-ratio
52 - Rich Club n. 2	0.0050	17.17

CA2

Community	p-value	Odds-ratio
58 - Rich Club n. 2	0.0061	16.17

ADRB2

Community	p-value	Odds-ratio
88 - Rich Club n. 3	0.0075	14.56
73 - Rich Club n. 2	0.0133	10.92

PDE4A

Community	p-value	Odds-ratio
89 - Rich Club n. 2	0.0085	13.39

ACHE

Community	p-value	Odds-ratio
13 - Rich Club n. 2	0.0090	12.97

TTR

Community	p-value	Odds-ratio
13 - Rich Club n. 2	0.0090	12.97

TOP2A

C. MODE OF ACTION ENRICHMENTS

Community	p-value	Odds-ratio
14 - Rich Club n. 4	0.0109	12.03

CALM1

Community	p-value	Odds-ratio
62 - Rich Club n. 3	0.0118	11.53
100 - Rich Club n. 1	0.0016	4.84

PLA2G4A

Community	p-value	Odds-ratio
90 - Rich Club n. 2	0.0010	11.53

GRIN2D

Community	p-value	Odds-ratio
100 - Rich Club n. 1	0.0131	8.71

pbpA

Community	p-value	Odds-ratio
47 - Rich Club n. 4	0.0206	8.71
29 - Rich Club n. 7	0.0577	5.04
32 - Rich Club n. 4	0.1070	3.55
5 - Rich Club n. 6	0.0828	2.87

SLC6A2

Community	p-value	Odds-ratio
60 - Rich Club n. 1	0.0239	8.08
100 - Rich Club n. 1	0.0000	4.98

ESR1

Community	p-value	Odds-ratio
43 - Rich Club n. 1	0.0262	7.71
100 - Rich Club n. 1	0.5988	1.02

ADRA2A

Community	p-value	Odds-ratio
89 - Rich Club n. 2	0.0339	6.69

CACNA1G

Community	p-value	Odds-ratio
90 - Rich Club n. 2	0.0368	6.15

CACNA1A

Community	p-value	Odds-ratio
100 - Rich Club n. 1	0.0364	5.80

DRD1IP

Community	p-value	Odds-ratio
100 - Rich Club n. 1	0.0364	5.80

DRD4

Community	p-value	Odds-ratio
100 - Rich Club n. 1	0.0364	5.80

SERPINA6

Community	p-value	Odds-ratio
90 - Rich Club n. 2	0.0117	5.76

SLC6A4

Community	p-value	Odds-ratio
32 - Rich Club n. 4	0.0451	5.76
100 - Rich Club n. 1	0.0000	4.90

HTR2A

Community	p-value	Odds-ratio
40 - Rich Club n. 3	0.0511	5.39
100 - Rich Club n. 1	0.0002	3.73
90 - Rich Club n. 2	0.1517	2.20

CHRM5

Community	p-value	Odds-ratio
100 - Rich Club n. 1	0.0125	5.22

HTR2C

Community	p-value	Odds-ratio
100 - Rich Club n. 1	0.0125	5.22

HRH1

Community	p-value	Odds-ratio

34 - Rich Club n. 1	0.0580	5.02
31 - Rich Club n. 7	0.0717	4.46
40 - Rich Club n. 3	0.1018	3.65
100 - Rich Club n. 1	0.0015	2.81
7 - Rich Club n. 3	0.1972	2.43
90 - Rich Club n. 2	0.1373	1.98

ALB

Community	*p*-value	Odds-ratio
43 - Rich Club n. 1	0.0701	4.52
5 - Rich Club n. 6	0.3263	1.72

CHRM4

Community	*p*-value	Odds-ratio
90 - Rich Club n. 2	0.0709	4.39
100 - Rich Club n. 1	0.0367	3.73

DRD1

Community	*p*-value	Odds-ratio
100 - Rich Club n. 1	0.0001	4.12

ADRB1

Community	*p*-value	Odds-ratio
19	0.1064	3.56

OPRD1

Community	*p*-value	Odds-ratio
100 - Rich Club n. 1	0.1038	3.48

DRD2

Community	*p*-value	Odds-ratio
7 - Rich Club n. 3	0.0581	3.33
100 - Rich Club n. 1	0.0002	3.07
19	0.2473	2.09

CHRM3

Community	*p*-value	Odds-ratio
100 - Rich Club n. 1	0.0538	3.26

CHRM1

Community	*p*-value	Odds-ratio
13 - Rich Club n. 2	0.1244	3.24
100 - Rich Club n. 1	0.5329	1.09

ANXA1

Community	*p*-value	Odds-ratio
90 - Rich Club n. 2	0.1809	2.56

CHRM2

Community	*p*-value	Odds-ratio
100 - Rich Club n. 1	0.1014	2.18

HTR1A

Community	*p*-value	Odds-ratio
100 - Rich Club n. 1	0.2318	2.18

ADRA1A

Community	*p*-value	Odds-ratio
90 - Rich Club n. 2	0.2015	1.92
100 - Rich Club n. 1	0.2939	1.45

Appendix D

ESF similarity and communities

Community	Enriched for a given:	n. drug with SMILES/n. drug in the community	Average ESF similarity
63	Literature-Evidence/ATC-codes/Direct-Target-Gene	2/11	0.55
77	Literature-Evidence/ATC-codes/Direct-Target-Gene	3/4	0.3922
43	Literature-Evidence/ATC-codes	5/9	0.2726
82	ATC-codes	5/10	0.235
100	Literature-Evidence/ATC-codes	42/76	0.2186
73	Literature-Evidence/ATC-codes	5/9	0.1894
65	Literature-Evidence/ATC-codes	2/3	0.1857
25	ATC-codes	5/10	0.1853
22	Literature-Evidence/ATC-codes	4/12	0.1754
104	Literature-Evidence/ATC-codes	5/17	0.1747
88	Literature-Evidence/ATC-codes	5/6	0.1717
34	Literature-Evidence/ATC-codes/Direct-Target-Gene	10/18	0.1708
14	Literature-Evidence/ATC-codes/Direct-Target-Gene	8/15	0.1697
13	Literature-Evidence/ATC-codes	15/25	0.1686
67	ATC-codes	4/10	0.1645
62	Literature-Evidence/ATC-codes/Direct-Target-Gene	11/30	0.1611
93	Literature-Evidence/ATC-codes	5/6	0.1544
58	ATC-codes	6/18	0.1518
53	Literature-Evidence	2/6	0.1463
42	Literature-Evidence/ATC-codes	10/15	0.1451
50	Literature-Evidence/ATC-codes	2/7	0.1449
89	Literature-Evidence/ATC-codes/Direct-Target-Gene	13/33	0.1448
74	ATC-codes	9/14	0.1433
29	ATC-codes/Direct-Target-Gene	12/21	0.143
46	ATC-codes	2/7	0.1429
60	Literature-Evidence	11/17	0.1414
90	Literature-Evidence/ATC-codes	38/79	0.1323
75	Literature-Evidence/ATC-codes/Direct-Target-Gene	3/4	0.1318
61	ATC-codes	7/20	0.1318
32	Literature-Evidence/ATC-codes	13/39	0.1317
6	ATC-codes	8/19	0.1314
81	Literature-Evidence	2/5	0.1277
52	Literature-Evidence/ATC-codes/Direct-Target-Gene	14/30	0.1273
16	Literature-Evidence/ATC-codes	9/16	0.1243
69	Literature-Evidence/ATC-codes	4/10	0.1232
28	Literature-Evidence	2/5	0.122
5	ATC-codes	15/24	0.1193
96	Literature-Evidence/ATC-codes	2/5	0.119

102	ATC-codes	7/16	0.1175
7	ATC-codes	14/25	0.1175
40	Literature-Evidence/ATC-codes	13/35	0.1158
99	Literature-Evidence/ATC-codes	10/14	0.1143
106	ATC-codes	3/9	0.1126
26	Literature-Evidence/ATC-codes	5/6	0.1106
76	ATC-codes	4/5	0.1084
3	Literature-Evidence	9/19	0.1062
49	Literature-Evidence/ATC-codes	6/12	0.0968
97	ATC-codes	6/9	0.0944
36	ATC-codes	3/6	0.0886
48	Literature-Evidence/ATC-codes	2/9	0.0805
91	Literature-Evidence	2/5	0.0741
39	Literature-Evidence	2/5	0.0577

Appendix E

cMap online tool results

*True Positives: drugs sharing the mode of action with the testing one

PHA-793887 on A2780			**NMS-Flavopiridol on A2780**			**PHA-690509 on A2780**		
cmap name	*cs*	*p*	*cmap name*	*cs*	*p*	*cmap name*	*cs*	*p*
apigenin*	0.9990	0.0000	alsterpaullone*	0.9990	0.0000	camptothecin	0.9990	0.0000
irinotecan	0.9980	0.0000	doxorubicin	0.9980	0.0000	GW-8510*	0.9990	0.0000
alsterpaullone*	0.9950	0.0000	camptothecin	0.9980	0.0000	alsterpaullone*	0.9980	0.0000
phenoxybenzamine	0.9930	0.0000	H-7*	0.9970	0.0000	H-7*	0.9970	0.0000
luteolin*	0.9800	0.0000	GW-8510*	0.9970	0.0000	0175029-0000	0.8910	0.0000
ellipticine	0.9750	0.0000	apigenin*	0.9690	0.0000	vorinostat	0.7280	0.0000
GW-8510*	0.9600	0.0000	luteolin*	0.9400	0.0000	LY-294002	0.5630	0.0000
vorinostat	0.9520	0.0000	trichostatin-A	0.3430	0.0000	trichostatin-A	0.5480	0.0000
trichostatin-A	0.7580	0.0000	thioguanosine	0.9220	0.0000	sirolimus	0.3860	0.0000
thioridazine	0.5500	0.0000	chrysin*	0.9570	0.0001	proscillaridin	0.9760	0.0000
LY-294002	0.4980	0.0000	8-azaguanine	0.8880	0.0002	hexestrol	0.9080	0.0001
camptothecin	0.9880	0.0000	meticrane	0.8100	0.0006	lanatoside-C	0.8190	0.0001
thioguanosine	0.9370	0.0000	phthalylsulfathiazole	0.7870	0.0009	thioguanosine	0.8870	0.0002
chrysin*	0.9710	0.0000	medrysone	0.7300	0.0010	helveticoside	0.7830	0.0003
acacetin	0.8250	0.0001	rimexolone	0.8290	0.0013	digoxigenin	0.7850	0.0010
proscillaridin	0.9590	0.0001	LY-294002	0.2400	0.0015	irinotecan	0.9170	0.0012
resveratrol	0.6810	0.0001	vorinostat	0.5050	0.0023	astemizole	0.7470	0.0024
digitoxigenin	0.8900	0.0001	sanguinarine	0.9610	0.0026	resveratrol	0.5710	0.0025
lanatoside-C	0.7790	0.0003	trazodone	0.8860	0.0029	menadione	0.9610	0.0026
8-azaguanine	0.8800	0.0003	hexestrol	0.7980	0.0032	MS-275	0.9550	0.0038
piperlongumine	0.9870	0.0003	mitoxantrone	0.8820	0.0032	digoxin	0.7830	0.0041
bisacodyl	0.8690	0.0004	piperlongumine	0.9510	0.0044	ouabain	0.7830	0.0041
parthenolide	0.8670	0.0004	milrinone	0.8640	0.0047	digitoxigenin	0.7790	0.0045
scriptaid	0.9340	0.0005	gliclazide	0.7740	0.0049	1.4-chrysenequinone	0.9440	0.0060
pyrvinium	0.7660	0.0005	0175029-0000	0.6440	0.0061	scriptaid	0.8500	0.0063
MS-275	0.9840	0.0005	daunorubicin	0.7570	0.0066	mefloquine	0.6950	0.0064
helveticoside	0.7530	0.0006	sulconazole	0.7570	0.0066	SC-19220	0.7550	0.0069
astemizole	0.8060	0.0006	phenoxybenzamine	0.7550	0.0069	strophanthidin	0.7440	0.0080
ciclopirox	0.8470	0.0008	resveratrol	0.5220	0.0079	5707885	0.7400	0.0088
digoxin	0.8450	0.0009	ellipticine	0.7460	0.0079	sanguinarine	0.9280	0.0098
bepridil	0.8400	0.0010	oxprenolol	0.7430	0.0082	rifabutin	0.8230	0.0111
dilazep	0.7690	0.0014	trioxysalen	0.7300	0.0105	thioridazine	0.3390	0.0155
tiabendazole	0.8260	0.0014	irinotecan	0.8260	0.0107	troglitazone	0.3750	0.0158
CP-690334-01	0.6170	0.0019	DL-thiorphan	0.9250	0.0110	harmine*	0.6970	0.0173
azacitidine	0.8980	0.0021	verteporfin	0.8190	0.0117	piperlongumine	0.9010	0.0202
withaferin-A	0.8140	0.0023	famprofazone	0.5990	0.0132	benzethonium-chloride	0.7700	0.0244

E. CMAP ONLINE TOOL RESULTS

amrinone	0.8120	0.0024	N-acetyl-L-leucine	0.7160	0.0133	sulfametoxydiazine	0.6700	0.0263
0297417-0002B	0.8900	0.0025	tyloxapol	0.7010	0.0164	skimmianine	0.6660	0.0279
spironolactone	0.7320	0.0032	antimycin-A	0.6420	0.0168	AR-A014418	0.7570	0.0283
mefloquine	0.7160	0.0042	levonorgestrel	0.5860	0.0171	guaifenesin	0.5550	0.0295
suloctidil	0.7820	0.0043	bisacodyl	0.6980	0.0171	bisacodyl	0.6550	0.0329
trifluoperazine	0.4160	0.0048	levamisole	0.6900	0.0193	quinostatin	0.8540	0.0433
cloperastine	0.6530	0.0052	procaine	0.6320	0.0194	clorgiline	0.6320	0.0454
0175029-0000	0.6510	0.0054	trifluoperazine	0.3660	0.0197	betazole	0.5720	0.0456
rottlerin	0.8420	0.0078	morantel	0.6210	0.0229	isotretinoin	0.6310	0.0463
harmine*	0.7470	0.0078	ethaverine	0.6790	0.0230	alclometasone	0.6280	0.0479
CP-645525-01	0.8420	0.0079	ebselen	0.7730	0.0236	harmol*	0.6250	0.0496
terfenadine	0.8410	0.0080	doxazosin	0.6770	0.0236			
fluphenazine	0.3760	0.0085	clomipramine	0.6750	0.0245			
daunorubicin	0.7360	0.0094	repaglinide	0.6710	0.0258			
oxetacaine	0.6720	0.0099	triflusal	0.7650	0.0259			
methylbenzethonium-chloride	0.6150	0.0101	eucatropine	0.5550	0.0295			
prochlorperazine	0.3920	0.0101	etofenamate	0.6600	0.0308			
ouabain	0.7290	0.0108	ronidazole	0.7500	0.0309			
fulvestrant	0.2500	0.0108	azacitidine	0.7460	0.0320			
hycanthone	0.7220	0.0118	norethisterone	0.6530	0.0337			
mycophenolic-acid	0.8160	0.0122	etomidate	0.7390	0.0348			
15-delta-prostaglandin-J2	0.3960	0.0123	demeclocycline	0.5420	0.0356			
trazodone	0.8070	0.0145	difenidol	0.7340	0.0369			
pimozide	0.7090	0.0148	lorglumide	0.5870	0.0372			
1.4-chrysenequinone	0.9120	0.0156	skimmianine	0.6440	0.0386			
PNU-0251126	0.5880	0.0166	acetylsalicylic-acid	0.3720	0.0389			
digoxigenin	0.6390	0.0176	omeprazole	0.6420	0.0403			
strophanthidin	0.6930	0.0185	liothyronine	0.6410	0.0406			
econazole	0.6880	0.0198	ginkgolide-A	0.6400	0.0407			
sanguinarine*	0.9000	0.0205	1.4-chrysenequinone	0.8570	0.0414			
tretinoin	0.3110	0.0214	Prestwick-1084	0.6380	0.0421			
clioquinol	0.6220	0.0226	zomepirac	0.6360	0.0428			
alexidine	0.6770	0.0238	chlorpromazine	0.3070	0.0434			
epiandrosterone	0.6760	0.0240	meclofenoxate	0.5270	0.0447			
disulfiram	0.6170	0.0244	carbachol	0.6310	0.0462			
PNU-0293363	0.7650	0.0258	apramycin	0.6280	0.0480			
MG-262	0.7640	0.0261	amiodarone	0.5680	0.0483			
pyridoxine	0.6670	0.0275	alvespimycin	0.3750	0.0485			
prenylamine	0.6650	0.0286	nipecotic-acid	0.6240	0.0498			
niclosamide	0.5920	0.0350						
cetirizine	0.6490	0.0359						
harmol*	0.6460	0.0377						
fluspirilene	0.6400	0.0408						
tonzonium-bromide	0.6370	0.0427						
griseofulvin	0.5720	0.0457						
phenazopyridine	0.6320	0.0458						
puromycin	0.6300	0.0465						
etacrynic-acid	0.7080	0.0486						
5182598	0.8440	0.0487						
5707885	0.6250	0.0494						

*True Positives: drugs sharing the mode of action with the testing one

PHA-848125 on U251			**NMS-SN38 on MCF7**			**NMS-Doxorubicin on MCF7**		
cmap name	*cs*	*p*	*cmap name*	*cs*	*p*	*cmap name*	*cs*	*p*
irinotecan	1.0000	0.0000	irinotecan*	0.9980	0.0000	resveratrol	0.7610	0.0000
alsterpaullone*	0.9980	0.0000	phenoxybenzamine	0.9950	0.0000	thioridazine	0.5980	0.0000
camptothecin	0.9970	0.0000	thioguanosine	0.9620	0.0000	trichostatin-A	0.5720	0.0000
phenoxybenzamine	0.9960	0.0000	resveratrol	0.7490	0.0000	camptothecin	0.9770	0.0000
apigenin*	0.9940	0.0000	15-delta-prostaglandin-J2	0.6710	0.0000	trifluoperazine	0.5740	0.0000
luteolin*	0.9900	0.0000	trichostatin-A	0.4080	0.0000	trifluridine	0.9270	0.0000
thioguanosine	0.9890	0.0000	camptothecin*	0.9940	0.0000	15-delta-prostaglandin-J2	0.5720	0.0001
ellipticine	0.9880	0.0000	mycophenolic-acid	0.9660	0.0000	LY-294002	0.2660	0.0001
8-azaguanine	0.9670	0.0000	mebendazole	0.8420	0.0002	mycophenolic-acid	0.9590	0.0001
thiostrepton	0.9510	0.0000	8-azaguanine	0.8760	0.0003	proscillaridin	0.9590	0.0001
parthenolide	0.9500	0.0000	menadione	0.9860	0.0004	digitoxigenin	0.8850	0.0002
vorinostat	0.9010	0.0000	5194442	0.8690	0.0004	fluphenazine	0.4760	0.0003
geldanamycin	0.6000	0.0000	trifluridine	0.8670	0.0004	bufexamac	0.8790	0.0003
tanespimycin	0.5590	0.0000	pyrvinium	0.7110	0.0016	thiostrepton	0.8660	0.0004
trichostatin-A	0.4830	0.0000	prochlorperazine	0.4450	0.0021	phenoxybenzamine	0.8640	0.0005
piperlongumine	0.9970	0.0000	ciclopirox	0.8090	0.0025	irinotecan	0.9180	0.0011
proscillaridin	0.9870	0.0000	hexestrol	0.8070	0.0026	cloperastine	0.7120	0.0016
thioridazine	0.5380	0.0000	vorinostat	0.5020	0.0026	digoxin	0.8130	0.0024
chrysin*	0.9690	0.0000	lomustine	0.7990	0.0031	vorinostat	0.5030	0.0024
digitoxigenin	0.9290	0.0000	hycanthone	0.7950	0.0035	norcyclobenzaprine	0.7890	0.0039
harmine*	0.9120	0.0001	fluorometholone	0.7900	0.0039	scriptaid	0.8690	0.0041
fluphenazine	0.5130	0.0001	corbadrine	0.7880	0.0040	antimycin-A	0.7160	0.0043
acacetin	0.8140	0.0001	antimycin-A	0.7170	0.0041	hycanthone	0.7750	0.0049
lanatoside-C	0.7950	0.0002	piperlongumine	0.9500	0.0047	monobenzone	0.7710	0.0053
menadione	0.9860	0.0003	digitoxigenin	0.7760	0.0048	withaferin-A	0.7590	0.0064
helveticoside	0.7720	0.0003	cloperastine	0.6410	0.0064	helveticoside	0.6230	0.0089
alvespimycin	0.5600	0.0004	flupentixol	0.7570	0.0067	pinacidil	0.7370	0.0094
CP-690334-01	0.6640	0.0006	daunorubicin	0.7390	0.0089	quinostatin	0.9270	0.0101
atropine	0.8500	0.0007	syrosingopine	0.7350	0.0098	daunorubicin*	0.7320	0.0102
oxetacaine	0.7960	0.0008	luteolin*	0.7340	0.0098	MS-275	0.9270	0.0102
0297417-0002B	0.9170	0.0012	pipenzolate-bromide	0.7340	0.0098	0297417-0002B	0.8200	0.0117
withaferin-A	0.8320	0.0012	trifluoperazine	0.3920	0.0102	alimemazine	0.7180	0.0131
mebendazole	0.7770	0.0012	clomipramine	0.7300	0.0105	quercetin	0.5960	0.0141
amiloride	0.7680	0.0015	astemizole	0.6670	0.0108	pimozide	0.7090	0.0148
digoxigenin	0.7670	0.0015	gliclazide	0.7280	0.0109	zalcitabine	0.7010	0.0164
STOCK1N-35215	0.9080	0.0017	thiostrepton	0.7220	0.0120	cefotetan	0.7950	0.0175
dilazep	0.7580	0.0019	cycloserine	0.7200	0.0124	cinchocaine	0.6370	0.0179
oxantel	0.8000	0.0030	hydralazine	0.5980	0.0137	etoposide*	0.6880	0.0198
resveratrol	0.5630	0.0031	SC-19220	0.7010	0.0165	methyldopate	0.6870	0.0201
PHA-00851261E	0.5870	0.0036	harmine	0.6970	0.0176	zuclopenthixol	0.6860	0.0209
1.4-chrysenequinone	0.9540	0.0039	methotrexate	0.5100	0.0185	strophanthidin	0.6800	0.0225
digoxin	0.7750	0.0048	tolfenamic-acid	0.6910	0.0190	fenoterol	0.7760	0.0227
suloctidil	0.7720	0.0051	monocrotaline	0.6890	0.0195	clomipramine	0.6750	0.0245
ouabain	0.7710	0.0052	apigenin*	0.6880	0.0198	parthenolide	0.6730	0.0251
F0447-0125	0.7700	0.0054	parthenolide	0.6870	0.0204	5230742	0.8880	0.0258
5182598	0.9410	0.0065	ellipticine	0.6820	0.0218	procarbazine	0.7620	0.0268
mefloquine	0.6900	0.0072	fluspirilene	0.6700	0.0261	ouabain	0.6620	0.0298
nocodazole	0.6350	0.0072	pimethixene	0.7570	0.0282	dosulepin	0.6580	0.0314
hycanthone	0.7470	0.0078	harmol	0.6650	0.0284	lanatoside-C	0.5490	0.0320
pimozide	0.7460	0.0079	5155877	0.6640	0.0289	harpagoside	0.6510	0.0348
nifuroxazide	0.7450	0.0080	nocodazole	0.5550	0.0295	telenzepine	0.6500	0.0354
etacrynic-acid	0.8400	0.0080	proadifen	0.6570	0.0319	prochlorperazine	0.3420	0.0354
harmol*	0.7440	0.0080	semustine	0.6550	0.0330	mometasone	0.6480	0.0364

E. CMAP ONLINE TOOL RESULTS

0173570-0000	0.6240	0.0086	suloctidil	0.6540	0.0334	flupentixol	0.6450	0.0384
daunorubicin	0.7300	0.0105	pimozide	0.6480	0.0368	ciclopirox	0.6430	0.0390
5155877	0.7190	0.0125	oxantel	0.6460	0.0375	tretinoin	0.2890	0.0395
PHA-00767505E	0.7030	0.0161	STOCK1N-35215	0.7310	0.0383	nifenazone	0.5790	0.0411
bisacodyl	0.7000	0.0166	CP-645525-01	0.7260	0.0402	ebselen	0.7220	0.0421
6-bromoindirubin-3'-oxime	0.5390	0.0186	1.4-chrysenequinone	0.8570	0.0412	MG-262	0.7220	0.0425
Y-27632	0.9040	0.0190	tretinoin	0.2860	0.0433	syrosingopine	0.6360	0.0431
cotinine	0.5730	0.0224	phenformin	0.4840	0.0490	isotretinoin	0.6330	0.0448
oxyphenbutazone	0.6750	0.0245				fluspirilene	0.6280	0.0481
loperamide	0.5600	0.0270				cefuroxime	0.6260	0.0490
0179445-0000	0.4890	0.0275				methylbenzethonium-chloride	0.5200	0.0491
PNU-0293363	0.7560	0.0288						
levobunolol	0.6500	0.0354						
scriptaid	0.7360	0.0361						
bepridil	0.6480	0.0363						

*True Positives: drugs sharing the mode of action with the testing one

NMS-Tanespimycin on MCF7			**NMS-E973 on MCF7**			**NVP-AUY922 on MCF7**		
cmap name	*cs*	*p*	*cmap name*	*cs*	*p*	*cmap name*	*cs*	*p*
geldanamycin*	0.9810	0.0000	alvespimycin*	0.9800	0.0000	5155877	0.9660	0.0000
alvespimycin*	0.9740	0.0000	geldanamycin*	0.9790	0.0000	tanespimycin*	0.7800	0.0000
tanespimycin*	0.9700	0.0000	tanespimycin*	0.9730	0.0000	alvespimycin*	0.7760	0.0000
securinine	0.9640	0.0000	securinine	0.9670	0.0000	geldanamycin*	0.7730	0.0000
monorden*	0.9560	0.0000	monorden*	0.9170	0.0000	diethylstilbestrol	0.7170	0.0014
thiostrepton	0.9490	0.0000	vorinostat	0.7840	0.0000	withaferin-A*	0.7790	0.0045
15-delta-prostaglandin-J2	0.8020	0.0000	15-delta-prostaglandin-J2	0.7450	0.0000	lomustine	0.7720	0.0051
vorinostat	0.6670	0.0000	trichostatin-A	0.6450	0.0000	thiostrepton	0.7720	0.0052
mycophenolic-acid	0.9520	0.0001	thiostrepton	0.9380	0.0000	STOCK1N-35874	0.9470	0.0053
diethylstilbestrol	0.7970	0.0001	rifabutin*	0.9590	0.0001	semustine	0.7710	0.0053
rifabutin*	0.9410	0.0003	diethylstilbestrol	0.8000	0.0001	5194442	0.7690	0.0056
idazoxan	0.8250	0.0015	oxolamine	0.8810	0.0003	oxyphenbutazone	0.7610	0.0063
STOCK1N-35215	0.8890	0.0026	mycophenolic-acid	0.9350	0.0004	alcuronium-chloride	0.9420	0.0063
withaferin-A*	0.8020	0.0029	idazoxan	0.8490	0.0008	betulinic-acid	0.7560	0.0067
parthenolide	0.8020	0.0029	withaferin-A*	0.8190	0.0019	3-acetamidocoumarin	0.7500	0.0075
lomustine	0.7990	0.0030	parthenolide	0.8170	0.0021	PHA-00745360	0.5340	0.0118
scoulerine	0.7990	0.0031	scoulerine	0.8170	0.0021	atracurium-besilate	0.8180	0.0119
PNU-0251126	0.6750	0.0032	lomustine	0.8160	0.0022	podophyllotoxin	0.7200	0.0125
semustine	0.7970	0.0033	semustine	0.8120	0.0024	MK-886	0.9180	0.0132
5155877	0.7840	0.0041	5182598	0.9560	0.0034	thiamphenicol	0.6480	0.0152
proguanil	0.8650	0.0046	1.4-chrysenequinone	0.9550	0.0037	cefamandole	0.7010	0.0164
sodium-phenylbutyrate	0.6020	0.0055	piperlongumine	0.9530	0.0039	heptaminol	0.6410	0.0171
5182598	0.9440	0.0059	STOCK1N-35215	0.8660	0.0045	C-75	0.6920	0.0188
genistein	0.3990	0.0060	5155877	0.7700	0.0054	W-13	0.9010	0.0202
etacrynic-acid	0.8480	0.0067	Prestwick-1103	0.7470	0.0079	MG-262	0.7840	0.0206
piperlongumine	0.9380	0.0074	dinoprost	0.7420	0.0084	eticlopride	0.6810	0.0222
F0447-0125	0.7400	0.0088	nordihydroguaiaretic-acid	0.4070	0.0092	idazoxan	0.6780	0.0232
oxolamine	0.7300	0.0105	vigabatrin	0.8260	0.0107	tocainide	0.6780	0.0234
1.4-chrysenequinone	0.9250	0.0110	rifampicin	0.7270	0.0111	diprophylline	0.6160	0.0244
dinoprost	0.7170	0.0131	MG-262	0.8190	0.0117	naltrexone	0.6150	0.0252
MG-262	0.8030	0.0157	F0447-0125	0.6920	0.0188	nadolol	0.6710	0.0258
5194442	0.6890	0.0197	furosemide	0.6830	0.0217	levobunolol	0.6690	0.0266
epitiostanol	0.6850	0.0210	isotretinoin	0.6780	0.0234	amiprilose	0.6670	0.0276
CP-320650-01	0.4990	0.0231	halofantrine	0.7490	0.0312	thioperamide	0.6040	0.0296
carbimazole	0.7670	0.0254	iobenguane	0.6580	0.0314	Prestwick-1103	0.6610	0.0301
scriptaid	0.7570	0.0283	CP-690334-01	0.4800	0.0321	isoniazid	0.5890	0.0359
halofantrine	0.7520	0.0300	0317956-0000	0.4770	0.0337	indoprofen	0.6480	0.0366
0317956-0000	0.4780	0.0329	proguanil	0.7420	0.0337	metrizamide	0.6430	0.0390
cinchonine	0.6450	0.0384	indoprofen	0.6530	0.0338	canadine	0.6400	0.0411
Prestwick-967	0.6330	0.0446	Prestwick-983	0.7380	0.0353	chenodeoxycholic-acid	0.6380	0.0419
ribavirin	0.6240	0.0499	Prestwick-967	0.6440	0.0386	Prestwick-642	0.6350	0.0437
			metixene	0.6340	0.0443	flecainide	0.5280	0.0442
						trichlormethiazide	0.6340	0.0445
						cinchonine	0.6320	0.0454
						halofantrine	0.7140	0.0461
						Prestwick-692	0.6250	0.0497

Appendix F

Neighborhood of the tested compounds in the drug network

F. NEIGHBORHOOD OF THE TESTED COMPOUNDS IN THE DRUG NETWORK

Figure F.1: Nighborhood of the tested compounds 1

Neighbors of the tested compounds

PHA-848125

well Known CDK2 inhibitors			well Known CDK2 inhibitors		
Distance	**Compound**	**Community**	**Distance**	**Compound**	**Community**
0.6213	0175029-0000	32	0.7431	phenoxybenzamine	104
0.6352	apigenin	32	0.7440	digoxigenin	63
0.6504	harmine	32	0.7507	etoposide	3
0.6672	thioguanosine	32	0.7521	digoxin	63
0.6712	GW-8510	14	0.7539	harmol	32
0.6746	luteolin	32	0.7602	tyrphostin_AG-825	14
0.6795	daunorubicin	14	0.7610	doxorubicin	14
0.6828	irinotecan	32	0.7654	H-7	14
0.6877	camptothecin	14	0.7689	lanatoside_C	63
0.6887	piperlongumine	40	0.7700	mebendazole	69
0.6893	ellipticine	14	0.7818	bisacodyl	63
0.6905	alsterpaullone	14	0.7843	anisomycin	53
0.7072	0297417-0002B	100	0.7853	sanguinarine	89
0.7121	proscillaridin	63	0.7889	resveratrol	3
0.7201	digitoxigenin	63	0.7905	menadione	89
0.7250	8-azaguanine	89	0.7938	STOCK1N-35215	104
0.7277	staurosporine	14	0.7957	thiostrepton	104
0.7296	helveticoside	63	0.7969	azacitidine	14
0.7297	ouabain	63	0.8012	fluspirilene	100
0.7353	nocodazole	69			
0.7372	chrysin	32			

Distance	**Community**
0.0722	14
0.0846	32
0.0927	63
0.2554	89
0.2590	104
0.3762	69
0.3764	100
0.3848	3

PHA-793887

well-known CDK2 inhibitors			well-known CDK2 inhibitors		
Distance	**Compound**	**Community**	**Distance**	**Compound**	**Community**
0.4715	0175029-0000	32	0.7509	digoxigenin	63
0.4846	GW-8510	14	0.7510	digitoxigenin	63
0.5145	alsterpaullone	14	0.7547	mitoxantrone	14
0.5370	apigenin	32	0.7548	digoxin	63
0.5694	daunorubicin	14	0.7550	ciclopirox	3
0.5976	doxorubicin	14	0.7590	etoposide	3
0.6014	ellipticine	14	0.7621	harmine	32
0.6353	tyrphostin_AG-825	14	0.7760	5109870	3
0.6582	luteolin	32	0.7781	lanatoside_C	63
0.6607	camptothecin	14	0.7792	bisacodyl	63
0.6825	staurosporine	14	0.7810	methotrexate	75
0.6940	H-7	14	0.7831	proscillaridin	63
0.7018	irinotecan	32	0.7859	8-azaguanine	89
0.7040	thioguanosine	32	0.7876	vorinostat	16
0.7097	chrysin	32	0.7925	thiostrepton	104
0.7131	piperlongumine	40	0.7938	sanguinarine	89
0.7165	fisetin	14	0.7942	resveratrol	3
0.7314	ouabain	63	0.8004	PNU-0293363	21
0.7369	helveticoside	63	0.8028	sulconazole	90
0.7403	phenoxybenzamine	104			
0.7447	azacitidine	14			

Distance	**Community**
0.0527	14
0.0916	32
0.0947	63
0.1927	3
0.383	104

Community Enriched for CDK/Topo inhibition

Figure F.2: Nighborhood of the tested compounds 2

F. NEIGHBORHOOD OF THE TESTED COMPOUNDS IN THE DRUG NETWORK

Neighbors of the tested compounds

NMS-Flavopyridol					
	well-known CDK2 inhibitors			well-known CDK2 inhibitors	
Distance	**Compound**	**Community**	**Distance**	**Compound**	**Community**
0.4540	alsterpaullone	14	0.7810	scopoletin	90
0.4857	GW-8510	14	0.7811	digitoxigenin	63
0.5374	apigenin	32	0.7812	skimmianine	32
0.5534	0175029-0000	32	0.7851	ginkgolide_A	89
0.5789	daunorubicin	14	0.7875	hexestrol	90
0.5966	doxorubicin	14	0.7884	celastrol	40
0.5976	camptothecin	14	0.7890	meticrane	74
0.6196	ellipticine	14	0.7898	MS-275	14
0.6270	H-7	14	0.7898	tropicamide	90
0.6301	tyrphostin_AG-825	14	0.7905	etamsylate	89
0.6429	luteolin	32	0.7911	thiostrepton	104
0.6602	irinotecan	32	0.7915	antimycin_A	62
0.6722	chrysin	32	0.7924	verteporfin	89
0.6868	mitoxantrone	14	0.7939	omeprazole	90
0.6873	thioguanosine	32	0.7939	oxprenolol	85
0.7071	staurosporine	14	0.7953	tobramycin	95
0.7080	sanguinarine	89	0.7954	beta-escin	90
0.7089	azacitidine	14	0.7965	estriol	85
0.7098	8-azaguanine	89	0.7976	milrinone	85
0.7131	piperlongumine	40	0.7979	propylthiouracil	95
0.7159	fisetin	14	0.7987	dipyridamole	89
0.7228	phenoxybenzamine	104	0.7988	trioxysalen	32
0.7341	harmine	32	0.7991	napelline	55
0.7455	sulconazole	90	0.7991	fluorometholone	90
0.7493	doxazosin	52	0.7992	flunisolide	59
0.7578	carbachol	52	0.7998	cetirizine	90
0.7583	cantharidin	89	0.8002	denatonium_benzoate	90
0.7596	rimexolone	32	0.8009	ebselen	46
0.7616	proscillaridin	63	0.8011	reserpine	99
0.7654	bisacodyl	63	0.8027	ethaverine	52
0.7686	menadione	89	0.8033	PHA-00665752	19
0.7712	phthalylsulfathiazole	90	0.8042	repaglinide	85
0.7717	methotrexate	75	0.8043	bepridil	100
0.7743	0297417-0002B	100	0.8046	sulfametoxydiazine	32
0.7747	triflusal	61	0.8050	metyrapone	
0.7747	gliclazide	61	0.8051	morantel	
0.7750	1,4-chrysenequinone	40	0.8051	cinchocaine	
0.7750	medrysone	90	0.8056	clomipramine	
0.7775	tyloxapol	90	0.8058	riboflavin	
0.7778	Prestwick-860	52	0.8059	chlorzoxazone	
0.7781	5109870	3	0.8059	N-acetyl-L-leucine	
0.7795	DL-thiorphan	13			
0.7809	norethisterone	90			

Distance	**Community**
0.0480	14
0.0603	90
0.0625	32
0.0954	89
0.1929	52
0.1995	85
0.2527	40
0.2564	63
0.3781	104
0.3874	61
0.3946	100
0.3983	95

Community Enriched for CDK/Topo inhibition

Figure F.3: Nighborhood of the tested compounds 3 -

Neighbors of the tested compounds

PHA-690509												
	well-known CDK2 inhibitors			well-known CDK2 inhibitors			well-known CDK2 inhibitors			well-known CDK2 inhibitors		
Distance	**Compound**	**Community**	**Distance**	**Compound**	**Community**	**Distance**	**Compound**	**Community**	**Distance**	**Compound**	**Community**
0.3838	GW-8510	14	0.7572	ginkgolide_A	89	0.7822	methazolamide	42	0.7960	methotrexate	75
0.4613	doxorubicin	14	0.7593	bisacodyl	63	0.7823	verteporfin	89	0.7970	hexestrol	90
0.4794	alsterpaullone	14	0.7601	oxprenolol	85	0.7823	1,4-chrysenequinone	40	0.7975	levonorgestrel	58
0.5001	H-7	14	0.7632	carbachol	52	0.7840	trazodone	90	0.7975	ebselen	46
0.5593	daunorubicin	14	0.7658	lorglumide	90	0.7840	scopoletin	90	0.7978	ursodeoxycholic_acid	85
0.5873	camptothecin	14	0.7673	methacholine_chloride	13	0.7844	Prestwick-559	61	0.7979	piperidolate	90
0.5956	ellipticine	14	0.7673	apramycin	90	0.7850	norethisterone	90	0.7979	roxithromycin	74
0.6048	mitoxantrone	14	0.7682	Prestwick-1084	90	0.7851	thiostrepton	104	0.7983	moxonidine	95
0.6144	tyrphostin_AG-825	14	0.7684	meticrane	74	0.7868	etomidate	85	0.7987	acepromazine	90
0.6274	fisetin	14	0.7703	trioxysalen	32	0.7868	fusaric_acid	8	0.7992	cefalexin	90
0.6375	0175029-0000	32	0.7703	medrysone	90	0.7874	naftopidil	106	0.7993	molindone	106
0.6548	staurosporine	14	0.7704	phthalylsulfathiazole	90	0.7877	natamycin	52	0.7996	napelline	55
0.7091	azacitidine	14	0.7705	gliclazide	61	0.7879	etamsylate	89	0.7997	aminocaproic_acid	35
0.7120	sanguinarine	89	0.7716	chrysin	32	0.7879	cetirizine	90	0.7998	propafenone	61
0.7210	apigenin	32	0.7720	milrinone	85	0.7884	fluorometholone	90	0.8003	bezafibrate	90
0.7326	skimmianine	32	0.7726	propylthiouracil	95	0.7889	rimexolone	32	0.8004	3-nitropropionic_acid	18
0.7397	sulconazole	90	0.7739	aminohippuric_acid	58	0.7889	ethaverine	52	0.8005	difenidol	90
0.7399	triflusal	61	0.7767	flunixin	32	0.7892	thioguanosine	32	0.8009	urapidil	102
0.7401	cantharidin	89	0.7769	ethotoin	32	0.7905	atovaquone	99	0.8025	ipratropium_bromide	89
0.7435	DL-thiorphan	13	0.7772	denatonium_benzoate	90	0.7906	Prestwick-860	52	0.8029	5114445	4
0.7439	sitosterol	89	0.7777	pentoxyverine	90	0.7908	sulfametoxydiazine	32	0.8031	roxarsone	41
0.7439	proxyphylline	74	0.7789	zomepirac	46	0.7913	talampicillin	89	0.8041	pancuronium_bromide	32
0.7450	dexverapamil	49	0.7795	hydrocotarnine	89	0.7924	halcinonide	59	0.8043	butoconazole	62
0.7476	repaglinide	85	0.7800	HC_toxin	16	0.7924	0297417-0002B	100	0.8046	oxetacaine	31
0.7518	doxazosin	52	0.7801	morantel	74	0.7940	procaine	82			
0.7528	piperlongumine	40	0.7803	omeprazole	90	0.7940	6-azathymine	8			
0.7536	riboflavin	90	0.7807	eucatropine	90	0.7941	dimethyloxalylglycine	51			
0.7544	estriol	85	0.7809	menadione	89	0.7941	cycloserine	90			
0.7547	N-acetyl-L-leucine	89	0.7810	amiodarone	90	0.7941	metacycline	32			
0.7560	tyloxapol	90	0.7810	pargyline	89	0.7953	atropine	13			
0.7562	luteolin	32	0.7821	liothyronine	90	0.7955	piromidic_acid	52			

Distance	**Community**
0.0300	90
0.0464	14
0.0585	32
0.0639	89
0.1283	85
0.1299	52
0.1931	74
0.1933	61
0.2561	13
0.3837	40
0.3927	95
0.3928	58
0.3941	46
0.3952	8
0.3967	106

Community Enriched for CDK/Topo inhibitior

Figure F.4: Nighborhood of the tested compounds 4

F. NEIGHBORHOOD OF THE TESTED COMPOUNDS IN THE DRUG NETWORK

Neighbors of the tested compounds

NMS-SN38			NMS-Doxorubicin		
	well-known TopoI inhibitors			well-known TopoII inhibitors	
Distance	**Compound**	**Community**	**Distance**	**Compound**	**Community**
0.3215	irinotecan	32	0.5587	daunorubicin	14
0.5641	camptothecin	14	0.6495	GW-8510	14
0.6158	apigenin	32	0.6536	hycanthone	3
0.6251	phenoxybenzamine	104	0.6555	ellipticine	14
0.6363	etoposide	3	0.6689	irinotecan	32
0.6596	luteolin	32	0.6900	camptothecin	14
0.6675	tyrphostin_AG-825	14	0.6921	etoposide	3
0.6877	daunorubicin	14	0.6926	mycophenolic_acid	75
0.6882	thioguanosine	32	0.6996	phenoxybenzamine	104
0.6903	hycanthone	3	0.7175	doxorubicin	14
0.7012	0175029-0000	32	0.7258	0175029-0000	32
0.7155	chrysin	32	0.7336	mepacrine	16
0.7251	resveratrol	3	0.7431	5151277	7
0.7360	GW-8510	14	0.7438	apigenin	32
0.7402	8-azaguanine	89	0.7521	5109870	3
0.7525	5109870	3	0.7579	vorinostat	16
0.7557	mycophenolic_acid	75	0.7598	scriptaid	16
0.7599	piperlongumine	40	0.7626	alsterpaullone	14
0.7635	methotrexate	75	0.7640	resveratrol	3
0.7805	5162773	7	0.7724	cytochalasin_B	100
0.7828	sanguinarine	89	0.7824	piperlongumine	40
0.7891	exemestane	49	0.7858	tyrphostin_AG-825	14
0.7901	ciclopirox	3	0.7896	HC_toxin	16
0.7902	digitoxigenin	63	0.7934	trifluridine	3
0.7937	harmine	32	0.7996	MG-262	40
0.8000	alsterpaullone	14			
0.8006	ellipticine	14			
0.8010	menadione	89			
0.8027	dopamine	44			
0.8041	sulindac_sulfide	91			
Distance	**Community**		**Distance**	**Community**	
0.0888	32		0.0978	14	
0.1174	14		0.1458	3	
0.1434	3		0.19	16	
0.2581	89		0.2374	32	
0.3798	75		0.3955	40	

Community Enriched for CDK/Topo inhibition

Figure F.5: Nighborhood of the tested compounds 5 -

Appendix G

Impact of rank merging on the performances

Figure G.1: Impact of rank merging on the performances 1

PHA-848125 on MCF7

Distance	Compound
0.5318	0175029-0000
0.5493	GW-8510
0.5973	alsterpaullone
0.6098	apigenin
0.6163	daunorubicin
0.6346	camptothecin
0.6399	irinotecan
0.6474	H-7
0.6508	thioguanosine
0.6607	ellipticine
0.6796	doxorubicin
0.6840	tyrphostin_AG-825
0.6847	luteolin
0.6848	staurosporine
0.7146	azacitidine
0.7152	digitoxigenin
0.7167	harmine
0.7194	proscillaridin
0.7202	helveticoside
0.7206	piperlongumine
0.7248	8-azaguanine
0.7295	sanguinarine
0.7305	ouabain
0.7365	digoxin
0.7397	phenoxybenzamine
0.7465	etoposide
0.7497	digoxigenin
0.7498	dexverapamil
0.7498	chrysin
0.7507	lanatoside_C
0.7531	mitoxantrone
0.7576	thiostrepton
0.7577	fisetin
0.7595	methotrexate
0.7634	triflusal
0.7653	phthalylsulfathiazole
0.7655	5114445
0.7666	0297417-0002B
0.7669	sulconazole
0.7673	5109870
0.7692	rimexolone
0.7716	bisacodyl
0.7726	Prestwick-1084
0.7730	medrysone
0.7737	menadione
0.7749	tyloxapol
0.7776	ginkgolide_A
0.7785	DL-thiorphan
0.7816	trazodone
0.7817	doxazosin
0.7831	norcyclobenzaprine
0.7842	meticrane
0.7858	sulfametoxydiazine
0.7860	omeprazole
0.7871	celastrol
0.7882	skimmianine
0.7887	beta-escin
0.7901	5152487
0.7901	trioxysalen
0.7921	benperidol
0.7933	piperidolate
0.7989	repaglinide
0.7990	oxamic_acid
0.7992	Prestwick-559
0.7998	hexestrol
0.8000	0179445-0000
0.8000	tobramycin
0.8005	metyrapone
0.8005	carbachol
0.8013	eucatropine
0.8020	gliclazide
0.8021	solanine
0.8031	zoxazolamine
0.8032	procaine
0.8037	6-azathymine
0.8041	verteporfin
0.8042	etomidate
0.8050	liothyronine
0.8062	natamycin
0.8074	MS-275
0.8075	cycloserine

Figure G.2: Impact of rank merging on the performances 2 -

PHA-848125 on MCF7, SF539

Distance	Compound
0.6181	0175029-0000
0.6350	GW-8510
0.6412	alsterpaullone
0.6538	thioguanosine
0.6636	daunorubicin
0.6675	apigenin
0.6690	ellipticine
0.6801	harmine
0.7062	irinotecan
0.7174	proscillaridin
0.7209	staurosporine
0.7229	luteolin
0.7248	camptothecin
0.7260	ouabain
0.7263	digitoxigenin
0.7280	helveticoside
0.7406	tyrphostin_AG-825
0.7413	8-azaguanine
0.7421	piperlongumine
0.7425	doxorubicin
0.7540	digoxin
0.7541	digoxigenin
0.7657	phenoxybenzamine
0.7723	H-7
0.7786	thiostrepton
0.7790	lanatoside_C
0.7796	bisacodyl
0.7844	fisetin
0.8005	chrysin

Figure G.3: Impact of rank merging on the performances 3 -

PHA-848125 on A2780, MCF7, U251

Distance	Compound
0.5674	apigenin
0.5765	0175029-0000
0.5997	irinotecan
0.6167	harmine
0.6223	thioguanosine
0.6405	camptothecin
0.6451	GW-8510
0.6451	luteolin
0.6462	alsterpaullone
0.6533	daunorubicin
0.6629	ellipticine
0.6692	phenoxybenzamine
0.6851	piperlongumine
0.6999	tyrphostin_AG-825
0.7013	8-azaguanine
0.7091	H-7
0.7095	0297417-0002B
0.7114	staurosporine
0.7143	proscillaridin
0.7186	chrysin
0.7198	ouabain
0.7230	digitoxigenin
0.7307	helveticoside
0.7376	etoposide
0.7387	digoxigenin
0.7432	harmol
0.7439	doxorubicin
0.7562	digoxin
0.7570	nocodazole
0.7592	sanguinarine
0.7657	menadione
0.7695	STOCK1N-35215
0.7730	hycanthone
0.7765	resveratrol
0.7766	1,4-chrysenequinone
0.7788	azacitidine
0.7801	mebendazole
0.7805	lanatoside_C
0.7807	methotrexate
0.7826	thiostrepton
0.7842	bisacodyl
0.7880	PNU-0293363
0.7932	mitoxantrone
0.7975	Prestwick-559
0.8036	PHA-00665752

Well Known cdk2-inhibitors

PHA-848125 on A2780, MCF7, U251, SF539

Distance	Compound
0.5673	apigenin
0.5765	0175029-0000
0.5997	irinotecan
0.6164	harmine
0.6223	thioguanosine
0.6405	camptothecin
0.6451	GW-8510
0.6451	luteolin
0.6462	ellipticine
0.6533	daunorubicin
0.6629	alsterpaullone
0.6692	phenoxybenzamine
0.6851	piperlongumine
0.7000	tyrphostin_AG-825
0.7013	8-azaguanine
0.7091	H-7
0.7095	0297417-0002B
0.7114	staurosporine
0.7143	proscillaridin
0.7186	chrysin
0.7198	ouabain
0.7230	digitoxigenin
0.7307	helveticoside
0.7376	etoposide
0.7387	digoxigenin
0.7432	harmol
0.7439	doxorubicin
0.7562	digoxin
0.7570	nocodazole
0.7592	bisacodyl
0.7657	menadione
0.7695	STOCK1N-35215
0.7730	hycanthone
0.7765	resveratrol
0.7766	1,4-chrysenequinone
0.7788	azacitidine
0.7801	mebendazole
0.7805	lanatoside_C
0.7807	methotrexate
0.7826	thiostrepton
0.7842	sanguinarine
0.7880	PNU-0293363
0.7932	mitoxantrone
0.7975	Prestwick-559
0.8036	PHA-00665752
0.8038	luteolin

Figure G.4: Impact of rank merging on the performances 4 -