UNIVERSITA' DEGLI STUDI DI VERONA

DEPARTMENT OF COMPUTER SCIENCE

GRADUATE SCHOOL OF NATURAL SCIENCES AND ENGINEERING

DOCTORAL PROGRAM IN COMPUTER SCIENCE

CYCLE XXXIV

Artificial Intelligence Techniques Integrate Biological Omics into Graphs for the Prediction and Pathway Analysis of Patient's Disease

S.S.D. INF/01

Coordinator: Prof. Massimo Merro

Tutor: Prof. Rosalba Giugno

Doctoral Student: Luca Giudice

2021/2022
Artificial Intelligence Techniques Integrate Biological Omics into Graphs for the Prediction and Pathway Analysis of Patient's Disease — LUCA GIUDICE PhD Thesis

LIST OF CONTENT

ABSTRACT (ENGLISH)

A person's phenotype refers to the observable physical properties of the organism. The phenotype is determined by the genotype, which is the set of organism's genes. The latter do not act alone but are regulated by other molecules in the same cell. The correlation between genotype and phenotype is the statistical relationship that binds one or multiple genes and their regulators with an observable physical property. When the phenotype is the status induced by the disease in a patient, finding a correlation helps diagnosis, prognosis, and treatment. However, accomplishing such task is not trivial because dysfunctional genes and regulators change among patients even if they share same disease and clinical conditions. One approach consists into looking for the altered cellular functions (i.e., pathways) instead of single dysfunctional actors.

The standard pathway analysis requires data describing the cell's molecular profiles of two classes of patients. One class composed by patients with the disease in study and one which includes control. A molecular profile is captured with wet-lab protocols from a patient and describes the activity of its molecules. A pathway is found significantly deregulated if the molecules performing that cellular function are co-ordinately more active or less active in the diseased profiles with respect the control ones.

Enrichment methods for the pathway analysis are easy to use, understand and provide biologically meaningful results. However, they suffer of two limits. They do not consider the interconnected nature of the molecules which are involved both inside and outside their own pathway. They are not able to learn the resulting pathways found in a patient class and to use this information to recognise or test profile of new patients which class is unknown or insecure. On the contrary artificial intelligence techniques, more precisely machine learning algorithms, can overcome these limitations but are not under development or interesting because they are difficult to use, tune, understand and to design for providing different insights with respect the simple enrichment counterpart.

This thesis focuses on pathway-based patient classifiers based on patient similarity networks. A recent concept that benefits of two characteristics; it learns pathway

information to predict the patient phenotype and it works with patient similarity networks as features to classify. These characteristics allow to build a classifier which is interpretable, able to accept different type of biological data and to provide new insights about the patients and the phenotypes in study; resulting a valid alternative or even better than enrichment methods.

As last, this thesis describes a state-of-art artificial intelligence algorithm for predicting the effect of an altered molecule over the rest of the cell and how this strategy is integrated into pathway analysis methods both of enrichment and of machine learning for considering the interconnected nature of the molecules.

ABSTRACT (ITALIAN)

Il fenotipo di una persona si riferisce alle sue proprietà fisiche osservabili. Il fenotipo è determinato dal genotipo, che è l'insieme dei geni dell'organismo. Questi ultimi non agiscono da soli, ma sono regolati da altre molecole nella stessa cellula. La correlazione tra genotipo e fenotipo è una relazione statistica che lega uno o più geni e i loro regolatori ad una proprietà fisica osservabile. Quando il fenotipo è lo stato indotto di una malattia, trovare questo tipo di correlazione aiuta la diagnosi, la prognosi e il trattamento del paziente affetto. Tuttavia, trovare tale connessione non è banale perché i geni e i regolatori disfunzionali (non più capaci di svolgere la loro funzione normale nella cellula per il bene dell'organismo) causati della malattia non sono sempre gli stessi e possono cambiare tra pazienti. Un approccio per risolvere questo problema consiste nel trovare le funzioni cellulari alterate (cioè i pathway) invece che i singoli attori disfunzionali. Questa operazione si chiama analisi dei pathway o pathway analysis.

L'analisi standard dei pathway richiede dati che descrivono i profili molecolari delle cellule di due classi di pazienti. Una classe composta da pazienti con la malattia/fenotipo in studio e una che include il controllo. Un profilo molecolare è ottenuto con protocolli di laboratorio da un paziente e descrive l'attività delle sue molecole. Un pathway è significativamente e statisticamente deregolato se le molecole che svolgono quella funzione cellulare sono coordinatamente più attive o meno attive nei profili malati rispetto a quelli di controllo.

I metodi standard per l'analisi dei pathway (enrichment methods in inglese) sono facili da usare, da capire e forniscono risultati biologicamente significativi. Tuttavia, soffrono di due limiti. Non considerano la natura interconnessa delle molecole che sono coinvolte sia all'interno che all'esterno del loro pathway. Non sono in grado di imparare i pathway che trovano significativi in una classe di pazienti e di utilizzarli per riconoscere o testare il profilo di nuovi pazienti la cui classe è sconosciuta o insicura. Al contrario, le tecniche di intelligenza artificiale, più precisamente gli algoritmi di apprendimento automatico, possono superare queste limitazioni, ma non sono né in fase di sviluppo né di interesse perché sono

difficili da usare, capire e progettare per ottenere informazioni diverse rispetto a quelle che già ottengono i metodi standard.

Questa tesi si concentra su classificatori di pazienti che integrano le informazioni dei pathway e lavorano su grafi di similarità. Un concetto recente di algoritmo che beneficia di due caratteristiche: impara quali sono i pathway che permettono di predire una malattia o fenotipo e lavora con grafi di similarità dove ciascun paziente in studio è rappresentato come un nodo. Queste caratteristiche permettono di costruire un classificatore interpretabile, capace di accettare diversi tipi di dati biologici e di fornire nuove informazioni sui pazienti e i fenotipi in studio, risultando una valida alternativa o addirittura migliore dei metodi di enrichment.

Infine, questa tesi descrive un algoritmo di intelligenza artificiale allo stato dell'arte che predice l'effetto di una molecola alterata sul resto della cellula e come questa strategia è integrata nei metodi di analisi dei pathway sia standard che di apprendimento automatico per considerare la natura interconnessa delle molecole.

1. INTRODUCTION

A person's phenotype refers to the observable physical properties of the organism; these include the appearance, development, behaviour, and status due to a disease. The phenotype is determined by the genotype, which is the set of organism's genes (cell's basic physical and functional units made up of DNA encodes the synthesis of RNA and protein molecules).

The correlation between genotype and phenotype is the statistical relationship that binds one or multiple genes with an observable physical property; its study is priority for understanding and defining always better a patient's disease; finding it between genes and a patient's clinical information helps diagnosis, prognosis, and treatment. A mutated TP53 gene is a marker for cancer because helps altered cells to grow and proliferate. However, proving a correlation is not trivial.

Genes must be annotated in structure and functionality. Same gene's features vary between cell type, tissue, organ, and person; reason why the four factors are chosen due to the disease in research. For example, Lung cancer study leads to sample cells from the neoplasm present in the lungs of diagnosed patients. Genes do not act alone but are regulated (promoted or repressed) by other molecules that are present in the same cell. The latter can be altered by the same disease due to different deregulated genes; reasons why, correlating genotype and disease's phenotype is currently passing through two levels of findings.

First, correlation with the structural and functional properties of the genes and their regulators playing a role in altering the normal functions of the patient's cell. Second, correlation with the functions performed by the altered cells such that is possible to recognise and explain the disease also if the actors change between individuals. For example, the TP53 gene is mutated in almost every type of cancer at rates from 38%–50%. The mutation leads cancer cells to grow uncontrolled and to quickly proliferate. A high activity in cell proliferation and division is one signature of cancer cells also if TP53 is not mutated.

In fact, TP53 contains the information that allows a cell to synthetize the transcription factor, protein and molecule called p53. The latter controls the cell

division and cell death. p53 exerts its function through regulation of genes as DR4, DR5, DCR1, DCR2, FASL/CD95L and FAS. p53 production is controlled by a group of regulators called miRNAs as miR-143, miR-145 and miR-34a. TP53 (gene), p53 (protein), the target regulated genes, the regulatory miRNAs form a complicated p53 pathway. In unstressed cells, p53 production is kept low through a continuous degradation of the protein. It is activated and increased in response to stressors, including but not limited to DNA damage, oxidative stress, osmotic shock, ribonucleotide depletion, and deregulated oncogene expression. In this last case, p53 tries to lead the death of the altered cells, however especially cancer can alter the normal function of one element in p53 pathway and allow the cells to grow and proliferate. As advantage, the pathway space is more robust and less dependent by both individual and environmental variables. Pathways are strongly used to annotate a disease phenotype which exhibits high variability among different persons at the level of the single altered molecules.

We can redefine the correlation between genotype and a disease phenotype as the statistical relationship that binds one or multiple deregulated pathways shown by the altered cells of the diseased patients and their status/clinical condition.

A deregulated pathway is determined by studying the cell's components. This means to get the molecules which are presents in the cell, how much they have been produced (i.e., expressed, codified) and which functions are performing. We can divide the molecular information in layers called omics. Each layer details a specific type of molecules. For example, the omic called proteomic provides information about which proteins are present in the cell, which are the most important and how they are regulated. The advent of the next generation sequencing techniques boosted and supported the study of omics. This due the fact that, next generation sequencing is a technology capable of capturing the activity and the presence of thousands of genomic (molecules are genes, mutations, and variants), transcriptomic (molecules are transcripts), proteomic (proteins) and metabolomic (metabolites) elements.

At this stage, bioinformatics enrichment tools already allow to perform an omic study to find correlations between genotype and a disease phenotype. They require

the sequencing of the diseased and control patients. For example, a bulk mRNA sequencing 75 DNA base pairs generates a patient's molecular profile composed of at least 20 thousand genes which activity and expression is measured with an continuous value. Follow two standard operations of analysis called differential expression analysis and pathway analysis. The first one finds the molecules, genes in this case, which are statistically different between the disease patient class and the reference one in term of expression. It allows to understand which molecules are deregulated due to the disease. The second operation finds the altered pathways. It requires a list of pathways P, obtainable from databases which collect associations between molecules and cellular functions, and the ranked list of the differentially expressed genes R (from the most upregulated to the most downregulated). Then, it determines whether members of a pathway P tend to occur together toward the top (or bottom) of the list R, in which case the gene set is correlated with the phenotypic class distinction.

The just described operation for performing the pathway analysis is one of the many strategies that have been developed to get the deregulated pathways by a patient class and so associated to its representing disease phenotype. However, enrichment tools share two limitations. They do not consider the interconnected nature of the molecules which are biologically and functionally involved inside but also outside their own pathway. They are not able to learn the resulting pathways found in a patient class and to use this information to recognise or test new patients which class is unknown or insecure; task which is then required to perform manually by biomedical scientists and clinicians.

The first limitation leads the enrichment method to biologically consider only what is happening inside a pathway and to miss the participation of other molecules which impact the biological process indirectly. In fact, cell's molecules do not work in isolation: they communicate and coordinate with each other to carry out their function even with elements outside their pathway. For example, this interconnected nature causes dysfunctional elements to propagate alterations or erroneous signals throughout the entire cell, they do not stop with affecting only the pathway in which they perform a task.

About this point, system biologists are unveiling molecular interactions and allowing the creation of biological networks providing an intuitive data structure to investigate the signaling cascades. For example, a protein-protein interaction network represents proteins as person of a social network and the relationships between them are represented as edges. A relationship is defined based on the physical interaction. Proteins that are close in the network have a similar function or even performing a task for the same pathway. This permits to understand the effect of an alteration on the overall cell and to infer new knowledge from a priori information. When a protein as p53 is found altered in a patient's cancer cell is possible to predict that also its interactors are going to be affected and deregulated.

A biological network can be seen as a graph G defined as pair G=(V,E) where V is a set of vertices (aka nodes) and E is a (multi)set of unordered pairs of vertices. The elements of E are called edges. The vertices represent molecules and the edges their connections. This aspect boosted the interest and the development of artificial intelligence algorithms that, accounting for the global structure of the network and the interconnected nature of the molecules, predict the impact of an altered molecule over the others. These methods share the paradigm of network-based propagation. Prior information indicating the activity, as the expression, associated to sequenced molecules is imposed on the respective vertices of the network and then is propagated through the edges to nearby nodes. In this way, not sequenced molecules get a "guilty" score based on the information received by the others. While sequenced molecules see their prior information adjusted based on their neighbours; for example, their expression increases in case they are connected to other molecules with a high expression or the opposite. The final score defines how much the molecule is involved and active due to the overall alterations of the patient's cell.

Enrichment methods for the pathway analysis are starting to consider propagation methods and in general biological networks to overcome their limitation. However, they cannot overcome the learning issue.

At the same time, Bioinformatics is evolving as an integrative field between computer science and biology. The network-based propagation is one example in which bioinformatics exploits graph theory applied to biological data. Another mathematics and computer science branch that is becoming more and more used in bioinformatics is the machine learning.

Machine learning (ML) studies the use of computers to simulate human learning by exploring patterns in the data and applying self-improvement to continually enhance the performance of learning. Machine learning algorithms can be roughly divided into supervised learning algorithms, which learn to map input example into their respective output, and unsupervised learning, which identify hidden patterns in unlabelled data. The advances made in machine-learning over the past decade transformed the landscape of data analysis. However, clinical applications have been limited, and follow-up mechanistic investigation of ML-based predictions is often lacking due to their difficulty in being used and understood from non-experts. As result, machine learning algorithms are not currently cited in the best bioinformatics practices for analysing sequencing data. There are mostly algorithms to reduce false positive findings obtained from the standard differential expression analysis and only one algorithm able to perform a pathway analysis.

Entering more in details, this thesis presents the study and development of software for the pathway analysis based on machine learning. Given a dataset of patient's profiles described by a biological omic such that a subset represents the disease of interest, and the remaining ones represent the control group, the software learns the pathways that most differentiate the two groups and then predicts the class of an unknown patient based on them. The correctness of the prediction measures the quality of the selected pathways and the ability of the machine learning classifier to generalize.

Pathway-based patient classification can be defined as a supervised learning task which goal is to support the decision-making process of human experts in biomedical applications providing signature pathways associated to a patient class characterized by a specific clinical phenotype.

However, there are currently very few published pathway-based patient classifiers due to specific drawbacks in developing such type of machine learning software. It is prone to lack a formal statistical basis, be computationally expensive, includes a not trivial hyper-parameter setting and does not handle neither imbalances classes nor structured feature types as the biological pathways. It does not hold the comparison against standard pathway analysis tools which are fast, popular, easy to use and understand. The only pathway-based classifier is of Hao et al. . A deep neural network called PASNet for the prediction of Glioblastoma patients. The method builds a network model by leveraging prior biological knowledge of pathway databases and predicts considering hierarchical nonlinear relationships between the biological processes and the patient classes in comparison.

This thesis illustrates three software of machine learning that on the contrary of enrichment methods, they offer new insights about pathway, generate data regarding the patients, natively support multiple data types, accept different biological patient data in input and are able to use the significative pathways to test and predict the class of unknown patients. While at the same time, they are more and more reducing the gap with the main benefits of enrichment methods in being easy to use and understand.

The first one is netDx. A patient classifier based on the patient similarity network (PSN) paradigm. In a PSN, each node is an individual patient and an edge between two patients corresponds to pairwise similarity for a given datum (gender, height, gene expression …). The paradigm brings many advantages. Analysing the similarities to gather new information is conceptually intuitive. A PSN can lead to the identification of patient subgroups or the prediction of a patient's class/outcome. Similarity networks can represent any datum, naturally handle missing and heterogenous data, have a history of successes in gene and protein function prediction.

netDx converts any available patient's datum (e.g., age, gender, gene expression, …) into a PSN. The method proceeds by testing the similarity networks. It scores each input PSN based on how well it classifies the training patients (patients which class is known are used to learn the best PSNs). A linear combination of the best

PSNs is then used to create a composite network on which the unknown patients (i.e., testing patients) are classified based on their similarity with the training ones. In case, netDx receives a biological omic about the patients, it creates a PSN for each pathway, if the classification provides sufficient results, then it returns the best PSNs/pathways used for the prediction of the testing patients.

The second software is Pratic. A module of netDx that has been developed to learn pathways and classify the patients described by sparse molecular profiles. Precisely, the patient's profiles are binary, and the genes have only a value either equal to 1 indicating a mutation or 0 otherwise. Given the biological fact that mutated genes are rare and do not compose even the 1% of a profile usually composed by 20 thousand genes, patient's profiles are difficult to compare because they rarely share mutations. Pratic extends netDx and uses a network-based propagation algorithm to convert the patient's profiles from binary to dense.

A network-based propagation algorithm exploits the interconnected nature of the molecules. It gives a guilty score to every molecule based on how much they are close and so likely to be altered by the mutated genes. Then it replaces the original binary value with the guilty scores. Pratic then creates the patient similarity networks using an ad-hoc measure which evaluates how much two patients are similar due to the guilty scores and passes them to netDx.

Simpati is the third classifier, and it evolves netDx and Pratic. It requires the patient's biological profiles (e.g. genes per patients) divided into classes based on a clinical information (e.g. cases versus controls). It prepares the profiles singularly applying guilty-by-association approach called network-based propagation to determine how much each molecule and single biological feature is associated and involved with the other ones and so to the overall profile. Higher is the guilty score and more the feature is involved in the patient's biology. Simpati proceeds by building a pathway-specific patient similarity network (psPSN). It determines how much each pair of patients is similarly involved in the pathway with an ad-hoc novel measure. If the members of one class are more similar (i.e. stronger intra-similarities) than the opposite patients and the two classes are not similar (i.e. weak inter-similarities), then Simpati recognizes the psPSN as signature. If the classes

are likely to contain outlier patients (i.e. patients not showing the same pathway activity as the rest of the class), then Simpati performs a filtering to keep only the most representative members of each class and re-test the psPSN for being signature. For this task, it exploits a graph-theory concept called cohesive subgroup detection. Unknown patients are then classified in the best pathways based on their similarities with known patients and on how much they fit in the representative subgroups of the classes (i.e. more they are similar to the representatives of a class and more they fit). As results, Simpati provides the classes of the unknown patients, the tested statistically significant signature pathways divided into up and down involved (new pathway activity paradigm based on similarity of propagation scores), the biological features which contributed the most to the similarities of interest, the guilty scores associated to the biological features and all the data produced during the workflow in a vectorial format easy to share or analyse.

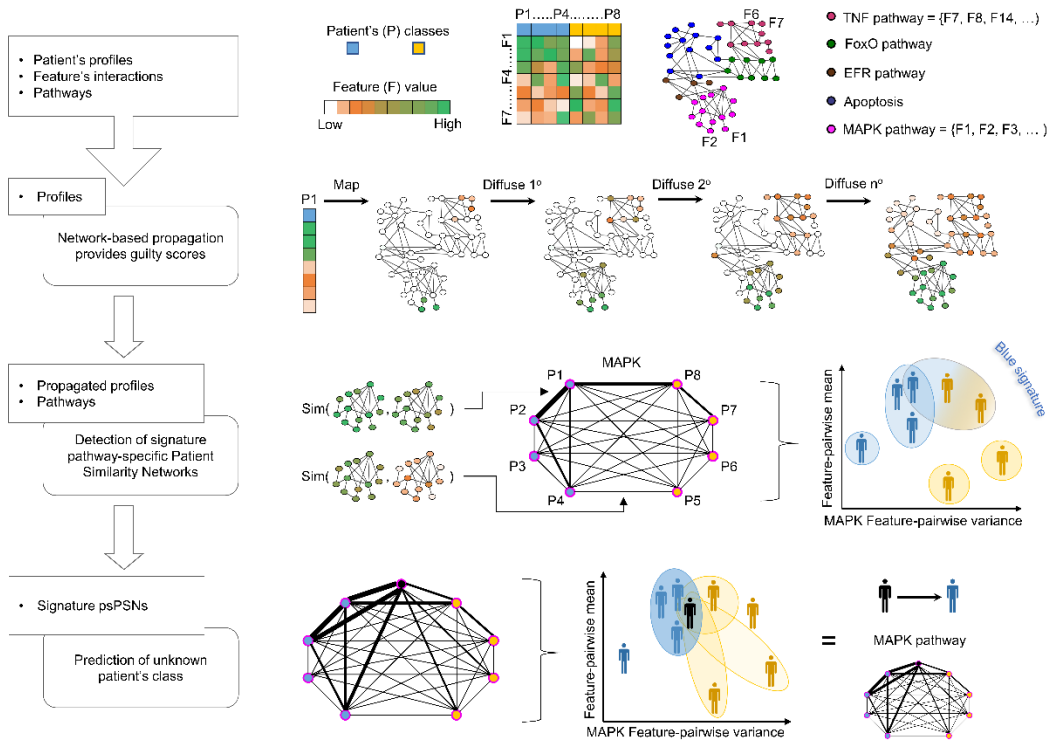


Fig.1.1 Workflow of Simpati. Patient profiles are divided in two classes and are described by biological features. A feature-feature interaction network together with pathways are further input data required by the software (e.g., gene-gene interaction network and KEGG pathways). All profiles are individually propagated over the network. The profile's values are replaced by scores that reflect the feature's starting information and interactions. Simpati proceeds by creating a patient similarity network for each pathway (psPSN). The pairwise similarity evaluates how much two

patients have a similar pathway activity. It evaluates how much the features between two patients are close and high in term of propagation values. Two patients that act on a pathway with the same features and same expression values get the maximum similarity. In the figure, more an interaction is thick and more the two corresponding patients are similar. The psPSN is recognised as signature if one class is cohesive, one is sparse, and the two classes are not similar. In case of signature pathway, an unknown patient is classified based on how much is like the other patients.

2. BIOLOGICAL BACKGROUND

This chapter summarizes concepts about molecular biology; they are considered by machine learning classifiers to integrate biological data and cellular functions for learning biologically meaningful features based on the patient's information.

- Section 2.1 introduces the terminology and the biological subjects of a pathway analysis and patient classification: organism, sample, and cell. It describes the structure of the cell and its role in enclosing the DNA. It presents the central dogma of molecular biology and its limitations. It delineates the roles of the genes and their regulators.
    - Pathway-based classifiers refer to features as molecules (e.g. genes, proteins, etc..) and to biological data as information describing the biology of the patients in study. This section introduces the biological objects and subjects that the machine learning algorithms have to handle and connect for performing a pathway analysis able to provide biologically meaningful results.
    - Pathway-based classifiers are tools which aim to perform a biological oriented analysis. This section highlights the reasons of pursuing such aim.
- Section 2.2 introduces the biological data types also called omics that are used to describe an organism's biology. It highlights the differences between the types and provides examples about how they are represented from a mathematical perspective.
    - Artificial intelligence techniques consider this knowledge to decide which operation to perform for which biological datum describing the subjects in study. For example, when patients are described by biological molecules (i.e. features) that are measured with a continuous numeric value, then the classifier can perform a large variety of techniques in the processing and feature selection phase. This does not hold when the classifier has to elaborate patient's features that are measured with a binary value equal to 0 or 1.

    - Pathway-based classifiers proposed in this thesis can work with multiple biological data types. However, they handle the diversity of the data in different ways. netDx requires to set up an ad-hoc similarity function for each biological datum. Pratic handles binary somatic mutation data and proposes one specific patient similarity function. Simpati performs a standardization on any input datum to always get the same meaning and type of information for building the patient similarity networks.
    - Pathway-based classifiers require specific meta-information based on the biological datum describing the patients.
- Section 2.3 introduces high-throughput technologies. They allow to massively generate information about all the biological molecules involved in the activity of the patient's cell. It is due to high-throughput technologies that machine learning algorithms can be developed. For example, a high-throughput technology as the bulk mRNA sequencing allows to get the activity (i.e. expression) of twenty thousand genes in a patient. Other high-throughput technologies allow to get the quantification of the gene's regulators and proteins.
    - Pathway-based classifiers can currently work only on high-throughput data. The reason is that more molecules describe a patient's biology and more pathways can be defined and characterized.
    - High-throughput data allow also to better model the interconnected nature of the molecules. Having only few hundreds of molecules detected would mean to necessarily infer in-silico the activity of the others in the same cell or to take assumptions of activation and inhibition due to lack of knowledge with respect the overall cell biology which counts for 50 thousand annotated genes, 400 thousand annotated proteins and so on.
- Section 2.4 introduces the concept of biological networks and explains the differences between: signaling networks, protein-protein interaction

networks, metabolic networks, gene regulatory networks and genetic interaction networks.

  - The interconnected nature of molecules can be represented by a network. The latter is a mathematical model also defined as graph that is used to understand how the patient's altered molecules have an impact on the organism's cell functions independently by their specific single role. In other words, the network allows to track the signaling cascade of a molecule and to understand which pathways are affected and regulated by it.
  - Pratic and Simpati classifiers use biological networks to infer a new activity score out the raw one obtained from a high-throughput technology for each captured molecule. A molecule gets a new imputed score that summarizes its original activity value and how all the other molecules have an impact on its function. Pratic uses biological networks to infer information, while Simpati to standardize biological data.

- Section 2.5 introduces the biological processes and functions which a cell performs during its lifetime. Looking at the pathways instead to the single components eases the recognition of what activity the cell is running.
  - The pathway space leads an analysis which is more robust, accurate and easy to understand.
  - Pathway-based classifiers integrate this information to get features for classifying the patients in study and to perform a pathway analysis; the latter is a task that allows to get strong correlating findings between genotype and the patients phenotypes.

2.1 Organism, Sample and Cell

An organism refers to a living thing that has an organized structure, can react to stimuli, reproduce, grow, adapt, and maintain homeostasis. An organism would, therefore, be any animal, plant, fungus, protist, bacterium, or archaeon on earth. These organisms may be classified in various ways. One of the ways is by basing

upon the number of cells that make it up. The two major groups are the single-celled (e.g. bacteria, archaea, and protists) and the multicellular (animals and plants).

A biological sample refers to a biological material (such as blood, urine, tissue, cells, cell cultures or saliva) collected from a subject and organism in study, in this thesis mainly human and mouse. A sample differs by the organism from which has been captured. A sample captures the cells (basic building blocks) of the organism in a specific time point, organ and tissue. Biological research uses a sample to acquire information about an organism's cell and this is why the terms are often interchangeable. For example, the cell in the sample obtained from the lungs of a patient has altered molecules affecting the normal function of respiration pathway; the cell respiration pathway is assumed to be part of the cause or consequence of the patient's clinical status in study.

Cells are the basic building blocks of all organisms [1]. They provide the structure and the fundamental functions to keep all living organisms alive. Cells are autonomous and self-sustaining, but they exchange information with the environment and with the other cells, through a variety of chemical and mechanical signals, in order to accomplish various biological tasks. In general, cells can be divided into two main types: prokaryotic and eukaryotic. Both eukaryotic and prokaryotic cells share several features: a layer, called cell membrane, that separates the cell from the external environment, a medium, called cytoplasm, in which the biochemical reactions of the cell take place and the use of deoxyribonucleic acid (DNA) to store their genetic information. In both eucaryotic and procaryotic cells, DNA is transcribed into ribonucleic acid (RNA) and translated into proteins (aka specific type of molecules) through the process of translation that is facilitated by specialized macromolecular machines called ribosomes. The main difference between prokaryotic and eukaryotic cells is that the latter have a membrane-bound nucleus. The nucleus is surrounded by a membrane, called nuclear envelope, which contains and protect the DNA. The DNA determines the entire structure and function of the cell deciding if the cell has to grow, mature, divide or die.

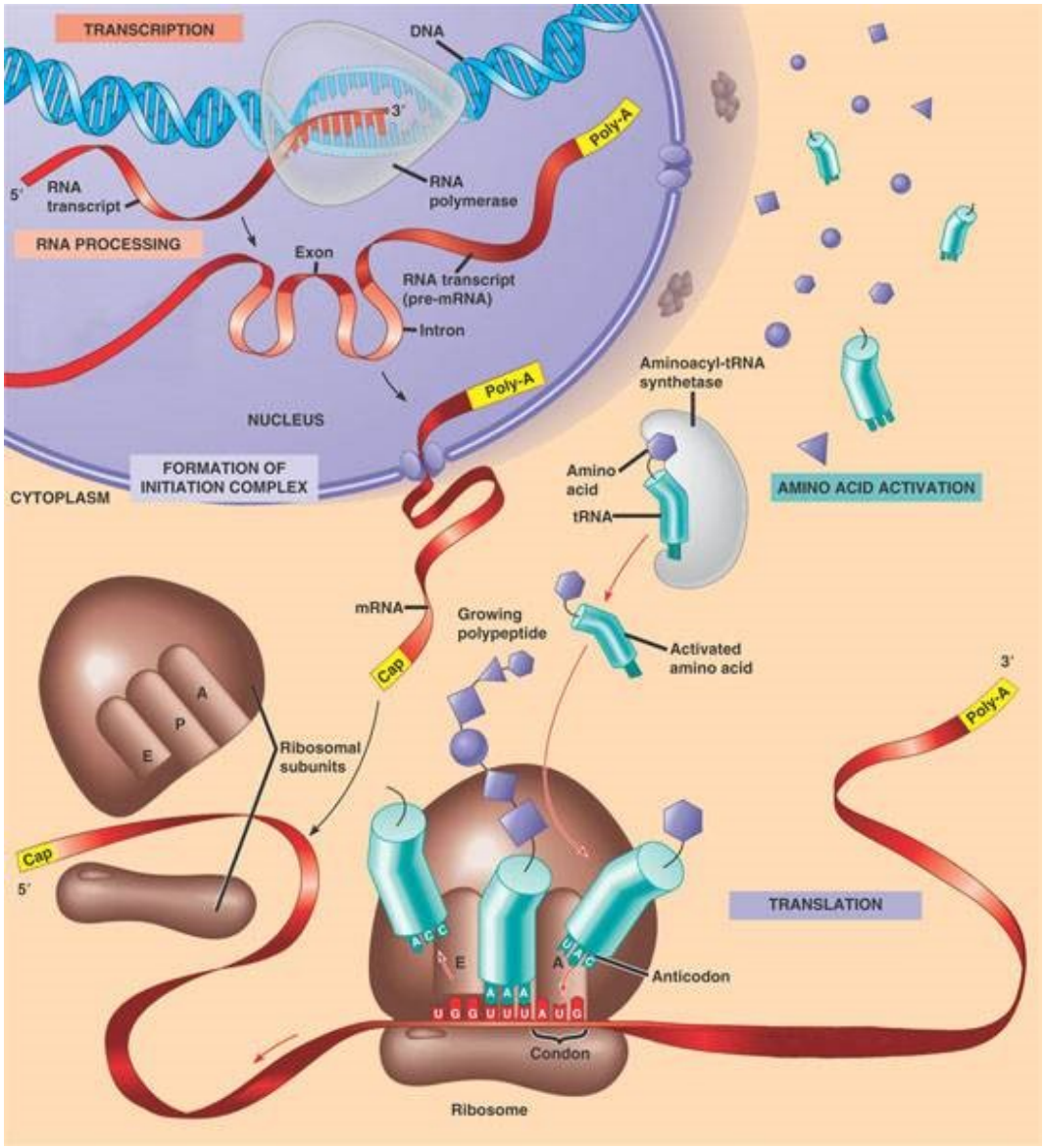


Figure 2.1 Representation of the main actions that are performed in and out a eukaryotic cell nucleus for producing proteins.

DNA is a double-stranded molecule characterized by subunits called nucleotides. Each nucleotide is composed by a phosphate group, a sugar group and a nitrogen base. The nitrogen bases that identify each nucleotide are Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). The order in which the bases appear in the DNA sequence encodes the instructions contained in the genes, that are the fundamental units used to make proteins which carry out the biological functions. The set of all genes is called genome and when a cell requires to perform a specific function, then it expresses the genes which codify for proteins able to fulfil the task. More the cell has to perform that function and more it expresses the associated genes to get

working proteins able to satisfy the requirement.[1] The flow of information that involves DNA is described by the central dogma of molecular biology [1]. This term was used for the first time by the biologist Francis Crick to express the idea that the processes that can involve DNA are the creation of new DNA from existing DNA (duplication process), the creation of RNA from DNA (transcription process) and the creation of proteins from RNA (translation process) (Figure 2.2). The central dogma states that the genes contain all the necessary information to make proteins and that the RNA, and in particular messenger RNA (mRNA), is the medium to transport this information to the ribosomes. DNA is converted into proteins through the process of gene expression that produces proteins depending on the needs of the cell.

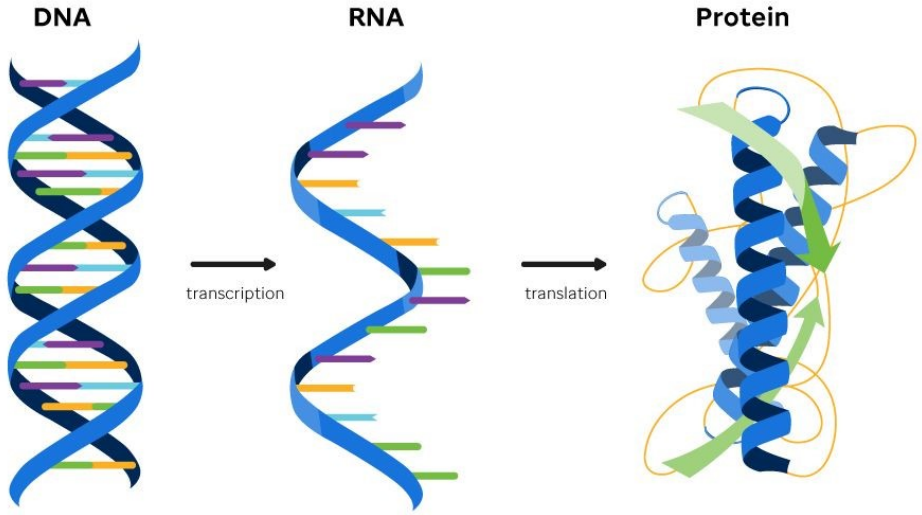


Fig. 2.2. The central dogma of molecular biology. DNA replication is the basis of biological inheritance in all living organisms, and it is the process by which a molecule of DNA is copied to produce two identical DNA molecules. DNA transcription is the process by which the DNA is transformed in RNA through the action of the enzyme RNA polymerase. In the RNA translation process, the ribosomes transform RNA into proteins.

However, the central dogma does not take into account an important process that strongly influences gene expression called gene regulation [2]. This process includes a wide range of mechanisms used by the cell to control the production of specific gene products. Regulation can affect gene expression at each stage, for example at the transcriptional level and post-transcriptional level. Recently, non-coding RNAs (ncRNAs), that are RNA molecules not translated into proteins, have emerged as important regulators of gene expression. Among them, two important

categories of functional ncRNAs include long non-coding RNAs (lncRNAs) and small noncoding RNAs such as microRNAs (miRNAs). The cell can be considered as a dynamic system in which the different types of molecules such as DNA, mRNA, ncRNA, proteins and metabolites are linked together into a complex network of biochemical reactions and interactions; this allows the cell sustenance and its interaction with the environment.

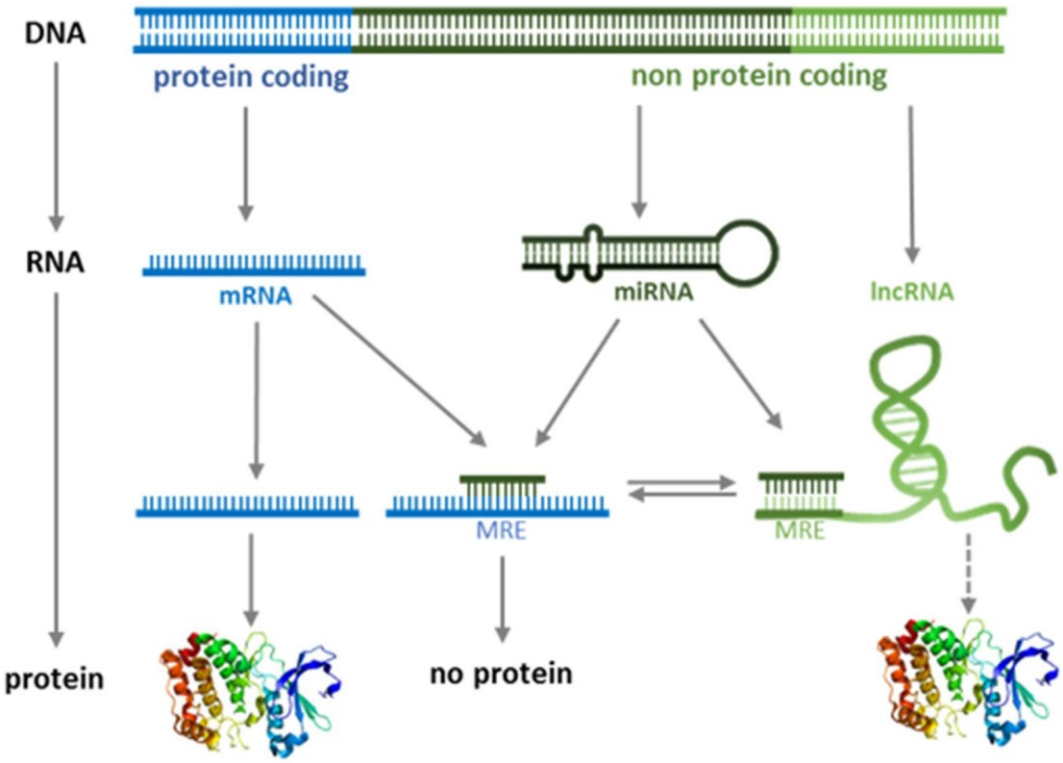


Fig. 2.3. [2] The new central dogma of molecular biology that includes the regulators of the protein coding regions of the DNA (aka genes). The classical "DNA-RNA-protein" model is extended by functional role of non-coding RNAs (ncRNAs), regions of the DNA which codify for molecules that are not used to synthetize proteins but are used to regulate the protein production.

2.2 Omics

Omics can be seen as information layers that describe a cell activity. Each layer details how a specific group of molecules is involved in the cell [3,4]. For example, the omic called proteomic provides information about which proteins are present in the cell and how they are participating. An omic aims at the universal detection of a specific group of molecules such as genes (genomics), mRNA (transcriptomics), proteins (proteomics) and metabolites (metabolomics) in a specific biological sample in a non-targeted and non-biased manner. This can also be referred to as

high-dimensional biology. Omics experiments differ from traditional studies, which are largely hypothesis driven or reductionist. Omics experiments are hypothesis-generating, using holistic approaches where no hypothesis is known or prescribed but all data are acquired and analysed to define a hypothesis that can be further tested. They can be performed not only for the greater understanding of normal physiological processes but also in disease processes where can play a role in screening, diagnosis and prognosis as well as aiding our understanding of the aetiology of diseases [3]. They lead themselves to biomarker discovery as they investigate multiple molecules simultaneously.

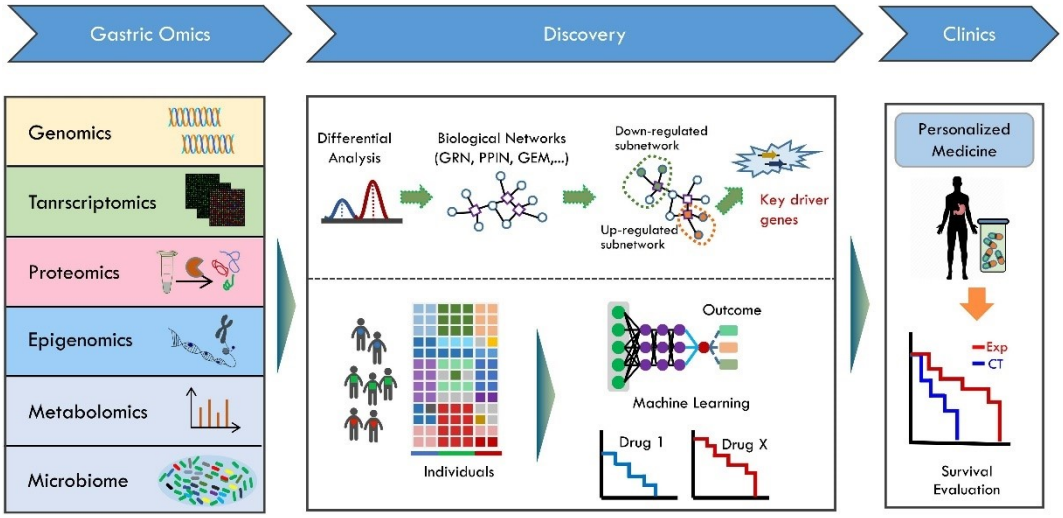


Fig. 2.4. [5] Omics describing the same patient and its cell's components at different levels are the starting information of any bioinformatics analysis which may involve both enrichment tools and machine learning algorithms. They allow to find significantly altered molecules and pathways correlated to the disease phenotype of the patient.

Genomics is the systematic study of an organism's genome. The genome is the total DNA of a cell. The human genome contains 3.2 billion bases and an estimated 30 000–40 000 protein coding genes. Traditionally, genes have been analysed individually but microarray technology has advanced substantially in recent years until the advent of next generation sequencing technologies. They measure differences in DNA sequences between individuals and the expression of thousands of genes simultaneously.

Measuring the differences between the DNA sequence of a patient and the one of the reference organism (i.e. patient vs reference of the human population) allows to reveal mutations. A mutation is a permanent change in the nucleotide sequence of

DNA. DNA can be mutated either by substitution, deletion, or insertion of base pairs. Mutations, for the most part, are harmless except when they lead to cell death or tumor formation.

For example, a mutation may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in a certain stretch of DNA. A mutation shares the same definition with Single nucleotide polymorphisms (SNPs). The difference between the two terms lie on their frequency. SNPs occur normally throughout a person's DNA and mostly have no effect on health or development. Mutations are rare, not shared by the population and are disease related.

From mathematical and genomics perspective, a patient's profile is commonly represented by a binary vector; if a gene is mutated that it has a value equal to 1, while 0 otherwise.

	P1	P2	P3	P4	P5	P6
G1	1	0	0	0	1	0
G2	0	0	0	1	0	1
G3	1	0	0	1	0	0
G4	0	0	1	0	0	1
G5	0	1	0	0	1	0

Fig. 2.5. Heatmap and matrix of patient's somatic mutation profiles. Rows are genes and columns are patients.

Epigenomics is the total epigenetic changes in a cell. Epigenetic changes are changes in the way genes are switched on and off without changing the actual DNA sequence. They may be caused by age and exposure to environmental factors, such as diet, exercise, drugs, and chemicals. Epigenetic changes can affect a person's risk of disease and may be passed from parents to their children.

The transcriptome is the total mRNA in a cell and the template for protein synthesis in the process called translation. The transcriptome reflects the genes that are actively expressed at any given moment. Transcriptomics examines RNA levels

genome-wide, both qualitatively (which transcripts are present) and quantitatively (how much each transcript is expressed). The central dogma of biology viewed RNA as a molecular intermediate between DNA and proteins, which are considered the primary functional read-out of DNA. However, the advent of large transcriptomic studies in the past decade has shown that only ~3% of the genome encodes proteins, up to 80% of the genome is transcribed. A discovery that led to the development of the non-coding RNA field. It is now clear that thousands of non-coding RNAs (long non-coding RNA, small non-coding RNA, circular non-coding RNA) transcribed in cells play essential roles in gene regulation, protein synthesis and pathways. Consequentially, dysregulation of non-coding RNAs had been implicated in various diseases, such as myocardial infarction, diabetes, cancer, and others.

From mathematical and transcriptomics perspective, a patient's profile is commonly represented as a continuous numeric vector such that for each transcript (e.g. gene, miRNA, lncRNA, etc ..) there is a value indicating how much has been transcribed in case of ncRNAs and expressed in case of genes.

	P1	P2	P3	P4	P5	P6
G1	100	67	89	150	122	45
G2	58	35	52	99	12	1
G3	62	22	6	32	13	65
G4	21	54	45	19	64	1
G5	5	15	18	17	28	93

Fig. 2.4. Gene expression matrix of patient's profiles. Rows are genes and columns are patients. The value in a cell is a continuous number greater or equal than 0 which indicates how much a gene is expressed in a determined patient. Higher the expression and more the gene is considered active, important and request by the subject organism.

The proteome is defined as the set of all expressed proteins in a cell. Proteomics aims to characterise information flow within the cell and the organism, through protein pathways and networks, with the eventual aim of understanding the

functional relevance of proteins. While we can gain much information from proteomic investigation, it is complicated by its domain size (>100 000 proteins) and the inability to detect accurately low-abundance proteins. The proteome is a dynamic reflection of both genes and the non-coding regulators.

In the end from proteomics perspective, a patient's profile is commonly represented as a continuous numeric vector such that for each protein there is a value indicating how much has been produced.

Metabolomics can be defined as the study of metabolites in a system (cell, tissue or organism) under a given set of conditions. Metabolomics simultaneously quantifies multiple small molecule types, such as amino acids, fatty acids, carbohydrates, or other products of cellular metabolic functions. Metabolite levels and relative ratios reflect metabolic function, and out of normal range perturbations are often indicative of disease. Metabolomics has a number of theoretical advantages over the other omic approaches. The metabolome is the final downstream product of gene transcription and, therefore, changes in the metabolome are amplified relative to changes in the transcriptome and the proteome. Additionally, as the downstream product, the metabolome is closest to the phenotype of the biological system studied. Although the metabolome contains the smallest domain (~5000 metabolites), it is more diverse, containing many different biological molecules, making it more physically and chemically complex than the other 'omes'.

From mathematical and proteomics perspective, a patient's profile is commonly represented as a continuous numeric vector such that for each metabolite (e.g. glucose) there is a value indicating how much has been synthetised.

Given the enormous amount of data generated in omics studies, bioinformatics and statistics are fundamental. A common issue regards the high numbers of patient's features (gene, proteins, …) which complicate the statistics and increase the likelihood of false positives. This is why, findings of such analysis are validated using in-vitro wet lab protocols. The methods available for analysis comprise various statistical techniques including univariate and multivariate analysis, supervised and unsupervised learning tools and system-based analyses. The aim is to find data patterns that provide useful biological information which can be used

to generate further hypotheses for testing. In-silico data validation is also becoming more and more essential to ensure that findings are not just random. P-values can be corrected for multiple testing (false discovery rate). Other methods of model validation include the use of a 'hold-out 'or 'test' set. The set used in producing the model is called the training set. Models built using the training data can then be independently validated using the hold-out set. An alternative method of independent model validation is to use permutation testing. More robust methods include confirming the observations with a complementary technique and replicating the experiment in a different sample set.

2.3 High-Throughput Data

Biological phenomena relating to transcription, gene regulation or DNA mutation can be measured over the entire genome using high-throughput experimental techniques.

High-throughput methods [6] aim to quantify or locate all or most of the biological features (gene, proteins, ncRNAs) that define an omic. With respect the previous section, the difference between the concepts of omic and high-throughput data is not trivial. An omic refers to a set of specific molecules (e.g., genes) and their information but does not specify anything about the data associated to them. High-throughput refers to a type of datum and how it has been obtained in order to quantify a cell's molecule. For example. a transcriptomics dataset has the expression level of genes measured with bulk mRNA-sequencing high-throughput technology. Microarrays were the standard tool for the quantification of molecules until the spread of sequencing techniques. With microarray, one had to design complementary bases, called "oligos" or "probes", to the target biological material. If the target is complementary to the oligos, a light signal is produced due to their binding and the intensity of the signal is proportional to the amount of the material pairing with that oligo. The captured intensity becomes the value of abundance or expression related uniquely to the target of interest (gene, protein, metabolite, etc …). For this to be able to work, there is the need to know the target sequence for designing the corresponding probe.

This technology has been replaced with the sequencing. DNA sequencing is the process of directly determining the nucleic acid sequence. It can be metaphorically simplified with the process of reading directly a book instead of finding those parts that match with the sentences of interest. Next generation sequencing capable of generating massive amounts of genomic, transcriptomic, proteomic and metabolomic data, provided new opportunities to understand human diseases, identify potential biomarkers, and develop new treatments. This data intensive paradigm has fundamentally transformed biomedical science and enhanced the reductionism approach.

Sequencing allows to study each individual molecule, to study their composition and aberrations at the resolution of single nucleotide or amino acid and to inquire their structural and chemical properties. However, cell's molecules do not work in isolation: they communicate and coordinate with each other to carry out biological functions. This interconnected nature of the molecules propagates aberrations or erroneous signals throughout the cell, thus posing a great challenge to elucidate the true causes and underlying mechanisms of a disease phenotype. An altered molecule or pathway may be the consequence of a distant actor that provoked an abnormal and disease related signaling cascade.

Biological networks provide a conceptual and intuitive framework to represent investigate, characterize, and understand how molecules interact and the effect of their interactions. The idea of representing molecules as the persons of a social network connected if they perform a task together or combine to fulfil a cell function boosted the interest towards the research and annotation of the interactome (the set of all direct and indirect molecular interactions). By employing a holistic approach, network biology studies the "interactome" and currently provides experimental validated interactions to create biological networks representing the molecular "wiring" diagram of a cell's information processing system.

A biological network and its graphical representation allow to reveal the relationships among different cellular components and help to detect subtle patterns by "connecting the dots".

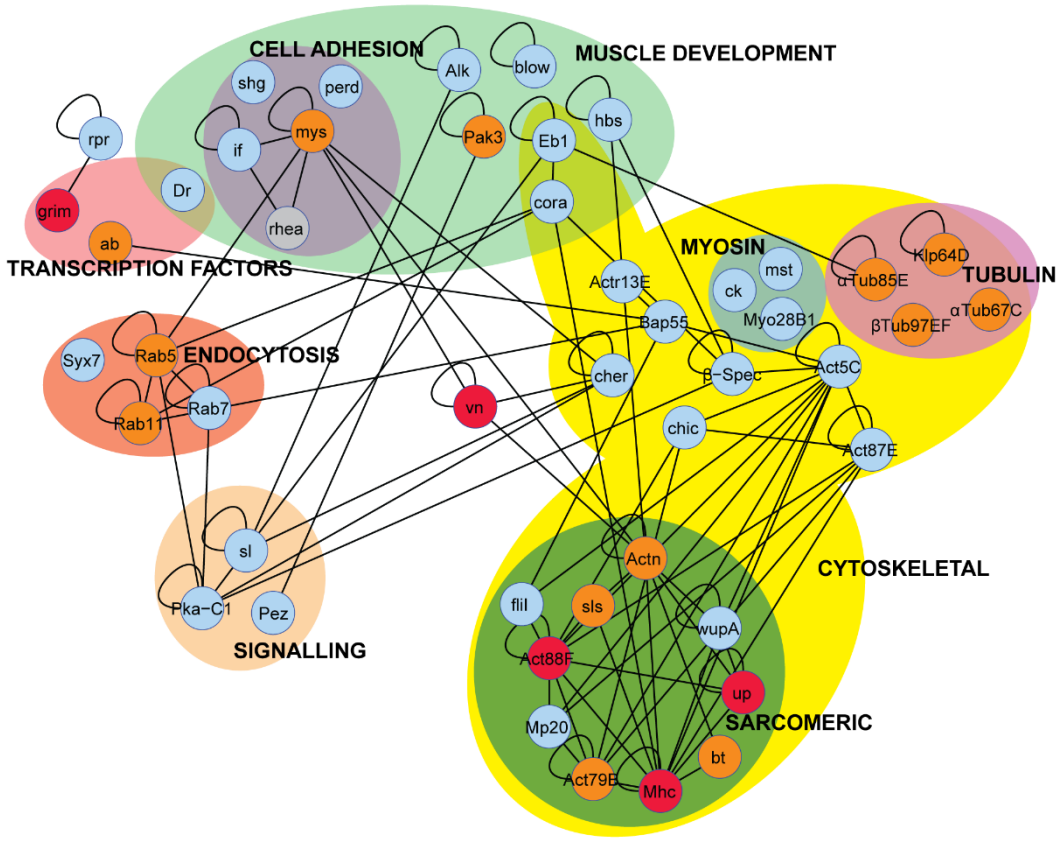


Fig.2.5 [7] Biological network of interacting genes. Lines (aka edges or connections) linking genes indicate direct genetic or protein-based interactions. Reflexive loops indicate a direct genetic or protein-based interaction with itself. Genes were then organized into pathways. Each gene in the network was colour-coded based how much have been severed affect by the disease phenotype in study. The intuition is that, the genes most affected by the disease in the sarcomeric pathway generated an erroneous signaling cascade outside the pathway and reached the muscle development and cell adhesion which alter the normal movements of the cell.

2.4 Networks

Omics experiments characterize transcripts, proteins and metabolites that are involved in the cell's activities. At the same time, all molecules are connected by direct and indirect connections. For example, a gene (transcriptomics) without mutation (genomics) is promoted by a long non-coding RNA (transcriptomics) to codify the synthesis of a protein (proteomics) which is used by the cell to produce glucose (metabolite). Representing the molecules as nodes of a network (i.e. persons of a social network) and connecting two nodes based on if there is a relationship between them produces a biological network that helps to understand

better how the cell performs its activities or how an alteration is helping the progress of a disease.

For this reason, the construction and analysis of biological networks has become crucial for the study of the cell. The interactions for importance can be physical, functional or predicted [8,9]. A physical interaction if there is a direct contact between the molecules (the atoms of the two molecules interact with one another), a functional interaction when they perform the same task and depend on one another (the two molecules do the same task and alterations affect both of them), while a predicted interaction if it has been found only in-silico and it has not been proved in-vitro. Among the three of them, only the physical interaction changes meaning based on the type of two molecules involved.

Depending by the types of interactions and by the molecular entities represented, the main biological and molecular interaction networks can be considered the following [8,9]: metabolic network, protein-protein interaction network (PPI), genetic interaction network (GI), and gene regulatory network (GRN). For each type, there is at least a consortium which takes the responsibility to collect and manually validate the interactions that are in research publications. This allows to have always access to the most recent network and interactions between the biological entities of interest. For example, STRING [10] collects protein-protein physical and functional interactions.

Metabolism is the chemical process by which cells break down food and nutrients into usable building blocks and then reassemble those building block to form the biological molecules the cell needs to complete its other tasks. Typically, this breakdown and reassembly involves chains of successive chemical reactions that convert initial inputs into useful end products by a series of steps. The complete set of all reactions included in the metabolism forms the metabolic network. The vertices in a metabolic network are chemicals produced and consumed by the reactions. These chemicals are known as metabolites. These are small molecules like carbohydrates, lipids, as well as amino acids and nucleotides. The metabolites consumed are called the substrates of the reaction, while those produced are called the products. Most metabolic reactions do not occur spontaneously or do so only at

a very low rate. To make reactions occur at a usable rate, the cell employs an array of chemical catalysts, referred to as enzymes. Enzymes are not consumed in the reactions they catalyse. By increasing or decreasing the concentration of the enzyme that catalyses a particular reaction, the cell can turn that reaction on or off, or moderate its speed. Enzymes tend to be highly specific to the reaction they catalyse. The most correct representation of a metabolic network is as a bipartite graph. The two types of vertices represent metabolites and metabolic reactions, with edges joining each metabolite to the reaction in which it participates. The edges are directed, since some metabolites (the substrates) go into the reaction and some (the products) come out of it. Enzymes can be incorporated by adding a third class of vertex to represent them, with undirected edges connecting them to the reactions they catalyse. The resulting graph is a mixed (directed and undirected) tripartite network.

Proteins do interact with one another and with other biomolecules, both large and small, but the interactions are not purely chemical. Proteins are long-chain molecules formed by the concatenation of a series of basic units called amino acids. Once created, a protein does not stay in a loose chain-like form, but folds on itself in a folded form whose shape depends on the amino acid sequence. The folded form dictates the physical interaction it can have with other molecules. Hence, the primary mode of protein-protein interaction is physical rather than chemical, their complicated folded shapes interlocking to create so-called protein complexes but without the exchange of particles that defines chemical reactions. In a protein-protein interaction network, the nodes are proteins and two nodes are connected by an undirected edge if they physically interact. In other words, it represents how proteins cooperate to perform biological processes within the cell. The physical interaction between proteins happens in specific binding regions. It can be stable or transient. Stable interactions allow two or more proteins to create a complex such as ribosomes, while transient interactions modify temporarily a protein to perform a specific action as in the case of protein kinases.

A gene regulatory network (GRN) represents how gene expression is controlled in a cell. In a GRN there are two actors that physically interact: genes and their regulators. Typical regulators are transcription factors (TF), i.e. proteins that are

able to bind specific regions of DNA in order to turn on or off the transcription of a specific gene. TFs act activating or inhibiting the recruitment of the RNA polymerase that is a protein responsible for the transcription of the genes. Further, TFs are themselves produced by the transcription and translation of a gene, therefore GRNs can be very complex. Other important regulators that act in GRNs are miRNAs. Several studies showed that there is an interplay between transcriptional regulators such as TFs and post-transcriptional regulators such as miRNAs in order to decrease or potentiate a gene expression.

A genetic interaction (GI) network represents functional associations between genes.

2.5 Cellular Pathway

A pathway or biological process describes a series of chemical reactions in which a group of molecules in a cell work together to control a cell function, such as cell division or cell death [11,12]. A cell receives signals from its environment when a molecule, such as a hormone or growth factor, binds to a specific protein receptor on or in the cell. After the first molecule in the pathway receives a signal, it activates another molecule. This process is repeated through the entire signaling pathway until the last molecule is activated and the cell function is carried out. Abnormal activation of signaling pathways may lead to diseases, such as cancer. Drugs are being developed to target specific molecules involved in these pathways.

There are many types of biological pathways. Among the most well-known, there are pathways involved in metabolism, in the regulation of genes and in the transmission of signals. Important to highlight that, on the contrary of the biological networks which are often kept separated by types, a biological pathway often includes all the types.

Metabolic pathways make possible the chemical reactions that occur in our bodies. An example of a metabolic pathway is the process by which cells break down food into energy molecules that can be stored for later use.

Gene-regulation pathways turn genes on and off. Such action is vital because genes provide the recipe by which cells produce proteins, which are the key components

needed to carry out nearly every task in our bodies. Proteins make up our muscles and organs, help our bodies move and defend us against germs.

Signal transduction pathways move a signal from a cell's exterior to its interior. Different cells are able to receive specific signals through structures on their surface called receptors. After interacting with these receptors, the signal travels into the cell, where its message is transmitted by specialized proteins that trigger a specific reaction in the cell. For example, a chemical signal from outside the cell might direct the cell to produce a particular protein inside the cell. In turn, that protein may be a signal that prompts the cell to move.

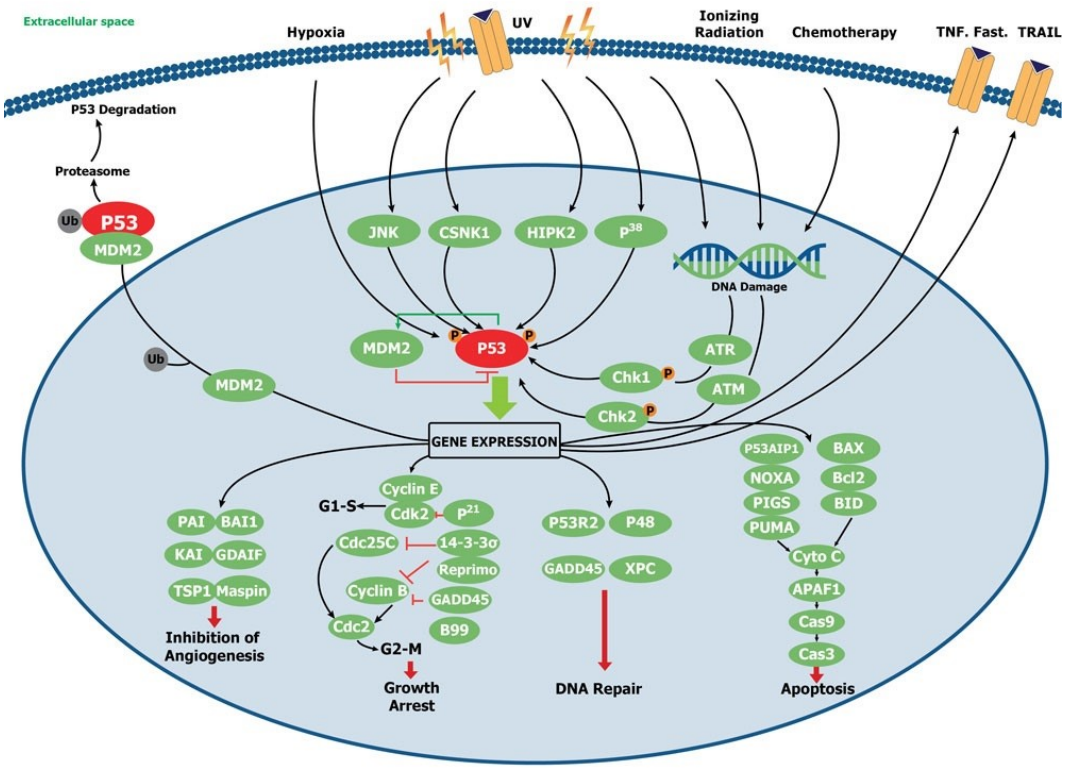


Fig. 2.6. [13] p53 is maintained at low protein levels during times of homeostasis, when the cell is not exposed to stress or DNA-damaging events, by its predominant negative regulator Mdm2 through the ubiquitin-proteasome pathway. The P53 pathway is important in cancer research. If the TP53 gene is damaged, tumor suppression is severely reduced. People who inherit only one functional copy of the TP53 gene will most likely develop tumors in early adulthood. The TP53 gene can also be damaged in cells by mutagens, increasing the likelihood that the cell will begin decontrolled division. More than 50 percent of human tumors contain a mutation or deletion of the TP53 gene. Restoring endogenous p53 function holds a lot of promise, while loss of p53

creates genomic instability that most often results in the aneuploidy phenotype. Certain pathogens can affect the p53 protein. One such example, human papillomavirus (HPV), encodes the E6 protein, which binds to the p53 protein and inactivates it, which in conjunction with the inactivation of a number of other regulating factors, leads to the clinical disease of warts. Certain HPV types, in particular types 16 and 18, can also lead to progression from a benign wart to low or high-grade cervical dysplasia, which are reversible forms of precancerous lesions. In healthy humans, the p53 protein is continually produced and degraded in the cell. The degradation of the p53 protein is associated with MDM2 binding. In a negative feedback loop, MDM2 is itself induced by the p53 protein, however, mutant p53 proteins often do not induce MDM2, and are thus able to accumulate at very high concentrations.

Researchers are learning that biological pathways are far more complicated than once thought. Most pathways do not start at point A and end at point B. In fact, many pathways have no real boundaries, and pathways often work together to accomplish tasks. When multiple biological pathways are considered as interacting with each other, they are modelled by a biological network.

3. GRAPH THEORY BACKGROUND

This chapter summarizes the concepts of graph theory that our classifiers exploit during their workflow. netDx, Pratic and Simpati use patient similarity networks (PSNs) as features to learn and to classify. These networks are mathematically defined as complete weighted graphs in which a patient is represented as a node and a connection between two patients is an edge associated to a value which measures how much the individuals are similar. All the three classifiers include a step of analysis of the properties of the PSNs and a recommender system which considers how much an unknown patient is similar to known ones. Further on, Pratic and Simpati use also biological networks to impute new information and standardize the starting patient data. Biological networks are mathematically defined as graph, but their properties depend by the represented molecules and by the meaning of the interactions. This chapter provides the basic definitions and knowledge to understand how features and meta information are manipulated and processed by the classifiers.

- Section 3.1 introduces the basic elements that compose a graph, how to mathematically define it, graph types and cases of special graphs
    - netDx includes a filtering step in which removes graph components to increase the signal-to-noise ratio, reduce information and decrease the system usage during the workflow.
    - Pratic and Simpati work on PSNs and biological networks in form of vectors or matrices. They both include a step of analysis of the components to understand if a PSNs can be considered a predictive feature based on its ability to characterize and separate the patient classes in comparison.
- Section 3.2 introduces graph properties and topology. A graph is described as a data structure in which all the elements are connected and cover a specific role. Understanding this data structure allows to formulate and perform query to get specific information about the elements.
    - Simpati uses the concepts about graph topology to predict the class of an unknown patient.

    - Simpati provides results in form of PSNs, so graphs that can be further topologically analysed to get extra information about the input patients.
- Section 3.3 introduces graph centrality measures. Tools of analysis to determine how much an element represented in a graph is crucial and important. Each measure ranks the elements based on a specific criterium of importance.
    - Simpati represents both molecules and patients as elements of networks. Centrality measures allow to prioritize the investigation of specific elements as well as to infer information about them. For example, how much a protein is considered multi-functional inside a protein-protein interaction network can be measured with the degree centrality. Further on, given a predictive patient similarity network is possible to use centrality measures to find the patient which is the most similar to the other members of its class; it is like to get a representative patient of the disease phenotype in study. These operations are not natively supported by enrichment tools for the pathway analysis and are advantages of the classifiers based on graphs.
- Section 3.4 introduces the graph models. Definitions used to categorize and identify graphs that exist in the real world (e.g., road networks) and that are produced in-silico (protein-protein interaction networks model partially and in a bias manner the biological interactions between proteins).
    - Biological networks follow a specific graph model in which the number of connections that a molecule possesses and the distance between two biological entities has a specific meaning. For example, it is commonly known the “four degrees of separation” theory in a social network. This one defines that any pair of persons are separated by at most 4 other individuals. Biological networks fall into a different model. For example, in a protein-protein interaction network, proteins that are within 3 interactions are considered functionally similar [14].

- Section 3.5 introduces the concept of cohesive subgroup. Theory and formalisms that are at the base of the detection of communities and clusters in graphs. From science perspective, there is no one-size-fit-all definition for a cohesive subgroup. It depends on the problem, application, and graph. However, there are notions at the basis of every algorithm that aims to identify a group of nodes based on how much they are close and cohesive.
    - Simpati includes a novel cohesive subgroup detection algorithm for identifying the set of patients that are the most cohesive in a class. The patients that are left out the selection are considered outliers. Simpati includes such operation in order to not assume that a class is well defined and that all the patients sharing a user-defined label have exactly the same alterations at the pathway level.
- Section 3.6 introduces the clustering analysis. Task of finding cohesive subgroups in a graph. Graph-theory provides multiple algorithms for the detection of clusters. They vary in applications, strengths, and weaknesses.
    - Cohesive subgroup detection algorithms evolved into clustering or community algorithms. In biology, they have been mainly developed for finding cohesive groups of molecules in biological and molecular networks. However, they never been applied to weighted complete networks as patient similarity networks; for this reason, we implemented a specific algorithm in Simpati.
- Section 3.7 introduces the graph layouts. One of the main advantage of graphs is that they can be represented graphically. Observing a graph allows to get information about the nodes as the elements of interest. However, if a graph includes many edges or nodes, it can be difficult to understand how the elements are connected without using a proper mapping of its structure in the 2D space. Therefore, a branch of graph theory is dedicated to find solutions about how to represent graphs.
    - Simpati has implemented a function to represent a patient similarity network. The latter has every node (aka patient) connected to any other node, so it cannot be easily represented. It requires a method which organizes the nodes and their connections in order to make

everything interpretable and understandable. This section describes a methodology that we integrated in our software.

- Section 3.8 introduces the network-based propagation. Graph-based technique to standardize, to infer new information and remove noise from the data associated to the node's weight of a graph.
    - Simpati uses a network-based propagation algorithm to integrate a biological omic with its corresponding molecular interaction network. It uses an algorithm called random walk with restart to standardize the values associated to the molecules by the omic and to impute a new value for those molecules that are not described from the a priori data.

3.1 FUNDAMENTALS

A graph G is a pair $G = (V, E)$ where V is a set of vertices and E is a (multi)set of unordered pairs of vertices [15–17]. The elements of E are called edges. We write $V(G)$ for the set of vertices and $E(G)$ for the set of edges of a graph G. Also, $|G| = |V(G)|$ denotes the number of vertices and $e(G) = |E(G)|$ denotes the number of edges. A loop is an edge (v, v) for some $v \in V$. An edge $e = (u, v)$ is a multiple edge if it appears multiple times in E. A graph is simple if it has no loops or multiple edges.

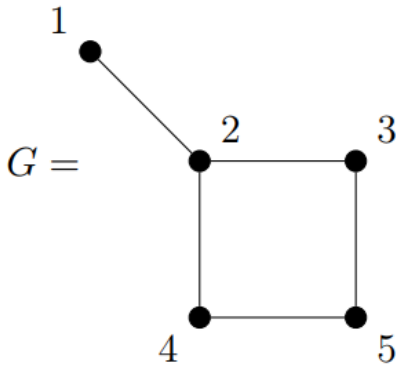
Definitions:

- Vertices u, v are adjacent in G if $(u, v) \in E(G)$.
- An edge $e \in E(G)$ is incident to a vertex $v \in V(G)$ if $v \in e$.
- Edges e, e' are incident if $e \cap e' \neq \emptyset$.
- If $(u, v) \in E$ then v is a neighbour of u.

The usual way to picture a graph is to put a dot for each vertex and to join adjacent vertices with lines. The specific drawing is irrelevant, all that matters is which pairs are adjacent.

Given $[n] = \{1, \dots, n\}$, let $G = (V, E)$ be a graph with $V = [n]$. The adjacency matrix $A = A(G)$ is the $n \times n$ symmetric matrix defined by

$$a_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$



$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}.$$

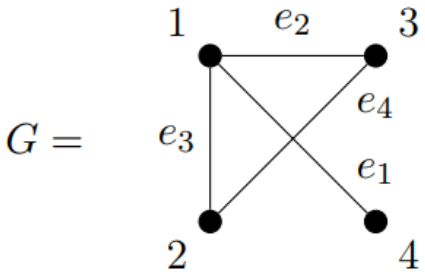
Fig. 3.1. Example of graph representation and its corresponding adjacency matrix A

The affinity matrix, also called similarity matrix, is like the adjacency matrix except that the value for a pair of nodes expresses how similar are the objects represented by them. If pairs of objects are very dissimilar then the affinity is 0. If the objects are identical, then the affinity is 1.

$$a_{ij} = \begin{cases} \text{similarity(i,j)} & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

Given $[n] = \{1, \dots, n\}$, let $G = (V, E)$ be a graph with $V = \{v_1, \dots, v_n\}$ and $E = \{e_1, \dots, e_m\}$. Then the incidence matrix $B = B(G)$ of G is the $n \times m$ matrix defined by

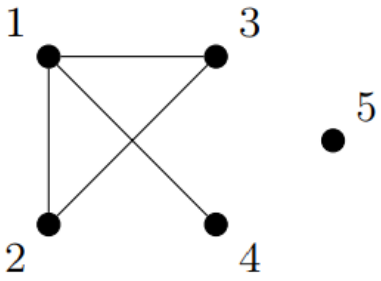
$$b_{ij} = \begin{cases} 1 & \text{if } v_i \in e_j \\ 0 & \text{otherwise} \end{cases}$$



$$B = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Fig. 3.2. Example of graph representation and its corresponding incidence matrix B

Given $G = (V, E)$ and a vertex $v \in V$, we define the neighbourhood $N(v)$ of v to be the set of neighbours of v. Let the degree $d(v)$ of v be $|N(v)|$, the number of neighbours of v. A vertex v is isolated if $d(v) = 0$



$d(1) = 3,\ d(2) = 2,\ d(3) = 2,\ d(4) = 1,\ d(5) = 0;$

5 is isolated.

Fig. 3.3. Example of graph representation and of annotation of the degree of its vertices

For any graph G on the vertex set [n] with adjacency and incidence matrices A and B, we have $BB^t = D + A$, where:

$$D = \begin{pmatrix} d(1) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & d(n) \end{pmatrix}$$

The minimum degree of a graph G is denoted $\delta(G)$; the maximum degree is denoted $\Delta(G)$. The average degree is:

$$\bar{d}(G) = \frac{\sum_{v \in G} d(v)}{|V(G)|}$$

A graph $H = (U, F)$ is a subgraph of a graph $G = (V, E)$ if $U \subseteq V$ and $F \subseteq E$. If $U = V$ then H is called panning. Given $G = (V, E)$ and $U \subseteq V (U \neq \emptyset)$, let $G[U]$ denote the graph with vertex set U and edge set $E(G[U]) = \{e \in E(G) : e \subseteq U\}$. Then $G[U]$ is called the subgraph of G induced by U.

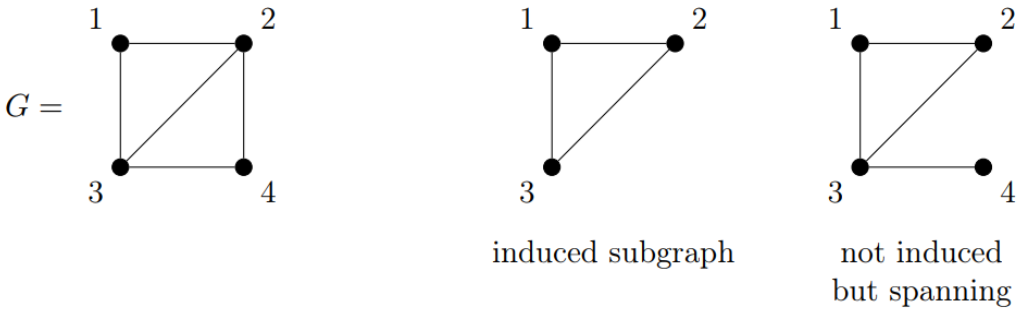


Fig. 3.4. Example of graphs induced and not induced

Special graphs:

- K_n is the complete graph, or a clique. Take n vertices and all possible edges connecting them.
- An empty graph has no edges.
- $G = (V, E)$ is bipartite if there is a partition $V = V_1 \cup V_2$ into two disjoint sets such that each $e \in E(G)$ intersects both V_1 and V_2.
- $K_{n,m}$ is the complete bipartite graph. Take $n + m$ vertices partitioned into a set A of size n and a set B of size m, and include every possible edge between A and B.

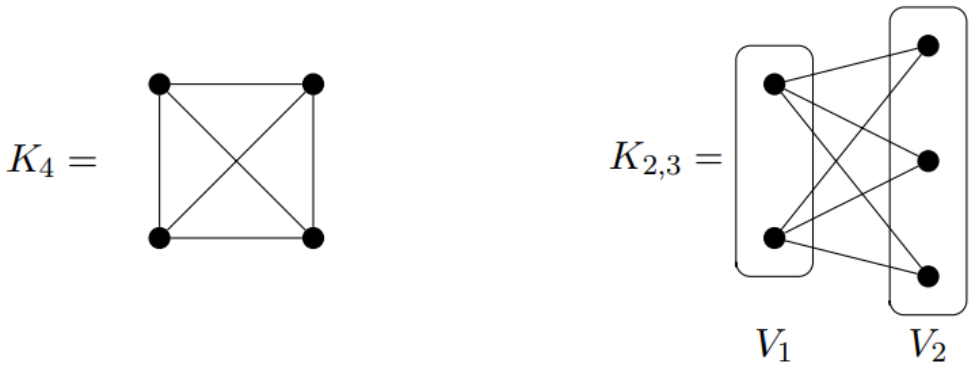


Fig. 3.5. Example of special graphs

A walk in G is a sequence of vertices $v_0, v_1, v_2, \ldots, v_k$, and a sequence of edges $(v_i, v_{i+1}) \in E(G)$. A walk is a path if all v_i are distinct. If for such a path with $k \geq 2$, (v_0, v_k) is also an edge in G, then $v_0, v_1, \ldots, v_k, v_0$ is a cycle. For multigraphs, we also consider loops and pairs of multiple edges to be cycles. The length of a path, cycle or walk is the number of edges in it.

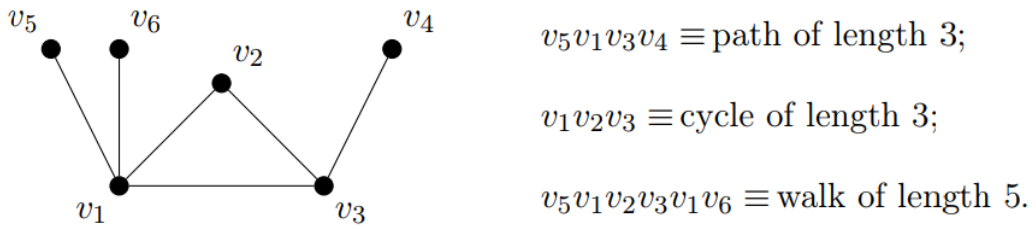


Fig. 3.6. Example of paths annotation on a graph

A graph G is connected if for all pairs $u, v \in G$, there is a path in G from u to v. A (connected) component of G is a connected subgraph that is maximal by inclusion. We say G is connected if and only if it has one connected component.

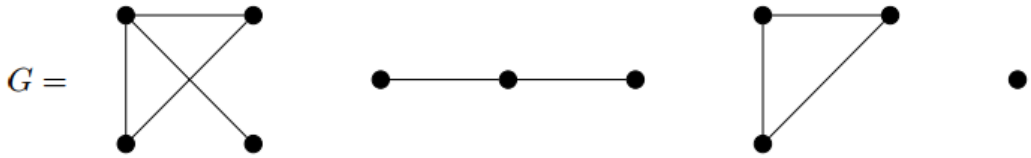


Fig. 3.7. Example of graph composed by four connected components

A graph is undirected if there is a single connection defined as $E = \{(i,j) |, i,j \in V\}$ between vertices i and j. In such case, vertices i and j are called direct neighbours.

A graph is called directed if an edge between vertices i and j is represented by an arrow, thus indicating a direction from vertex i to vertex j or vice versa. A directed graph is defined as an ordered triple $G = (V, E, f)$ where f is a function that maps each element in set E to an ordered pair of vertices in V.

A weighted graph is defined as a graph where E is a set of edges between the vertices i and j ($E = \{(i,j) \mid i,j \in V\}$) associated with a weight function $w: E \rightarrow R$, where R denotes the set of all real numbers. Most of the times, the weight w_{ij} of the edge between nodes i and j represents the relevance of the connection (e.g., sequence similarity network).

A subgraph is said to be maximal with respect to some property if that property holds for the subgraph but does not hold if additional nodes and the edges incident to them are added to the subgraph. For example, a component of a graph is a maximal connected subgraph. The presence of two or more components in a graph indicates that the graph is disconnected.

A clique in a graph is a maximal complete subgraph of three or more nodes. It consists of a subset of nodes, all of which are adjacent to each other, and there are no other nodes that are also adjacent to all of the members of the clique.

In a biological network, nodes are the biological entities while edges are the physical or functional interactions in which they are involved. In this context, the terms graph and network can be used interchangeably.

3.2 Properties

As degree $\deg_I$, we define the total number of edges adjacent to a vertex. In the case of a directed graph, we distinguish between the "indegree" $\left(\deg_i^{in}\right)$ and "outdegree" $(\deg_i^{out})$ [15–17]. The indegree refers to the number of edges, incident from the vertex, whereas the outdegree to the number of edges incident to the vertex. In a social network for example, the indegree would represent the followers, whereas the outdegree the people one follows. The total degree in a directed graph is the sum of the indegree and outdegree $\deg_i = \deg_i^{in} + \deg_i^{\text{out}}$ showing all connections (both followers and followed people). The average degree of the network is $\deg_{avg} = \frac{\Sigma deg_i}{|V|}$. Looking at all nodes in a network, to study the degree distribution $p(k)$, we consider the probability that a randomly selected vertex has degree equal to k. The same information can also be found as cumulative degree distribution $p_c(k)$ which shows the a-posterior probability of a randomly selected vertex to have degree larger than k. Notably, the degree distribution is one of the most important topological features and is characteristic to different network types. In the simplest case, $p(k)$ can be estimated by a histogram of degrees. Networks, whose degree distribution follow a power law, are called scale-free networks.

Density is the ratio between the number of edges in a graph and the number of possible edges in the same graph. In a fully connected graph (e.g., protein complex), the number of possible edges (pairwise connections) are $\frac{|V|(|V|-1)}{2}$. Therefore, the density can be calculated as $\frac{2|E|}{|V|(|V|-1)}$. If a graph has $|E| \simeq |V|^k, 2 > k > 1$, then this graph is considered as dense, whereas when a graph has $|E| \simeq |V|$ or $|E| \simeq |V|^k, k \leq 1$, it is considered as sparse.

The Clustering coefficient is a measure which shows whether a network or a node has the tendency to form clusters or tightly connected communities. The clustering coefficient of a node is defined as the number of edges between its neighbours divided by the number of possible connections between these neighbours. The clustering coefficient of a node i is defined as $Cl_i = \frac{2e}{k(k-1)}$ where k is the number of neighbors (degree) and e the number of edges between these k neighbors. The

average clustering of a network is given by $Cl_{\text{avg}} = \frac{\Sigma C_i}{|V|}$. The clustering coefficient takes values $0 \leq Cl_i \leq 1$, thus the closer to 1 , the higher the tendency for clusters to be formed.

3.3 Centrality Measures

In a biological network, nodes are the biological entities while edges are the interactions in which they are involved. In this context, the terms graph and network can be used interchangeably. Examples of undirected biological networks are protein-protein interaction networks and genetic interaction networks, while cell signalling networks, metabolic networks and gene regulatory networks are directed. In general, directed graphs are effective to model biological networks in which it is important to underline the sequential order of the interactions to obtain a specific product of reaction, as in the case of a signalling pathway.

This section shows how nodes can be ranked or sorted according to their properties [15–17]. In biological networks, a ranking based on a specific property can help in prioritizing the molecules in study. For example, a transcriptomic dataset can easily include twenty thousand genes and the selection of the most important ones can pass through the question: which genes have the highest number of interactions such that they can be recognized as crucial for the function of the other patient's genes? This question can be solved by ranking the genes based on their degree.

Degree Centrality shows that an important node is involved in many interactions. For a node i, the degree centrality is calculated as $C_a(i) = \deg(i)$. For directed graphs, each node is obviously characterized by two degree centralities. Nodes with very high degree centrality are called hubs since they are connected to many neighbours.

Closeness Centrality indicates important nodes that can communicate quickly with others of the network. Let $G = (V, E)$ be an undirected graph. Then, the centrality is defined as $C_{cb}(i) = \frac{1}{\sum_{I \in V}^{|V|} \text{dist}(i,j)}$ where dist(i, j) denotes the distance or else the shortest path p between the nodes i and j.

Betweenness Centrality shows that nodes which are intermediate between neighbours rank higher. Without these nodes, there would be no way for two neighbours to communicate with each other. Thus, betweenness centrality shows important nodes that lie on a high proportion of paths between other nodes in the network. For distinct nodes $i, j, w \in V(G)$, let σ_{ij} be the total number of shortest paths between i and j and $\sigma_{ij}(w)$ be the number of shortest paths from i to j that pass through w. Moreover, for $w \in V$ (G), let $V(i)$ denote the set of all ordered pairs, (i, j) in $V(G) \times V(G)$ such that i, j, w are all distinct. Then, the Betweenness Centrality is calculated as $C_b(w) = \sum_{(i,j) \in V(w)} \frac{\sigma_{ij}(w)}{\sigma_{ij}}$.

Eigenvector Centrality ranks higher the nodes that are connected to important neighbours. Let $G = (V, E)$ be an undirected graph and A the adjacency matrix of network G. The eigenvector centrality is the eigenvector $C_{\text{eiv-}}$ of the largest eigenvalue $\lambda_{\max}$ in absolute value such that $\lambda C_{eiv} = AC_{\text{eiv}}$. Formally, if A is the adjacency matrix of a network G with $V(G) = \{v_1, \ldots, v_n\}$, and $\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|$, then the eigenvector centrality $C_{eiv}(v_i)$ of the node v_I is given by the i^{th} coordinate x_i of a normalized eigenvector that satisfies the condition $Ax = \rho(A)x$.

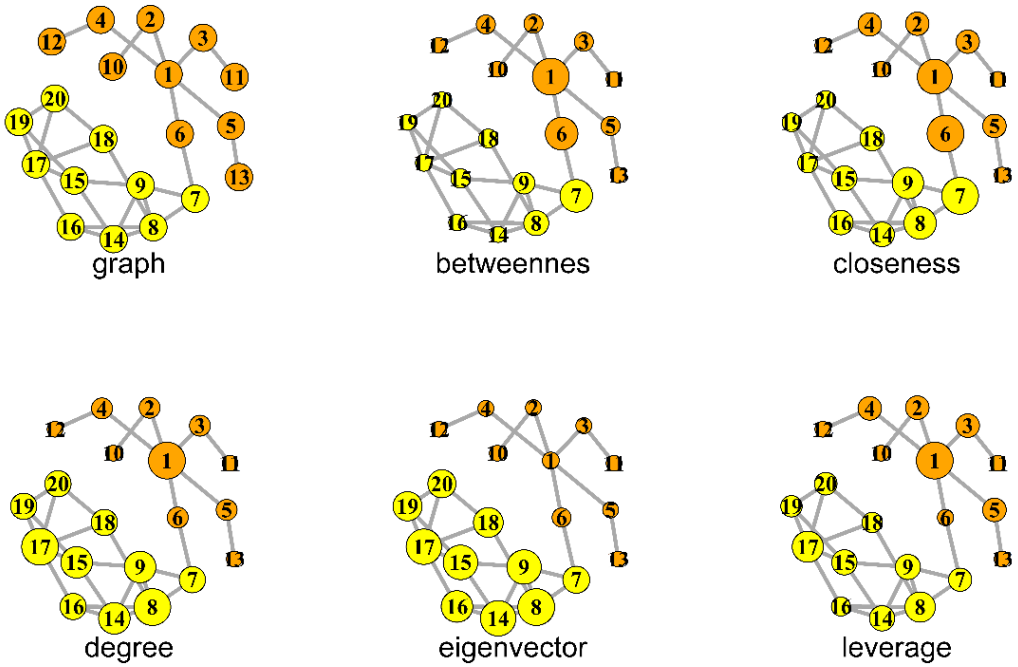


Fig. 3.7. Example of graph with two communities (yellow and orange) and 17 nodes where each

node is ranked by a centrality measure. The size of a node represents how much is central due to a specific measure. For the betweenness, only the nodes that are crucial for connecting two communities are important; node 1, 6 and 7 can be seen as composing a bridge and if are removed from the graph, then the two communities separate themselves.

3.4 Topological Models

In order to better understand a network's topology and come to the conclusion of whether observed features are network-specific or not, several models such as the Erdos-Rényi, Watts-Strogatz, and Barabási-Albert have been introduced [15–17].

The Erdos-Rényi model: It is one of the most popular models in graph theory and was mainly introduced to describe the properties of a random graph. According to this model, V, number of vertices are randomly connected with probability $p = \frac{2|E|}{|V|(|V|-1)}$. In general, in such a graph, each pair of vertices can be connected with approximately an equal probability $p \leq 1$, whereas the degree distribution is given by a binomial distribution. The probability of a vertex to have degree deg is $p(deg) = e^{-deg_{avg}} * \frac{deg_{avg}^{deg}}{deg!}$. Notably, for a network where $|V| \to \infty$ the distribution becomes approximately Poissonian. A typical characteristic of a random network is its homogeneity as most vertices have a similar number of connections. For small p, the network seems as disconnected, whereas for $p \approx \frac{1}{|V|}$, the network has a bigger component containing most of the network's connections. When $p \geq \frac{\log(|V|)}{|V|}$, then almost all vertices are connected homogeneously and at random. The clustering coefficient of this network is $C = p = \frac{deg_{avg}}{|V|}$ and shows that the probability of two nodes with a common neighbor to be connected is the same as the probability of two randomly paired vertices. In the case of biological networks, straightforward comparisons show if they have a certain topology or differ from any other random network. Thus, Erdos-Rényi is not a good model for biological networks with respect to degree distribution.

The Watts-Strogatz model: This model was introduced to describe random networks that follow a small world topology meaning that most nodes can be reached by any other node in a small number of steps. While random networks can

often capture this property too, they fail to account for highly connected regions like in most empirical networks (e.g., social networks). Therefore, Watts and Strogatz proposed a model for networks described by local structures (high clustering coefficient) as well as small average path lengths. Metabolic networks, in which metabolites are linked to each other with small steps, is a typical example. In a Watts-Strogatz network, if all vertices are placed on a circular ring, each vertex would be connected to its $\frac{|V|}{2}$ neighbours. In the real world, this indicates the form of small communities where people know other people from their close environment as well as friends of friends from nearby areas. Coexistence of high local clustering and short average path length are two main characteristics of this type of networks.

The Barabási-Albert model: This model describes random scale-free networks. These are networks whose degree distribution follows a power law considering their inhomogeneous degree distribution or otherwise networks with nodes which do not have a typical number of neighbours. According to this model, networks can evolve overtime and new edges do not appear randomly, whereas new nodes follow the existing degree distribution. At time point $t = 0$ for example, let's assume a network consisting of V_0 vertices and zero edges. A new vertex will connect with $e \leq V_0$ edges to the existing vertices, whereas after t time points, the network is expected to consist of $V = V_0 + et$ edges.

Notably, for $t \gg 1$, the Barabasi-Albert model will exhibit a scale-free distribution $p(k) \sim k^{-\gamma}, \gamma = 3$. Like in a social network, individuals who already have many friends are likely to acquire more friends overtime compared to individuals with a limited number of friends. When comparing the Erdos-Rényi and Watts-Strogatz networks of the same size and density, the Barabási-Albert networks were found to have shorter average path lengths.

Protein-Protein Interaction Networks follow a small-world property and are scale-free networks. Central hubs often represent evolutionarily conserved proteins, whereas cliques (fully connected subgraphs) have been found to have a high functional significance.

Gene and Genetic Regulatory Networks are directed, dynamic, and can be visualized as bipartite graphs. In such networks, most nodes have only a few interactions and only a few hubs come with a higher connectivity degree. In any case, such networks follow a power law degree distribution (scale-free). These networks are usually directed graphs and can be represented as Petri nets. They are scale-free, they carry small-world properties and can often be organized using hierarchies.

3.5 Cohesive Subgroups

One of the major concerns of social network analysis is identification of cohesive subgroups of actors within a network. Cohesive subgroups are subsets of actors among whom there are relatively strong, direct, intense, frequent, or positive ties (aka edges, lines, connections, relations). These methods attempt, in part, to formalize the intuitive and theoretical notion of social group using social network properties. However, since the concept of social group as used by social and behavioural scientists is quite general, and there are many specific properties of a social network that are related to the cohesiveness of subgroups, there are many possible social network subgroup definitions.

Social psychologists and sociologists have argued that individuals are most strongly influenced by the members of their primary groups - people with whom they engage in frequent interactions [18] - and anthropologists have argued that primary groups are integral to understanding people within the contexts of their communities [18]. Homans and other more recent organizational theorists have argued that large organizations are composed of essentially non-overlapping subgroups which contain dense interactions [19]. This conception is consistent with theories defined at the level of the individual, in which people influence each other through direct communication within their subgroups, and then integrate into the larger organization through interactions beyond the subgroup boundary [18]. An extensive literature based on laboratory-controlled interactions demonstrated that patterns of interaction are linked to actors' knowledge bases and resulting actions (behaviours), and that, in particular, members of cohesive subgroups are likely to share

sentiments (including beliefs) and exhibit similar actions related to the content of their interactions [18].

Although the literature on cohesive subgroups in networks contains numerous ways to conceptualize the idea of subgroups, the notion of subgroup is formalized by the general property of cohesion based on specified properties of the ties among the members. There are four general properties of cohesive subgroups that have influenced social network formalizations of this concept [20]:

- The mutuality of ties
- The closeness or reachability of subgroup members
- The frequency of ties among members
- The relative frequency of ties among subgroup members compared to non-members

Subgroups based on mutuality of ties require that all pairs of subgroup members "choose" each other (or are adjacent); subgroups based on reachability require that all subgroup members be reachable to each other, but not necessarily adjacent; subgroups based on numerous ties require that subgroup members have ties to many others within the subgroup; and subgroups based on the relative density or frequency of ties require that subgroups be relatively cohesive when compared to the remainder of the network.

Subgroups Based on Complete Mutuality

- Cliques
    - strict definition that limits the number of groups which can be examined. In fact, the absence of a single tie will prevent a subgraph from being a clique.
    - the sizes of the cliques are limited by the degree of the nodes. This can be a problem if the number of ties that an actor can have been limited by the data collection design If collection limits data to k ties, then no clique in the study can be larger than k+1 members.
    - All clique members are adjacent to all other clique members, thus there are no distinctions among members based on graph theoretic

properties within the clique. If we expect that the cohesive subgroups within a network should exhibit interesting internal structure, such as having some core actors who are more strongly identified with the subgroup and other peripheral actors who are less identified with it, then a clique might be an inappropriate definition of cohesive subgroup.

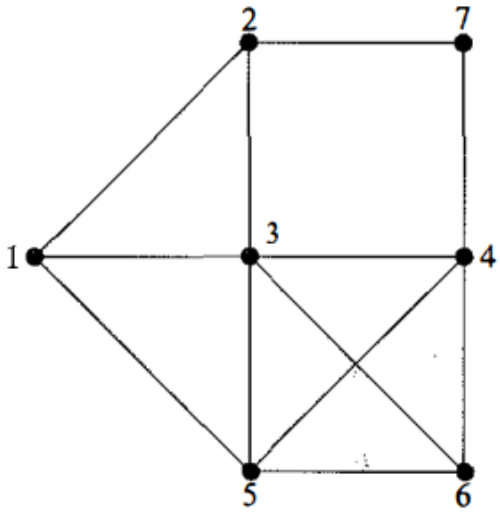
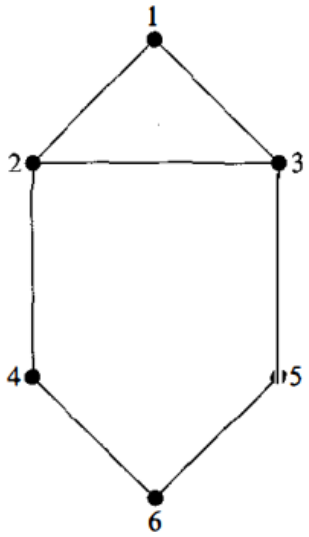


Figure 3.8. Example of graph and its cliques [20]

Subgroups Based on Reachability and Diameter: definition if the researcher hypothesizes that important social processes occur through intermediaries. Let's define the geodesic distance between two nodes, denoted by $d(i,j)$, as the length of a shortest path between them. Cohesive subgroups based on reachability require that the geodesic distances among members of a subgroup be small. Thus, we can specify some cutoff value, n, as the maximum length of geodesies connecting pairs of actors within the cohesive subgroup.

- n-clique:
    - is a maximal subgraph in which the largest geodesic distance between any two nodes is no greater than n. Maximal means: no

additional nodes that are also distance n or less from all nodes in the subgraph
    - All members might be connected through a non-member
    - n-cliques are not robust and subject to failure with the removal of an individual node [21]
- n-clans and n-clubs
    - n-clans exclude nodes connected by a geodesic distance greater than n and those connected through an intermediary
        - All n-clans are n-cliques [20]
    - n-clubs are maximal subgraphs with a diameter of n
        - All n-clubs are contained within n-cliques [22]



2-cliques: {1,2,3,4,5} and {2,3,4,5,6}
2-clan: {2,3,4,5,6}
2-clubs: {1,2,3,4}, {1,2,3,5}, and {2,3,4,5,6}

Figure 3.9. Graph illustrating n-cliques, n-clans, and n-clubs [20]

Subgroups based on nodal degree. The approaches that follow this definition are based on restrictions on the minimum number of actors adjacent to each actor in a subgroup. Since the number of actors adjacent to a given actor is quantified by the degree of the node in a graph, these subgroup methods focus on nodal degree. Subgroups based on nodal degree require actors to be adjacent to relatively numerous other subgroup members. Thus, unlike the clique definition that requires

all members of a cohesive subgroup to be adjacent to all other subgroup members, these alternatives require that all subgroup members be adjacent to some minimum number of other subgroup members. Subgroups based on adjacency between members are useful for understanding processes that operate primarily through direct contacts among subgroup members.

- k-plexes
    - A maximal subgraph that limits membership based on the number of ties k. In other words, each node in the subgraph may be lacking ties to no more than k subgraph members
- k-cores
    - A subgraph in which each node is adjacent to at least a minimum number, k, of the other nodes in the subgraph
    - In contrast to the k-plex, which specifies the acceptable number of lines that can be absent from each node, the k-core specifies the required number of lines that must be present from each node to others within the subgraph

Subgroups based on comparing Within to Outside the subgroup ties: this intuitive notion of cohesive subgroup considers both from the relative strength, frequency, density, or closeness of ties within the subgroup, and the relative weakness, infrequency, sparseness, or distance of ties from subgroup members to non-members [20] The ideal graph would show ties within the subgroup but no ties to actors outside the subgroup

- LS Sets
    - Proposition 1. Let G be a graph. Then $H \subset V(G)$ is a LS set if and only if for any proper $K \subset H$, fewer edges join K to $V(G) - H$ than join K to $H - K$.

Measures of Subgroup Cohesion

Descriptive measures of ties within a cohesive subgroup

- Bock and Husain (1950) [23] used the ratio of the strength of the ties within the subgroup to those outside the subgroup

- Numerator is average strength of ties within subgroup
- Denominator is average strength of ties outside subgroup
- With non-weighted ties the strength is based on the density of the subgroup
- If the ratio is equal to 1, then the strength of ties does not differ within the subgroup as compared to outside the subgroup. If the ratio is greater than 1, then the ties within the subgroup are more prevalent (or stronger) on average than are the ties outside the subgroup.

Variant on the subgroups definitions based on valued undirected ties (aka Weighted edges): In cohesive subgroups, members should have high valued ties.

Let's define the user-defined constant continuous value C and x_{ij} the value of the tie from i to j

- Values can range from 0 to C-1
- Highest possible value indicates the strongest tie between any pair of actors, smaller values of x_{ij} indicates weaker ties
- A threshold value of c can be used to examine cohesiveness

Cliques at level c

- Subgraph in which the ties between all have values of c or greater and there is no other actor outside the clique who also has ties of strength c or greater to all actors in the clique
- It is possible to adjust the threshold value of c from 0 to C-1
  - Increase to find more cohesion
  - Decrease to find less cohesion

n-cliques

- Based on the values of geodesics between members
- All paths between members have value of c or greater

k-plex

- Requires all members have ties with values of c or greater to no fewer than k of the other members

3.6 Clustering Analysis

The detection of cohesive subgroups falls under a bigger task called clustering or cluster analysis. The latter is defined as the operation of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). The first one is included in the second because objects described by features can be represented as nodes of a graph such that the edges between them are weighted based on how much the objects are similar. Cluster analysis [15–17] is a main task of exploratory data analysis, and a common technique for statistical data analysis, used in many fields, including pattern recognition, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning.

Similarly to cohesive subgroup also the notion of a cluster cannot be precisely defined. The definition varies based on the type of clustering algorithm that is applied:

Hierarchical clustering: based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. Connectivity-based clustering depends by the definition of distance between objects and the linkage criterion between candidate clusters (since a cluster consists of multiple objects, there are multiple candidates to compute the distance) to use: single-linkage clustering (the minimum of object distances), complete linkage clustering (the maximum of object distances), and UPGMA or WPGMA ("Unweighted or Weighted Pair Group Method with Arithmetic Mean", also known as average

linkage clustering). These methods will not produce a unique partitioning of the objects, but a hierarchy from which the user still needs to choose appropriate clusters. They are not very robust towards outliers, which will either show up as additional clusters or even cause other clusters to merge.

Centroid-based clustering: clusters are represented by a central vector, which may not necessarily be a member of the set of objects. When the number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized. Most k-means-type algorithms require the number of clusters k to be specified in advance, which is considered to be one of the biggest drawbacks of these algorithms.

Distribution-based clustering: Clusters are defined as objects belonging most likely to the same distribution. A convenient property of this approach is that this closely resembles the way artificial data sets are generated: by sampling random objects from a distribution. While the theoretical foundation of these methods is excellent, they suffer from one key problem known as overfitting, unless constraints are put on the model complexity. A more complex model will usually be able to explain the data better, which makes choosing the appropriate model complexity inherently difficult.

Density-based clustering: clusters as connected dense regions. Objects in sparse areas that are required to separate clusters are usually considered to be noise and border points.

Graph-based models: a clique, that is, a subset of nodes in a graph such that every two nodes in the subset are connected by an edge can be considered as a prototypical form of cluster. Relaxations of the complete connectivity requirement (a fraction of the edges can be missing) are known as quasi-cliques, as in the HCS clustering algorithm.

A "clustering" is essentially a set of such clusters, usually containing all objects in the data set. Additionally, it may specify the relationship of the clusters to each

other, for example, a hierarchy of clusters embedded in each other. Clustering models can be distinguished in:

- Hard clustering: each object belongs to a cluster or not
- Soft clustering (also: fuzzy clustering): each object belongs to each cluster to a certain degree (for example, a likelihood of belonging to the cluster)
- Strict partitioning clustering: each object belongs to exactly one cluster
- Strict partitioning clustering with outliers: objects can also belong to no cluster, and are considered outliers
- Overlapping clustering (also: alternative clustering, multi-view clustering): objects may belong to more than one cluster; usually involving hard clusters
- Hierarchical clustering: objects that belong to a child cluster also belong to the parent cluster
- Subspace clustering: while an overlapping clustering, within a uniquely defined subspace, clusters are not expected to overlap

3.7 Graph Layouts

For graph analysis and interpretation, it is important to be able to depict a graph whose structure, symmetries, and other main features become clear in a visually and aesthetically appealing way. This is especially true for graphs of large size, where many nodes and edges can have multiple clusters and interconnected areas. Graph drawing combines methods from mathematics and computer science to derive two- and three- dimensional representations of graphs, employing a number of strategies. Among the most successful layouts are the force-based layout approaches [24], where the nodes of the graphs are metaphorically modelled as point particles with attractive (spring) forces acting between nodes connected by an edge and repelling (electrical) forces acting between all pairs of nodes. The optimal layout is determined by the positions of nodes/particles that minimize the total energy of the system. Typically, such a state is found by simulating the forces of the many-particle physical system and arriving at a minimum energy state iteratively. Orthogonal layout methods allow the edges of the graph to run horizontally or vertically, parallel to the coordinate axes of the layout, while tree layout algorithms use tree-like structures and are suitable for visualizing ontologies

or hierarchies. Finally, circular layout methods place the vertices of the graph on a circle, carefully choosing the ordering of the vertices around the circle to reduce crossings and place adjacent vertices close to each other

3.8 Network-Based Propagation

The task of discovery the molecular basis promoting a patient's disease passes through the bioinformatics analysis of biological omics. However, an omic provides also noisy and incomplete observations. This requires the integration of multiple data within a single framework. A promising approach to boost the signal-to-noise ratio and overcome these hurdles is the analysis of the data in the context of molecular networks. In fact, a disease is rarely a consequence of an abnormality in a single gene but reflects the perturbations of the complex intracellular and intercellular network [25]. A better understanding of the effects of cellular connections may lead to identification of disease genes and disease pathways, which, in turn, may offer better targets for drug development [26]. These advances may also lead to better and more accurate biomarkers to monitor the functional integrity of networks that are perturbed by disease as well as to better disease classification [27]. A molecular network model consists of nodes, which are representative of molecules like genes or proteins, and edges, which represents relationships between the corresponding molecules like protein-protein interaction.

Recently, a new group of methods accounting for the global structure of the network has emerged as the state-of-the-art to improve the signal and reduce the noise provided by omics data [28]. These methods share the paradigm of network-based propagation, which amplifies a biological signal based on the assumption that molecules underlying similar phenotypes tend to interact with one another.

Prior information associated with molecules (aka seed or input) is imposed on the respective nodes of the network and then is propagated through the edges to nearby nodes (both seed and non-seed) iteratively for a fixed number of steps or until convergence. In this way, nodes that were not included in the prior information can be associated with the phenotypes by the means of proximity to the prior nodes [25]. The prior information of the seed nodes is adjusted based on their neighbours.

A seed node with prior information will keep is information if its neighbours have information, while it will lose information otherwise.

For example, in a bioinformatics analysis that studies transcriptomics, we often end up with a list of genes that previous studies have shown are associated with a disease, and we wish to prioritize other genes that may be associated with that disease. Given a network of interactions among these genes, we invoke the principle that disease-related genes are more likely to have biological interactions with each other than with randomly chosen genes. A straightforward analysis approach might be to predict that all the direct neighbours of disease genes in the network are also disease genes. However, such a naive approach would potentially introduce false predictions (false positives) that are connected to disease genes by irrelevant edges; it would also miss genes (false negatives) that do not directly interact with known disease genes, even if such genes are well connected to the known genes through multiple longer paths. To address this issue, one could examine longer paths in the network and could define the distance between pairs of genes by the length of the shortest path between them. One could then prioritize new genes on the basis of their distance to the prior list. However, the difficulty with this approach is that many genes will be near disease genes owing to the 'small world' property of most biological networks; that is, the property that most nodes can be reached from every other node in a small number of steps. Thus, such an approach might return many false-positive genes that connect to disease genes through paths that contain irrelevant or erroneous interactions. Network propagation offers a more refined approach by simultaneously considering all possible paths between genes. The application of network propagation to gene ranking can thus overcome some of the difficulties associated with shortest path-based approaches. Specifically, potentially spurious predictions (false positives) that are supported by a single (shortest) path are down-weighted, and true causal genes that are potentially missed, even though they are well connected to the prior list (false negatives), are promoted.

Network-based propagation techniques include random walks on a graph, diffusion processes on a graph and current computations in electric networks. The starting point is a vector $p^0(v)$ of scores on genes representing our prior knowledge or

experimental measurements. For example. ve could set $p^0(v) = 1$ for known disease genes and $p^0(v) = 0$ for all other genes. Alternatively, we could set $p^0(v)$ to represent some measure of confidence in the role of v in a disease, for example, its frequency of somatic mutations when studying cancer cohorts. Conceptually, one can think of $p^0(v)$ as an amount of heat, fluid or information that diffuses (or flows) over the edges of the network. At each time point t, the amount of information at each node v depends on the sum of the information at the neighbouring (adjacent) nodes $N(v)$ at time $t-1$, in proportion to the weights on the corresponding edges, according to the following equation:

$$p^t(v) = \sum_{v \in N_b} p^{t-1}(u) w(u, v)$$

where $w(\mathrm{u}, \mathrm{v})$ is the (normalized) weight or the confidence of the interaction between u and v. If we run this process for t steps, then the values in the resulting vector $p^t(v)$ give us a ranking of each node. When t is small, the ranking is close to the initial distribution $p^t(v)$, but when t is large, the information diffuses away from the initial distribution and reflects the network topology. The propagation process rewritten in matrix notation as follows:

$$p_t = W p_{t-1}$$

where W is a normalized version of the adjacency matrix of the network of interest. Repeated iteration of this equation yields $p_{\mathrm{z}} = W^t p^0$, where p^0 represents our initial, or prior, information on genes. If W is a stochastic matrix, that is, its columns sum to 1, this process is equivalent to a random walk on the network, where a walker traverses the nodes, each time moving :o a random neighbour of the present position with a probability given by (the transpose of) W. Alternatively, we can view the edges as representing conductance in an electric network with some designated source and target. If one unit of current flows through the source, then the amount of current flowing through any edge is the frequency with which a random walker traverses that edge on the way from the source of the target. Another version of the propagation process is the random walk with restart (RWR; also known as insulated diffusion and personalized PageRank) :

$$p_k = \alpha p^0 + (1-\alpha)Wp^{t-1}$$

where the parameter α describes the trade-off between prior information and network smoothing. When the network is connected and the eigenvalues of W are at most 1 in absolute value, then this process can be shown to converge to a steady-state distribution:

$$p = \alpha(I - (1-\alpha)W)^{-1}p^0$$

Different variants may use different ways of defining W based on the adjacency matrix A of the network (which could be weighted or unweighted) and the diagonal degree matrix D, the diagonal entries of which hold the node degrees and all other entries are 0 . The random walk above uses $W = AD^{-1}$. Other approaches set W to $D^{-1/2}AD^{-1/2}$, which also satisfies the convergence conditions.

In both cases, the final ranking can be obtained from the initial ranking by matrix multiplication: if we denote by p either the steady-state distribution or the diffusion at some time point t, then $p = Sp^0$ for some appropriately defined matrix S. This matrix can be interpreted as a (potentially asymmetric) similarity matrix, in which each entry S_{ij} gives the amount of information propagated to node i, given that the initial ranking p_0 is an elementary vector with 1 at entry j and 0 elsewhere.

Furthermore, if S is symmetric and positive semi-definite, then S defines a kernel. For example, the diffusion kernel is the continuous-time analogue of RWR, where $S = e^{-\alpha W}$ and $W = D - A$ is called the network's Laplacian matrix

Let t denotes the number of time steps; A denotes the adjacency matrix, which could be weighted or unweighted; D denotes the diagonal degree matrix; α is the smoothing parameter

Name	Similarity matrix	Weight normalization
Random walk	W^t	$W = AD^{-1}$
Random walk with restart	$\alpha(l-(1-\alpha)W)^{-1}$	$W = AD^{-1}; W = D^{-1/2}AD^{-1/2}$

Specifically addressing the random walk with restart (RWR) algorithm, it can be seen as a random walker, starting from one known node, travelling randomly through the graph with the addition that at each iteration t it has a probability of returning directly to the initial node. Given a connected weighted graph $G(V, E)$ with a set of nodes $V = \{v_1, v_2, \dots v_N\}$ a set of link $E = \{(v_i, v_j) \mid v_i, v_j \in V\}$ and a set of source/seed nodes $S \subseteq V$. Here, it is presented the RWR algorithm for measuring the importance of each nodes v_i to the source S. Random walk with restart can be computed iteratively according to the power iteration method (Page et al. 1999)

$$\boldsymbol{p}^t = (1 - a)\tilde{A}\boldsymbol{p}^{t-1} + c\boldsymbol{p}^0$$

A is the $N \times N$ adjacency matrix of the graph $G = (E, V)$ which represents the relations between cellular components, the binary matrix $\boldsymbol{A}$ is than converted into $\tilde{A} = AD^{-1}$, where D is the degree vector. $\tilde{A}$ is the transition matrix of the graph, the (u, j) element in $\tilde{A}$ denotes a probability with which a walker in position v_i moves to $v_j \in V \setminus \{v_i\}$. Formally, $\tilde{\boldsymbol{A}}_{i,j}$ is defined as follows (i.e. row normalization)

$$\tilde{\boldsymbol{A}}_{i,j} = \frac{\boldsymbol{A}_{i,j}}{\sum_j \boldsymbol{A}_{i,j}}$$

$\boldsymbol{p}^0$ is the initial probability vector defined as follows:

$$p^0 = \begin{cases} \frac{1}{|S|}, & \text{if } v_i \in S \\ 0, & \text{otherwise} \end{cases}$$

a is the probability of the random walker to restart at initial position. The $\boldsymbol{p}^0$ can be obtained from the expression levels of genes or from known genes related to a specific disease. At each time point $t - 1$, the random walk either flows from the current node u to a randomly chosen neighbour $v \in V$ or restarts at one vertex in p^0. The propagation of $\boldsymbol{p}^t$ strictly depends on $\boldsymbol{p}^{t-1}$ and it is run iteratively with sufficient step until $\boldsymbol{p}^t$ converges to a steady-state $\boldsymbol{p}$. The vector $\boldsymbol{p}$ is guaranteed to converge to a unique solution if $0 < a < 1$ (Google's PageRank and Beyond 2012). Iterative methods have expensive query cost, especially when there are lots of

queries since the whole iterations need to be repeated for each one. $\boldsymbol{p}$ can be instead obtained by solving the following linear equation:

$$\begin{aligned}(\boldsymbol{I} - (1-a)\tilde{\boldsymbol{A}})\boldsymbol{p} &= c\boldsymbol{p}^0 \\ \Leftrightarrow \boldsymbol{H}\boldsymbol{p} &= c\boldsymbol{p}^0\end{aligned}$$

When dealing with a matrix there are mainly two reason to represent it as a compressed matrix. First, if a sparse matrix of dimension N × M contains only a few nonzero elements, it is surely inefficient to store N M elements, second, it is often physically impossible to allocate storage for all N M elements. It is also important to note that even if we can allocate such storage, it would be inefficient or prohibitive in machine time to iterate over all of it in search for nonzero elements. For this job, an indexed storage scheme is required, one that stores only nonzero matrix elements, with the addition of auxiliary information to determine where the element belongs in the fully constructed matrix. One naive way would be to store the coordinate list (COO). This is nothing more than loading the edge list from the input file into two separate vectors plus a vector representing the data (in our case the probability of going from node u to its neighbours). This is however very inefficient since we cannot retrieve in constant time the neighbours of a node. Another obvious data structure is a list of the nonzero values for each row (or column, based on convenience) from which we can access using a vector of sparse row. This data structure is good, but it is not very efficient when one needs to loop over all elements of the matrix. A good general storage scheme is the compressed sparse row (CSR) format (Isaacson 1989), also called compressed row storage or Yale format. In this scheme, three vectors are used: val for nonzero values, col_ind for the corresponding col indices of each value, and row_ptr for the location in the other two arrays that start a row. In other words, if $val[k] = a[i][j]$, then $col\_ind[k] = j$. The first nonzero in row i is at $row\_ptr[i]$. The last is at $row\_ptr[i+1] = 1$. $row\_ptr[0]$ is always 0, and by definition $row\_ptr[N]$ equals to the number of nonzeros. Note that the dimension of row_ptr array is $N + 1$, not N. The advantage of this scheme is that it requires storage of only about two times the number of nonzero matrix elements.

In the original algorithm, source nodes have equal value. Instead in this case, source nodes can be assigned unequal weights. Assuming that the source set S can be categorized into M subsets $(S_1, S_2, \dots S_M)$ with the corresponding weights $(w_1, w_2, \dots \sum_{i=1}^{M} w_i = 1)$. Therefore, the initial probability vector is defined as follows:

$$p^0 = \begin{cases} w_1 \dfrac{1}{|S_1|}, & \text{if } v_i \in S_1 \\ w_2 \dfrac{1}{|S_2|}, & \text{if } v_i \in S_2 \\ \dots & \\ w_N \dfrac{1}{|S_M|}, & \text{if } v_i \in S_M \\ 0, & \text{otherwise} \end{cases}$$

This restart vector can be obtain from experimental data. For example, Ye et al. applied a network propagation method, based on RWR, to smooth expression values. The starting point is a vector p^0 of scores on genes which are derived from the gene expression profile of a given cell.

Example of application of the network-based propagation as artificial intelligence technique that predicts the activity of biological entities can be found at chapter 8 section: Esearch3D

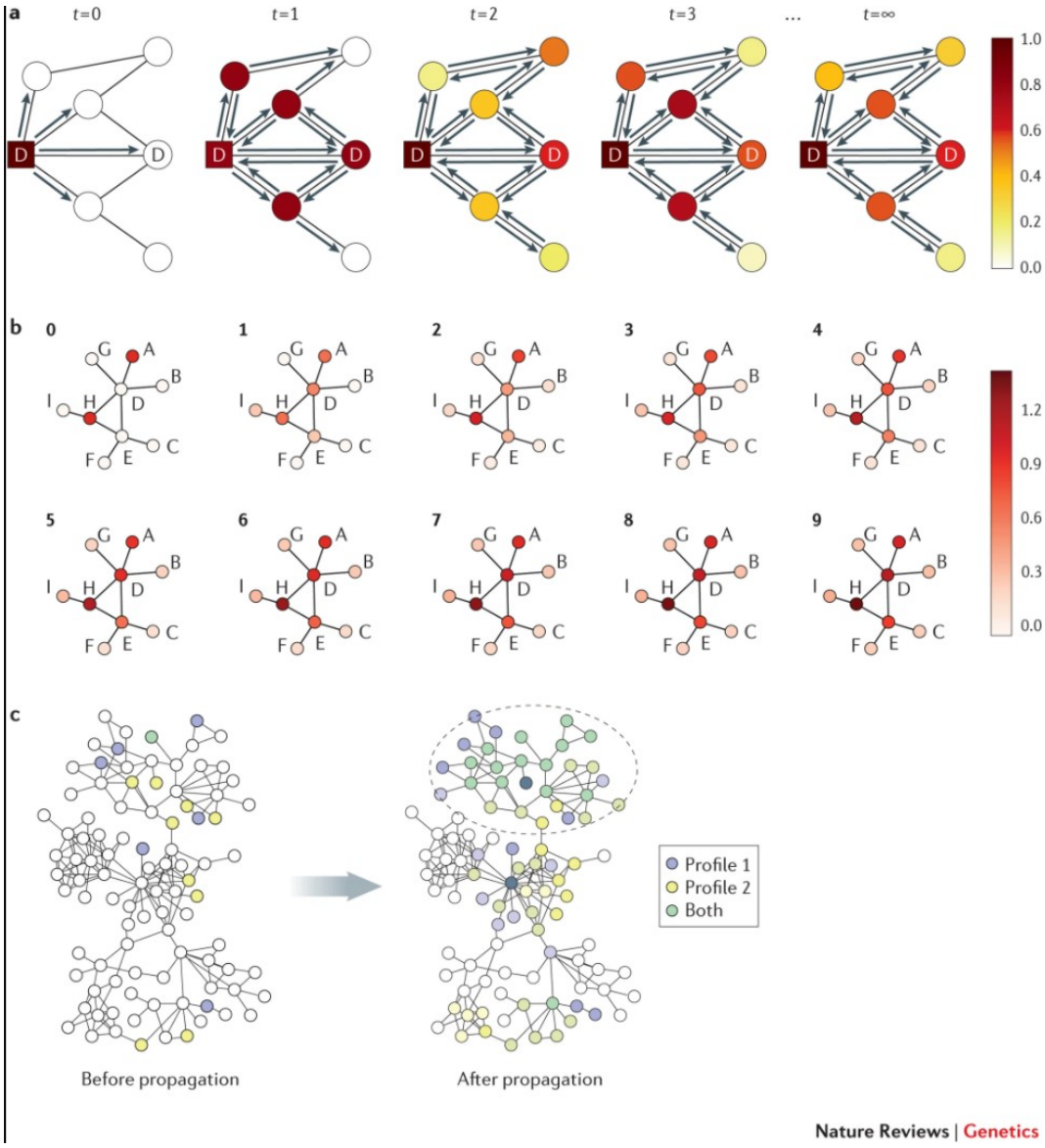


Fig.3.10 [25] Schematic illustration of network propagation. A: step-by-step demonstration of network propagation. The propagation process is depicted at different time points until convergence (steady-state (t =∞)). Arrows depict the direction of the flow or walk. Nodes are colour-coded according to the amount of flow that they receive. Letter "D" indicates nodes that are known (square node) or that are predicted (circular node) to be associated with a disease phenotype. B: example network with initial high scores for two of nine nodes (step 0, nodes A and H; score shown by colour bar). These scores are allowed to propagate over stepwise iterations 0–9; note that convergence is reached by approximately step 5 and thus the colours do not change markedly in subsequent steps. C: illustration of a biological network with gene scores before and after propagation, performed independently for two data sets (profile 1 and profile 2). Propagation results in greater concordance between the data sets, as is evident from the greater number of green nodes (dashed oval).

4. MACHINE LEARNING BACKGROUND

This chapter summarizes the concepts of machine learning. netDx, Pratic and Simpati are methods that fall into this computer science and mathematical branch. They learn patient data in form of similarity networks and solve a task called supervised classification. This chapter allows to put all the objects processed and operations performed by the methods in the machine learning context. Hence, it helps to improve the understanding of their workflows.

- Section 4.1. introduces the idea behind machine learning and the terminology associated to specific tasks and objects
    - netDx, Pratic and Simpati perform specific tasks which make them different from enrichment methods; they “learn”, “train”, “predict”, “simulate a human way of thinking” and “cross-validate” the patient information that they considerate meaningful.
- Section 4.2. introduces the classification as one type of machine learning that can be performed
    - netDx, Pratic and Simpati require to know the patient classes in comparison to find information which enable them to predict the class of an unknown patient. This task is called supervised classification.
- Section 4.3. introduces the measures used to assess the quality of a machine learning algorithm
    - Classification performances are described by an aptly named tool called the confusion matrix or truth table. The latter compare the predictions made by a method with the true labels (aka classes in our case) of the subjects (aka patients) of the classification. There are different measures which can be determined from the confusion matrix and each one provides a specific information about how good the classifier has classified.
- Section 4.4. introduces the techniques that are applied to get the quality measurements

    - netDx, Pratic and Simpati use the same measures to determine their classification performances but different techniques to get the measurements. netDx and Pratic perform multiple runs of classification, create a truth table for run, determine the performance of the single run and then summarize the classification results with their median. On the contrary, Simpati performs a single run and its only truth table provides the final performance.
- Section 4.5. introduces the task of selecting the most important data describing the subjects that the algorithm uses to learn and predict
    - netDx, Pratic and Simpati include a step of selection of the most important patient similarity networks. This allows them to increase their classification performances and reduce their system usage.

Machine learning consists in programming computers to optimize a performance criterion by using example data or past experience. The optimized criterion can be the accuracy provided by a predictive model in a modelling problem, and the value of a fitness or evaluation function in an optimization problem [29,30].

In a modelling problem, the 'learning' term refers to running a computer program to induce a model by using training data or past experience. Machine learning uses statistical theory when building computational models since the objective is to make inferences from a sample. The two main steps in this process are to induce the model by processing the huge amount of input data and to represent the model and making inferences efficiently. It must be noticed that the efficiency of the learning and inference algorithms, as well as their space and time complexity and their transparency and interpretability, can be as important as their predictive accuracy. The process of transforming data into knowledge is both iterative and interactive. The iterative phase consists of several steps. In the first step, we need to integrate and merge the different sources of information into only one format. In the second step, it is necessary to select, clean and transform the data. To carry out this step, we need to eliminate or correct the uncorrected data, as well as decide the strategy to impute missing data. This step also selects the relevant and non-redundant variables; this selection could also be done with respect to the instances. In the third

step, called data mining, we take the objectives of the study into account to choose the most appropriate analysis for the data. In this step, the type of paradigm for supervised or unsupervised classification should be selected and the model will be induced from the data. Once the model is obtained, it should be evaluated and interpreted, both from statistical and biological points of view, and, if necessary, we should return to the previous steps for a new iteration.

The model satisfactorily checked, and the new knowledge discovered are then used to solve the problem. Optimization problems can be posed as the task of finding an optimal solution in a space of multiple (sometimes exponentially sized) possible solutions. The choice of the optimization method to be used is crucial for the problem solution. Optimization approaches to biological problems can be classified, according to the type of solutions found, into exact and approximate methods. Exact methods output the optimal solutions when convergence is achieved. However, they do not necessarily converge for every instance. Approximate algorithms always output a candidate solution, but it is not guaranteed to be the optimal one. Optimization is also a fundamental task when modelling from data. In fact, the process of learning from data can be regarded as searching for the model that gives the data the best fitting. In this search, in the space of models any type of heuristic can be used. Therefore, optimization methods can also be seen as an ingredient at modelling.

4.1 Supervised Classification

In a classification problem, we have a set of elements divided into classes. Given an element (or instance) of the set, a class is assigned according to some of the element's features and a set of classification rules. In many real-life situations, this set of rules is not known, and the only information available is a set of labelled examples (i.e., a set of instances associated with a class). Supervised classification paradigms are algorithms that induce the classification rules from the data.

In two-group supervised classification, there is a feature vector $\mathrm{X} \in \Re^n$ whose components are called predictor variables and a label or class variable $\mathcal{C} \in \{0,1\}$. Hence, the task is to induce classifiers from training data, which consists of a set of

N independent observations $\mathcal{D}_N = \{(\mathbf{x}^{(1)}, \mathrm{c}^{(1)}), \ldots, (\mathbf{x}^{(N)}, \mathrm{c}^{(N)})\}$ drawn from the joint probability distribution $p(\mathbf{x}, c)$ as shown here:

	$\boldsymbol{X_1}$	...	$\boldsymbol{X_n}$	$\boldsymbol{c}$
$(\boldsymbol{x}^{(1)}; c^{(1)})$	$x_1^{(1)}$	...	$x_n^{(1)}$	$c^{(1)}$
$(\boldsymbol{x}^{(2)}; c^{(2)})$	$x_1^{(2)}$	...	$x_n^{(2)}$	$c^{(2)}$
$(\boldsymbol{x}^{(N)}; c^{(\mathrm{N})})$	$x_1^{(N)}$	...	$x_n^{(\mathrm{N})}$	$c^{(N)}$
$\boldsymbol{x}^{(N+1)}$	$x_1^{(N+1)}$	...	$x_n^{(N+1)}$	??

Table 4.1. Raw data in a supervised classification problem

The classification model will be used to assign labels to new instances according to the value of its predictor variables.

4.2 Quality Measurements

When a $0/1$ loss is used, all errors are equally bad, and our error calculations are based on the confusion matrix:

		Predicted class	Predicted class
		Positive	Negative
True class	Positive	TP: True positive	FN: False negative
True class	Negative	FP: False positive	TN: True negative

In this case, we can define the error rate as $(|FN| + |FP|)/N$, where $N = |TP| + |FP| + |TN| + |FN|$ is the total number of instances in the validation set. To fine-tune a classifier, another approach is to draw the receiver operating characteristics

(ROCs) curve, which shows hit rate versus false alarm rate, namely, 1 -specificity $= |FP|/(|FP| + |TN|)$ versus sensitivity $= |TP|/(|TP| + |FN|)$. For each classification algorithm, there is a parameter, for example, a threshold of decision, which we can play with to change the number of true positives versus false positives. Increasing the number of true positives also increases the number of false alarms; decreasing the number of false alarms also decreases the number of hits. Depending on how good/costly these are for the particular application we have, we decide on a point on this curve. The area under the receiver operating characteristic curve is used as a performance measure for machine learning algorithms.

4.3 Classification Error

An important issue related to a designed classifier is how to estimate its (expected) error rate when using this model for classifying unseen (new) instances. The simplest and fastest way to estimate the error of a designed classifier in the absence of test data is to compute its error on the sample data itself. This resubstitution estimator is very fast to compute and a usually optimistic (i.e., low-biased) estimator of the true error.

In k-fold cross-validation , $\mathcal{D}_N$ is partitioned into k folds. Each fold is left out of the design process and used as a testing set. The estimate of the error is the overall proportion of the errors committed on all folds. In leave-one-out cross-validation, a single observation is left out each time, which corresponds to N-fold cross-validation.

The bootstrap methodology is a general resampling strategy that can be applied to error estimation. It is based on the notion of an 'empirical distribution', which puts mass $1/N$ on each of the N data points. A 'bootstrap sample' obtained from this 'empirical distribution' consists of N equally likely draws with replacement from the original data set $\mathcal{D}_N$.

4.4 Feature Selection

One question that is common to all supervised classification paradigms is whether all the n descriptive features are useful when learning the classification rule. In

trying to respond to this question, the so-called feature subset selection (FSS) problem appears, which can be reformulated as follows: given a set of candidate features, select the best subset under some learning algorithm. This dimensionality reduction made by an FSS process can bring several advantages to a supervised classification system, such as a decrease in the cost of data acquisition, an improvement in the understanding of the final classification model, a faster induction of the final classification model and an increase in the classifier accuracy.

FSS can be viewed as a search problem, with each state in the search space specifying a subset of the possible features of the task. An exhaustive evaluation of possible feature subsets is usually unfeasible in practice because of the large amount of computational effort required. Four basic issues determine the nature of the search process: a search space starting point, a search organization, a feature subset evaluation function and a search-halting criterion. The search space starting point determines the direction of the search. One might start with no features and successively add them, or one might start with all the features and successively remove them. One might also select an initial state somewhere in the middle of the search space. The search organization determines the strategy of the search in a space of size 2^n, where n is the number of features in the problem. The search strategies can be optimal or heuristic. Two classic optimal search algorithms which exhaustively evaluate all possible subsets are depth-first and breadth-first [31]. Otherwise, branch and bound search [32] guarantees the detection of the optimal subset for monotonic evaluation functions without the systematic examination of all subsets. When monotonicity cannot be satisfied, depending on the number of features and the evaluation function used, an exhaustive search can be impractical. In this situation, heuristic search is interesting because it can find near optimal solutions, if not optimal. Among heuristic methods, there are deterministic and stochastic algorithms. On one hand, classic deterministic heuristic FSS algorithms are sequential forward and backward selection, floating selection methods [33] or best-first search [34]. They are deterministic in the sense that all runs always obtain the same solution and, due to their hill-climbing nature, they tend to get trapped on local peaks caused by interdependencies among features. On the other hand, stochastic heuristic FSS algorithms use randomness to escape from local maxima,

which implies that one should not expect the same solution from different runs. The evaluation function measures the effectiveness of a particular subset of features after the search algorithm has chosen it for examination. Each subset of features suggested by the search algorithm is evaluated by means of a criterion (accuracy, area under the ROC curve, mutual information with respect to the class variable, etc.) that should be optimized during the search. In the so-called wrapper approach to the FSS problem, the algorithm conducts a search for a good subset of features using the error reported by a classifier as the feature subset evaluation criterion. However, if the learning algorithm is not used in the evaluation function, the goodness of a feature subset can be assessed by only regarding the intrinsic properties of the data. The learning algorithm only appears in the final part of the FSS process to construct the final classifier using the set of selected features.

5. BIOINFORMATICS BACKGROUND

This chapter summarizes the enrichment methods that are applied in bioinformatics to analyse biological high-throughput data. It mathematically formulates the two most common operations: the differential expression and pathway analysis. Enrichment methods are not based on artificial intelligence or machine learning, but they offer a quick deterministic resolution to get the altered single molecules and cellular functions of a patient [35–37]. At the same time, they have limitations. They are parametric and base their operations on statistical assumptions, do not perform multivariate analysis, do not learn, do not simulate a human way of thinking and do not consider indirect relationships between biological features (molecules or pathways) and the patient classes in comparison [38].

- Section 5.1. introduces the standard differential expression analysis
    - Despite this thesis focus on the pathway analysis, the task of finding single altered molecules between the patient class of interest and the reference one has always been complementary to the task of discovering biologically relevant pathways. Differential expression analysis also offers the opportunity to understand better the meaning of comparing the patient class of interest against the reference one.
- Section 5.2. introduces the pathway analysis and the standard enrichment method to perform it.
    - netDx, Pratic and Simpati learn pathways and classify patients. If the classification performances are high, then the learnt pathways can be considered significantly associated to the disease phenotype in study and altered with respect the reference phenotype. In other words, the classifiers perform a pathway analysis. The best pathways, instead of being statistically significant as in enrichment methods, are considered biomarkers because are tested for holding a correlation between genotype and phenotype.
- Section 5.3. describes the pros and cons of enrichment methods for the pathway analysis with respect machine learning algorithm

    - Enrichment methods are the only used ones for the pathway analysis. They are easy to understand and to use, they require low system requirements, they are fast and they provide the most important results. They have few limitations which do not really affect the advantages.

- Section 5.4. describes the pros and cons of machine learning algorithms for the pathway analysis with respect enrichment methods
    - Machine learning methods for the pathway analysis are not used or developed. They are computational expensive, lack of statistics, difficult to understand, to be trust and to use. The few advantages are not enough for being relevant and to let classifiers to win the battle against enrichment methods. However, netDx, Pratic and Simpati are based on the paradigm of patient similarity networks. The latter allows them to be understandable and trusted. Simpati also generates and provides new information about the patient classes in comparison for being more useful than any enrichment method.

5.1 Differential Expression Analysis

Next-generation sequencing techniques enable researchers to produce and access massive amounts of data than previously available. One example is the bulk RNA-sequencing (RNA-seq) technology which provides information regarding the expression levels of thousands of genes in patients characterized by different phenotypes. Naturally arising from this information is the concept of differentially expressed genes (DEGs), which are genes having significantly different expression levels between the phenotypes (i.e. patient classes). The same concept can be extended to non-coding RNAs, proteins and other omics captured with high-throughput technologies; the only requirement is to have continuous values determining the level of abundance, activity, or expression of the molecule in the patients describing the phenotypes to compare.

There are currently many enrichment methods that perform the differential expression (DE) analysis. They are mainly developed with R language and make

heavy use of numerous statistical methods that have been developed and implemented over the past two decades to improve the power to detect robust changes based on extremely small numbers of samples (i.e., sequenced patients that compose the two conditions in comparison, for example cancer patients versus healthy or late stage cancer patients vs early stage) [35,39,40]. As it is possible to guess, each method developed its own statistical test for determining which genes or generic molecules have a statistically significant difference. At the same time, even if with different operations, they all aim to resolve two tasks:

1. Estimate the magnitude of the differential expression between two phenotypic conditions
2. Estimate the significance of the difference and correct for multiple testing

Enrichment methods for the DE analysis estimate the expression difference for a given gene using regression-based models and perform a statistical test based on the null hypothesis that the difference is close to zero, which would mean that there is no difference in the gene expression values that could be explained by the conditions [39,40]. Here, it is proposed the simplest scenario and methodology in which the estimation of the difference is done with a linear regression model.

The linear regression model can be defined with the form: $Y = b_0 + b_1 * x + e$

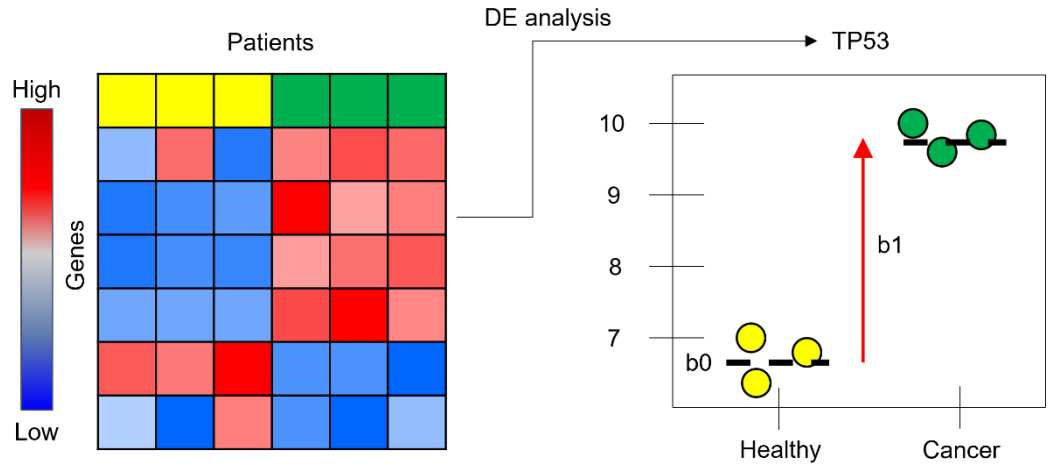


Figure 5.1: Example of comparison between two conditions. On the left, the gene expression and transcriptomic matrix composed by the patient's molecular profiles divided in two classes; yellow indicates the healthy class and represents the control phenotype, while green is the disease phenotype and composed by cancer patients. The legend maps a color to an expression level. Genes are on the rows and one of them called "TP53" undergoes the test of the DE analysis. In the frame of the right,

Y entails the expression values from all phenotypes which a gene describes, b_0 represents the average value of the baseline group and b_1 the difference between baseline and non-reference group: $Y = b_0 + b_1 * x + e$. x is a discrete parameter, which is set to 0 for expression values measured in the baseline condition (here: Healthy) and set to 1 for the non-reference group (here: $Cancer$). b_0 is called the intercept; x is the condition, e is a term capturing the error or uncertainty, and b_1 is the coefficient that captures the difference between the samples of different conditions.

The model shown above could be fitted in R using the function $1m(rlog.norm \sim genotype)$ (which will return estimates for both b_0 and b_1, so that the average expression values of the baseline genotype (e.g., $\text{Cancer} = 0$) would correspond to $Y = b_0 + b_1 * 0 + e$. This is equivalent to $Y = b_0$ (assuming that e is very small), thereby demonstrating why the intercept (b_0) can be interpreted as the average of our baseline group. b_1, on the other hand, is the coefficient whose closeness to zero is evaluated during the statistical testing step.

The null hypothesis is that there is no systematic difference between the average expression values of the different conditions for a given gene. T−test (How dissimilar are the expression means of the two conditions) or ANOVAs (How well does a reduced model capture the data when compared to the full model with all coefficients) are generically well-suited to offer the model for performing the test and providing the probability value. At this stage, it is obtained a list of probability values, one for each sequenced gene describing the patients. DE methods correct for the multiple tested hypotheses all related to the comparison of the same patient condition; for example the Benjamini-Hochberg formula provides final adjusted probability values and allows to reduce false positive differential expressed genes.

The result of a differential expression analysis is a set of molecules which are statistically different between the two conditions of patients because having an adjusted probability value lower than 0.05. A common scenario includes cases versus controls, the set is assumed to be composed of molecules which are altered due to the patient's disease, either upregulated (activated, expressed or abundant more than in a healthy cellular state) or downregulated. The latter explains the direction of change of a specific molecule and is quantitative measured with the fold change (i.e., fold change greater than zero means upregulation).

After this operation, the downstream analysis aims to elucidate the functions of the DE molecules and to identify possible patterns among them. Standard operations of downstream analysis are Over-representation analysis and the Pathway enrichment analysis [35–37].

The over-representation analysis (ORA) relies on a filtered list of molecules of interest, (e.g., all genes that pass the DE analysis). This list is compared to the molecules that are known to be part of a specific pathway or a generic set of interest (e.g., “Glycolysis"). A statistical test (e.g., hypergeometric test) is then used to determine whether the overlap between the list of interest and the known pathway is greater than expected by chance. While this approach is relatively straight-forward, there are serious limitations including the fact that both magnitude and direction of the change (up or down regulated) of individual molecules are completely disregarded; the only measure that matters is the presence of absence of a given molecule within the lists that are being compared.

The Pathway analysis addresses some of the limitations of the ORA approach, functional class scoring (FCS) algorithms typically do not require a pre-selected list of genes; instead, they require a fairly exhaustive list of all molecules that could make up your “universe". These molecules should have some measure of change by which they will be sorted. The basic assumption is that although large changes in individual molecules can have significant effects on pathways, weaker but coordinated changes in sets of functionally related molecules (i.e., pathways) can also have significant effects. Therefore, the molecular level statistics for all molecules in a pathway are aggregated into a single pathway-level statistic (e.g., the sum of all log-fold changes), which is then evaluated.

5.2 Pathway Analysis

High-throughput technologies such as DNA microarrays and RNA sequencing are widely used to monitor the activity of thousands of genes and molecules in a single experiment. The primary challenge to realizing the potential of these technologies is gaining biological insight from the generated data. The early approach was the single-gene analysis (extendable also to other types of omics and molecules), where expression measurements of each gene for case and control samples are compared

using a statistical test such as t-test or Wilcoxon rank-sum test and a p-value is calculated. Then, in order to reduce the number of false positives resulting from multiple comparisons, an adjustment for multiple comparison is made. Next, genes with an adjusted p-value smaller than 0.05 are predicted as being differentially expressed. Finally, a biological interpretation is attempted using these genes.

This approach suffers from several shortcomings[41–43]:

- In a high-throughput study, many single-gene tests are typically performed. Consequently, adjustment for multiple comparisons is performed for a large number of genes. Such adjustments may lead to many false negatives by detecting very few or even no gene as being differentially expressed. This issue is more pronounced when using conservative methods, such as Bonferroni.
- Cellular processes are often associated with changes in the expression patterns of groups of genes that share common biological functions or attributes. A meaningful change in a group of these genes is more biologically reliable and interpretable than a change in a single gene. A priori knowledge about some of these pathways is available through public online databases such as GO [44] and KEGG [45]. The single-gene approach disregards this information. Incorporating this information in the data analysis may provide valuable insight about underlying biological processes or functions.
- Although high-throughput technologies make the monitoring of expression of thousands of genes in a single experiment possible, they introduce a challenge of dealing with high dimensional data, often referred to as the "curse of dimensionality". To deal with high dimensional data, dimensionality reduction methods are used for downstream analyses and visualizations. Relying on sets of biologically related genes is the most intuitive and biologically relevant approach to dimensionality reduction in high-throughput studies.
- Multi-functional genes, i.e., genes that are involved in multiple biological activities, are commonplace. For example, it has been reported that multi-

functional genes make up 24, 26, and 19% of annotated genes in Drosophila melanogaster, Homo sapiens, and Saccharomyces cerevisiae, respectively. The presence of such a large number of multifunctional genes means single-gene analysis may lead to false or ambiguous conclusions.

- Single-gene approach may report several hundred to a few thousand genes as being differentially expressed. Interpreting a long list of differentially expressed genes is a cumbersome task prone to investigator bias toward a hypothesis of interest.

Pathway analysis, also called enrichment or gene set analysis, is an attempt to resolve these shortcomings. The aim is to identify differentially enriched pathways between two patient's conditions. Similarly, as for the differential expression analysis, this thesis mathematically introduces ORA and FCS methodology [41].

Let us assume to have data from a high-throughput case-control experiment. They can be organized as an expression matrix. This matrix is generated by joining the corresponding expression values for all samples in the experiment. Each column of the matrix corresponds to the expression measures for the sample of one patient and each row corresponds to the expression measures for one gene across all samples. This expression matrix is the input for expression analyses including single-gene and gene set analysis. Figure 5.2 shows an expression matrix with ||C|| control samples and ||T|| case samples.

$$\begin{array}{c} A^{(c_1)} \quad \dots \quad A^{(c_{\|C\|})} \quad A^{(t_1)} \quad \dots \quad A^{(t_{\|T\|})} \\ \begin{bmatrix} g_1^{(c_1)} & \dots & g_1^{(c_{\|C\|})} & g_1^{(t_1)} & \dots & g_1^{(t_{\|T\|})} \\ g_2^{(c_1)} & \dots & g_2^{(c_{\|C\|})} & g_2^{(t_1)} & \dots & g_2^{(t_{\|T\|})} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ g_{m-1}^{(c_1)} & \dots & g_{m-1}^{(c_{\|C\|})} & g_{m-1}^{(t_1)} & \dots & g_{m-1}^{(t_{\|T\|})} \\ g_m^{(c_1)} & \dots & g_m^{(c_{\|C\|})} & g_m^{(t_1)} & \dots & g_m^{(t_{\|T\|})} \end{bmatrix} \end{array}$$

Fig.5.2 [41] Expression matrix for a pairwise comparison where $A^{(c_1)}, \dots, A^{(c_{\|C\|})}$ columns represent control samples and $A^{(t_1)}, \dots, A^{(t_{\|T\|})}$ columns represent case samples. In this figure, $g_i^{(c_j)}$ and $g_i^{(t_j)}$

represent the expression measures for the i^{th} gene in the c_j^{th} control sample and t_j^{th} case sample, respectively. $g_i^{(c_j)}$ is the gene expression level for gene g_i in sample $A^{(c_j)}$

ORA uses a list L of genes each predicted as being differentially expressed by a single-gene analysis method. Given L and a gene set G_i that has n_i' genes in common with L, ORA considers G_i as being differentially enriched if the occurrence of n_i' differentially expressed genes in G_i is unlikely to be due to chance.

	Genes in L	Genes not in L	Total
Genes in G_i	n_i'	$\|G_i\| - n_i'$	$\|G_i\|$
Genes in $\bar{G}_i$	$\| L \| - n_i'$	$n - \|G_i\| - (\| L \| - n_i')$	$n - \|G_i\|$
Total	$\| L \|$	$n - \| L \|$	n

This table illustrates the contingency scheme for the over-representation of differentially expressed genes in G_i given L and U, where $\bar{G}_i$ is the set of all genes under study that are not members of G_i.

The set of n genes under study is called the reference set or background set and depicted by U, and $\bar{G}_i$ is the complement of G_i with respect to U. Under the null hypothesis that there is no association between differential expression and membership in G_i, we can assume that G_i is the result of a simple random sampling of $\|G_i\|$ genes from U; therefore, the probability of having n_i' differentially expressed genes within G_i can be calculated using the hypergeometric distribution as follows (Drăghici, 2016):

$$f(n_i'; n, \|G_i\|, \| L \|) = \frac{\binom{\|G_i\|}{n_i'} \times \binom{n - \|G_i\|}{\| L \| - n_i'}}{\binom{n}{\|L\|}}$$

The significance of the association between genes in G_i and genes in L can be assessed using Fisher's exact test, as follows:

$$p = \sum_{j=n_i'}^{\|G_i\|} f(j; n, \|G_i\|, \| L \|) = 1 - \sum_{j=0}^{n_i'-1} f(j; n, \|G_i\|, \| L \|)$$

The main assumptions of ORA are that genes are independent and equally effective in biological processes. Although these assumptions simplify problem modelling, they are not biologically valid. It is well-established that genes, proteins, and other biomolecules often act in concert. In addition, ORA only utilizes differentially expressed genes, which often are the result of applying a p-value cutoff, and all the quantitative measures for the rest of the genes are disregarded. However, a consistent change in the expression of genes—even those with a p-value slightly greater than the cutoff value—may contribute to the detection of pathway activities. In contrast to ORA, the main goal of FCS methods is to use all information from an expression matrix to address the enrichment problem without relying on the aforementioned biologically invalid assumptions. Therefore, FCS methods—instead of working with a list of differentially expressed genes— take advantage of an expression matrix of gene expression measures for all genes to discern differential enrichment of gene sets.

There are many FCS methods available. This thesis focuses on the main FCS methods which are univariate. A score is calculated for each gene using each row of the expression matrix. Then these scores are used to calculate a gene set score for each pathway. Finally, the significance of the gene set scores is assessed and differentially enriched gene sets are reported. An FCS method often consists of a set of common components such as a gene score that is a statistic summarizing the expression level of a gene across control and case samples, a gene set score that summarizes the expression level of genes within a gene set as a single statistic, a procedure for significance assessment, and an adjustment for multiple comparisons.

GSEA [46] is one of the most widely used univariate FCS methods. It uses a signal-to-noise ratio (SNR) difference between gene expression measures in control and case samples to calculate a gene score. The signal-to-noise ratio difference is as follows [47]:

$$\mathrm{SNR}\,(g_i) = \frac{\frac{\sum_{j=1}^{\|C\|} g_i^{(c_j)}}{\|\ C\ \|} - \frac{\sum_{j=1}^{\|T\|} g_i^{(t_j)}}{\|\ T\ \|}}{\sigma'_{c,i} + \sigma'_{t,i}}$$

$$\sigma'_{c,i} = \text{Max}\left(\sigma\left(g_i^{(c_1)}, \ldots, g_i^{(c_{\|c\|})}\right), 0.2 \times \frac{\sum_{j=1}^{\|C\|} g_i^{(c_j)}}{\| C \|}\right)$$

where $g_i^{(c_j)}$ is the gene expression level for gene g_i in sample $A^{(c_j)}$; $\sigma'_{c,i}$ is the standard deviation of expression levels for gene g_i among control samples; $g_i^{(t_j)}$ and $\sigma'_{t,i}$ are defined analogously using case samples.

GSEA ranks all genes according to their scores. Then to measure the association between members of a given gene set G_i and conditions/phenotypes, it calculates a gene set score, also called enrichment score (ES), using a Kolmogorov-Smirnov statistic. The ES value for G_i, denoted as ES (G_i), is calculated using a running sum initialized as 0 . Assume $g_1, \ldots, g_n$ is the sorted list of all genes according to SNR difference in decreasing order. For each gene in the sorted list starting with the first one the running sum (enrichment score) is updated by adding a value of $+\sqrt{\frac{n-\|G_i\|}{\|G_i\|}}$ when the gene belongs to G_i and by subtracting a value of $\sqrt{\frac{\|G_i\|}{n-\|G_i\|}}$ when the gene does not belong to G_i [46]. The ES value is calculated "as the maximum observed positive deviation of the running sum" [46] as shown in the following formula:

$$ES(G_i) = \max_{1 \le l \le n} \sum_{k=1}^{l} x_k$$

$$x_k = \begin{cases} +\sqrt{\dfrac{n - \|G_i\|}{\|G_i\|}} & R_k \in G_i \\ -\sqrt{\dfrac{\|G_i\|}{n - \|G_i\|}} & R_K \notin G_i \end{cases}$$

After calculation of the actual ES values for all gene sets, the method determines the maximum ES, denoted as MES. The significance of the calculated MES value is assessed using a permutation test. The sample labels are permuted 1,000 times, and for each permutation a MES value is calculated. Finally, the significance of MES of the actual data is calculated as the fraction of permutations that lead to an MES higher than the MES of the actual data. It should be mentioned that the significance of the MES does not provide any insight about the significance of the

enrichment score of a given gene set G_i although this is the main purpose of enrichment analysis. In fact, assessing the significance of the MES tests the null hypothesis that "no gene set is associated with the class distinction" [46], where the rank ordering is used as the measure of association. Therefore, rejection of this null hypothesis only suggests that there is at least one gene set for which the rank ordering of its members is associated with the sample classes, i.e., phenotypes. Since the enrichment score is defined as the "maximum observed positive deviation of the running sum" [46], it does not detect differential enrichment of gene sets that have the majority of their genes upregulated unless the phenotypes are swapped and the GSEA procedure is run again. Hence, this method should be considered as a one-sided test.

Based on the approach used for significance assessment, gene set analysis methods can be classified as parametric and nonparametric methods. In parametric methods, after calculating a gene set score for each gene set, a parametric distribution is used to assess the significance of this score. Non-parametric approaches, on the other hand, rely on an empirical distribution to assess the significance of the gene set scores. These methods often do not make any strong assumptions about the underlying distribution of the gene set scores. Phenotype permutation and gene sampling are the main non-parametric approaches used in gene set analysis. For example, methods such as GSEA offers both phenotype permutation and gene sampling for significance assessment.

The parametric approach is another way to assess the significance of gene set scores [48]. In this approach, first, a gene set score is proposed. Then, under the null hypothesis and by accepting some simplifying assumptions, a parametric distribution for the gene set statistic is proposed. Finally, the parametric distribution is used to assess the significance of gene set statistics. Parametric methods are built based on some knowledge or assumptions about the underlying distribution of the gene set scores. Although parametric approaches are not computationally demanding, they have been criticized as being too simplistic and unable to detect truly differentially enriched gene sets [47].

The non-parametric approaches are mainly two. Gene sampling and Phenotype permutation.

In gene sampling the significance of a gene set score S(G_i) for a given gene set $G_i$is assessed by comparing it to the scores of randomly assembled sets of ||Gi|| genes from the reference set U, i.e., all genes under study. In gene sampling method, a large number of random gene sets are assembled, and their scores are calculated. Then the significance value of the gene set score of $G_i$is calculated as the fraction of assembled gene sets that lead to stronger scores than the score of G_i, where a score in comparison to another is considered stronger if it is more in favour of rejecting the null hypothesis of interest. Since gene sampling does not depend on the number of samples, it has been widely used for gene set analysis of datasets with small sample sizes [46]. The main shortcoming of gene sampling is that it relies on the unrealistic assumption of independence between genes within a gene set. Usually, genes within a gene set show a highly correlated behaviour; therefore, a gene sampling method may incorrectly predict a gene set as differentially enriched only because of high correlation between its genes. In this regard, it may cause false positive predictions. Another shortcoming of gene sampling is being computationally demanding. For each gene set G_i, the whole process of gene set score calculation should be repeated for a large number of randomly assembled gene sets. In implementations of the gene sampling approach, usually the number of assembled gene sets is an order of magnitude of 1,000. This number of repetitions makes the significance evaluation computationally demanding.

Phenotype permutation, also known as sample permutation, assesses the significance of a gene set score of a given gene set $G_i$by permuting sample labels. First, the gene set score of $G_i$is calculated. Let S(G_i) denote the gene set score of $G_i$according to the actual gene expression profile. Then a large number of expression profiles are synthesized by permuting the sample labels, i.e., the column labels of the actual expression profile. For a synthesized expression profile, we expect no association between the expression patterns of genes in $G_i$and the phenotypes. Next, for each synthesized expression profile, the gene set score of $G_i$is calculated. Finally, the significance of S(G_i) is calculated as the fraction of the

synthesized expression profiles that lead to a stronger score than S(G_i), where a score in comparison to another is considered stronger if it is more in favour of rejecting the null hypothesis of interest. Phenotype permutation, unlike gene sampling, does not rely on the unrealistic assumption of gene independence, but it requires a large number of samples for each phenotype.

The last piece missing to define a gene set or pathway analysis method is the null hypothesis.

Defining a null hypothesis is an essential step in conducting any statistical inference. Different null hypotheses have been used in gene set enrichment analysis: competitive null hypothesis [49], and self-contained null hypothesis [49]. Competitive means that for a given gene set G_i, a competitive null hypothesis states that genes in G_i do not have a different expression pattern in comparison to the rest of the genes under study ($\bar{G}_i$). While self contained means that for a given gene set G_i, a self-contained null hypothesis states that genes in G_i do not have a different expression pattern across phenotypes.

Example of differential expression and pathway analysis can be found in the chapter 8 at the sections: miR-669c, Denovo Assembly for lncRNAs, Piezo1, Co-LCNEC and LErNet. This latter combines also biological network concepts.

5.3 Enrichment: pros and cons

Enrichment methods are an extremely popular approach to summarize the functional characteristics of seed gene sets. These methods present several advantages when compared to other approaches, as follows. First, they are quick and computationally light, often able to analyse large gene sets using only a laptop computer, especially given the large number of web tools available. This makes enrichment analysis very suitable for small labs that may not have access to high-power computing clusters or machine-learning experts or for situations where a quick summary of gene set functionality is sufficient and a more sophisticated method would be unnecessary and overly time consuming. Second, there are a wide variety of tools available covering multiple statistical methods. Many of these tools (for instance the highly popular DAVID tool [50]) are very user friendly with good

documentation and clear explanations of their methodology to allow users to determine the best method for their data. These tools tend to use methods based on classical statistical tests that non-statisticians are likely to have at least some understanding of. Finally, although less popular, Bayesian statistical methods have been incorporated into some enrichment analysis tools, allowing a more sophisticated statistical approach. The oldest of these is BayGO [51], which uses a Bayesian inference method to incorporate Goodman and Kruskal's Gamma score of association. The association of differential expression to each GO term is measured, and Monte Carlo simulations are employed to determine the probability of randomly observing a stronger level of GO term enrichment than the measured level. Other Bayesian tools are GO-Bayes [52], model-based gene set analysis [53] and multi-level ontology analysis [54], which all attempt to infer the probability that a given GO term is associated with a supplied gene set. These methods alleviate some of the concerns affecting most enrichment analysis methods, since the probability estimations account for some of the network characteristics inherent in biological data, while also considering all terms simultaneously thus removing the need for multiple hypothesis testing correction. Most Bayesian methods also have the advantage of not relying on classical tests of statistical significance, whose limitations were discussed earlier. Instead, they are based on the prior probability (before building the model) and the probability of observing the data given the model, which are, arguably, easier concepts for most people to grasp than p-values. The main disadvantages of enrichment methods are as follows. First, most enrichment methods are heavily based on tests of significance using p-values as the decision criterion. However, p-values by themselves are not adequate as the main basis for scientific conclusions, since they do not measure the effect size, importance and reproducibility of a result. For this reason, they should not be taken as definitive evidence for the existence or size of an effect [55]. Instead, researchers should use p-values to help guide a broader analysis, avoiding absolute conclusions based on them. In [56] the authors point out that the p-values of enrichment methods are often treated as a score of 'interestingness', and seldom the sensitivity and specificity of the list of 'interesting' properties are estimated. That is, little importance is given to the actual predictive power of the properties, giving more

value to differences in relative frequencies instead. The authors also make the interesting point that the definition of the seed genes (for SEA methods) and gene rankings (for GSEA methods) are based on the assumption that the higher the differential expression of a gene, the more important the gene should be considered in the analysis. This is often a valid assumption but not always; a small change in expression of a regulatory gene may be much more biologically relevant than larger changes in, for instance, a metabolism-related gene. Second, most 'traditional' statistical tests assume that the sampling units are independent. This is clearly not the case in most gene-expression experiments (where the sampling unit is usually a gene), a common application of enrichment methods. There are several regulatory genes that modulate the expression of other genes. When this assumption is not satisfied, the tests tend to make more type I errors than what would be expected (incorrectly rejecting the null hypothesis of 'no differential expression'). Third, SEA and GSEA enrichment methods (see the Overview of enrichment methods for bioinformatics section) ignore correlations between gene properties, analysing their enrichment significance independently. However, normally there are strong correlations among the gene properties; it is common that if a gene is annotated with a property, it is much more likely to be annotated with a 2nd property. This is particularly common when using GO terms, which are hierarchically structured (e.g. every gene annotated with the term 'detection of stimulus' is, by definition, also annotated with the term 'response to stimulus'). Arguably, this fact is not so detrimental to the enrichment methods as high gene correlation (mentioned in the previous paragraph), but it is still an important source of bias.

5.4 Classification: pros and cons

The main advantages of classification methods are as follows. First, most modern classification methods are nonparametric in the statistical sense (i.e., they do not assume that the data are distributed in a certain way). Instead, they adapt the learned model to the characteristics of the problem automatically during their training phase. Therefore, in principle, most classification algorithms can be used to discover very different types of relationships among variables in the data, including the discovery of highly non-linear correlations between the features (gene

properties) and the class labels (the phenotype of interest). Most enrichment methods, on the other hand, are parametric in the statistical sense, and each method performs the same statistical calculations regardless of the extent to which the data satisfies the assumptions of the statistical test used. Second, some types of classification models (e.g., decision trees) are relatively easily interpretable by users [57]. Such models can be used both for predictions and to gain insights about how the class label is related to the features in a relatively human-friendly fashion. Third, most classification methods consider multivariate interactions between the features and the class label. On the other hand, most enrichment methods analyse one feature at a time, ignoring the fact that, sometimes, two or more gene properties, when taken at the same time, can be much more predictive (or enriched) than the individual properties. The main disadvantages of classification methods are as follows. First, some classification methods lack formal statistical basis—several classification algorithms cannot make principled statistical assessments regarding the data. That is, the predictions are made without confidence intervals or p-values. Second, many classification methods are very computationally intensive. For instance, deep neural networks are very computationally demanding, often requiring the use of specialized hardware to run in reasonable times [58]. Note, however, that some well-known classification methods, like most decision tree algorithms and Naive Bayes, are relatively fast [59]. Third, hyper-parameter setting is not trivial. Recall that most classification algorithms have settings (hyper-parameters) that control important aspects of the learning process. A poor hyperparameter choice can lead to low (even close to random) predictive performance. Many classification algorithms are very sensitive to these settings, requiring either expert knowledge or computationally expensive hyper-parameter tuning methods. These tuning methods usually work by running the classification algorithm several times, with different hyper-parameter settings, estimating the predictive performance of the constructed models to determine which hyper-parameter setting is the best one. One must be careful while performing this hyper-parameter tuning to not measure the predictive performance in the ‘validation set’, where the final predictive performance estimation will be carried out, but rather in a subset of the ‘training set’. The predictive power of classification algorithms will

very likely be grossly overestimated if one uses the 'validation set' to tune the algorithm's hyper-parameters. Fourth, bioinformatics data sets often have two important particularities that can negatively impact the predictive performance of traditional classification algorithms: high class imbalance and structured biological descriptors. Regarding the issue of class imbalance, the data sets are often very unbalanced towards the negative class label—most whole-genome enrichment analyses involve thousands of genes without the phenotype of interest and only a few dozens with the phenotype of interest. Most classification algorithms do not cope well with this high level of class imbalance. However, there has been extensive research on methods for improving the performance of classification algorithms in this scenario, including the use of over(under) sampling of the minority (majority) class to create a more balanced training set [60]. Regarding the issue of structured biological descriptors, some descriptors (e.g. GO and FunCat terms) have a hierarchical structure. However, most classification algorithms treat them as unstructured, which may lead to problems due to the high correlation between terms.

6. METHODS

Supervised machine learning methods are useful in clinical genomics and biomedicine for the tasks of disease diagnosis, prognosis and treatment design. They are analytic methods that can identify patterns distinguishing and characterizing patient classes for the formulation of biomarkers; however, their interpretation remains an active area of research and an issue. Methods, that are not able to provide comprehensive explanations about how they learn and predict, suffer of the black box effect and are unlikely to be clinically successful. Clinicians and physicians must understand the operations and the features that the method used to recognize the patient's class in order to make a confident diagnosis or to take any decision. Separately, most machine learning methods neither handle missing data without prior data imputation nor consider system biology concepts when they learn biological omics data about the patients.

In this scenario, this chapter presents the patient classifiers and supervised machine learnings that have been developed based on the criteria that a classifier must be easy to understand about how it predicts the patient's classes to allow a clinician or physician to trust the results and must be able to use multiple omics data describing the patients both continuous numeric dense and sparse data types.

- Section 6.1 describes the patient classifier called netDx [61,62]. It has been developed in collaboration with the Gary's Bader research group of the University of Toronto. It is the very first supervised machine learning to classify patients based on patient similarity networks, proposes operations for using this graph type and allows the user to design the classification for the pathway analysis. It defined the workflow, introduced the practical disadvantages, and highlighted the benefits of working with PSNs to classify. For example, it is crucial to have a step of selection of the best PSNs to improve the classification performances, increase the interpretability and reduce the system usage. Patient similarity networks as features for classifying are not trivial to create, to process and to analyse. They naturally increase the model complexity. While, at the same time, they are easy to understand and to interpret when are used to predict the class of

an unknown patient. Despite researchers' effort to deal with the disadvantages during the development, the method is difficult to use, requires a non-trivial tuning of hyper-parameters and the interpretability achieved lacks accessibility and practicality.

    - netDx has been important to understand how to build a patient classifier that could exploit patient similarity networks for the pathway analysis. netDx has not been designed with the idea of performing such task but it born with the goal of integrating multiple data for the patient classification.

- Section 6.2 focuses on the netDx module called Pratic [62]. It has been implemented as module of netDx to allow the classification of sparse patient's profiles. It combines network-based propagation, biological networks, pathways and graph-theory concepts to build and find predictive pathway-specific patient similarity networks characterizing the patient classes in comparison. The method has been integrated in netDx because proved to achieve satisfying classification performances, introduced a first idea related to how to exploit network-based propagation scores associated to patient's single features and included a preliminary version of the algorithm that will be used by our main software Simpati to identify signature pathway-specific patient similarity networks.
    - Pratic has been our first attempt to improve netDx. It succeeded on certain aspects but failed in others.
- Section 6.3 describes the main software called Simpati [63]. It is a second-generation machine learning patient classifier based on patient similarity networks. It evolves netDx, integrates concepts of Pratic, tackles issues discovered implementing the previous two classifiers and is the first native pathway-based classifier.
    - Simpati abandoned the idea of netDx to integrate multiple omics for the classification and has been implemented to perform uniquely the task of pathway analysis.

6.1 netDx

6.1.1 Overview

netDx [61,62] is a supervised patient classification framework based on patient similarity networks. Each patient datum (e.g., gene expression profiles, age) is represented as a patient similarity network (PSN). A node is patient, and an edge between two patients corresponds to pairwise similarity for a given datum. This means that patients of unknown class can be classified based on their similarity to patients with known class. It applies the idea of recommender systems, similar to those used in Amazon or Netflix ("find movies like this one"), to precision medicine ("find patients who don't respond to therapy"). This process is clinically intuitive because it is analogous to clinical diagnosis, which often involves a physician relating a patient to a mental database of similar patients they have seen.

Given a binary classification problem as to correctly predict patients belonging to two classes (i.e., worse prognosis, better prognosis), netDx takes in input the available data of the patients and uses them to build PSNs. Next, it removes low similarities with a sparsification step and learns the similarity networks which are potentially more useful to correctly predict the class of unknown patients in a cross-validation setting. For the prediction, netDx applies GeneMANIA [64] on the sparsified networks. The external method scores each PSN using a query-driven regression. The query can be interpreted in the following way: "how well the similarity network is able to predict correctly the class of known patients belonging to the training set?". As result of this step, netDx keeps only PSNs who got a high GeneMANIA score, while it discards the rest.

netDx enters in the prediction phase. It computes the similarity between the testing patients and the training ones such that they can be all represented as nodes in the kept PSNs and applies GeneMANIA again. The latter combines the networks to produce one consensus integrated network and it runs a network-based propagation to predict the class of the testing patients. The class of the training patients is used as prior information mapped to their corresponding nodes and propagated. A testing patient is assigned to the training class from which received more information.

As result, netDx returns the best PSNs, the predicted classes of the testing patients and classification performance measures. For example, given two classes of patients with and without lung cancer, netDx could detect the patient similarity network built with the number of cigarettes smoked by a person. The pairwise similarity measure could be the difference between the number of cigarettes smoked by two patients. Close the difference and more two patients are similar. The final network could be relevant and predictive since is likely that the cigarette number is correlated to the risk of lung cancer and patients with the disease would have the highest numbers.

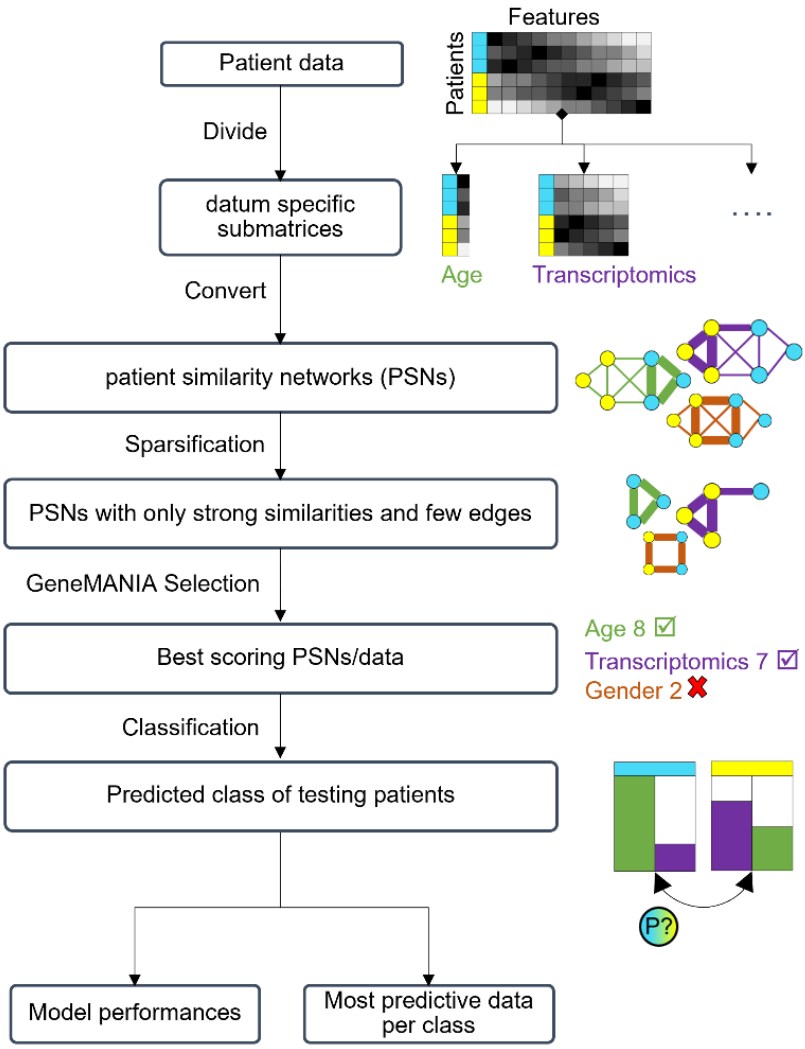


Fig. 6.1 netDx workflow showed as flow diagram. Patient classes are yellow and blue. The available data are divided based on their specific information type. For each datum, netDx generates a patient similarity network with only training patients as nodes. The edges have different thickness based on

the weight associated to them which is equal to the pairwise patient similarity. More two patients are similar due to a specific datum and more their connection and edge is thick. netDx then removes low similarities and keeps only the strongest ones for each node. The PSNs are scored and ranked based on their topology to separate the patient classes in study. Higher the separability and higher the score. The best scoring PNSs are used to classify unknown patients. netDx results include how many predictions were right and the data represented by the best PSNs.

6.1.2 Input Data Preparation

The first operation of netDx is the creation of the database of patient similarity networks from the available data. A PSN can be generated from any kind of datum, using a pairwise similarity measure. For example, how much two patients are similar based on their gene expression profiles can be measured using Pearson correlation, while patient age similarity can be measured with the difference. A reasonable design is to define one similarity network per datum, such as a single network based on correlating the expression of all genes in the transcriptome, or a single network based on similarity of responses to a clinical questionnaire. If a datum is multivariate, defining a network for each individual variable will result in more interpretable output. However, this approach may lead to too many networks generated (e.g., one network for each gene in a transcriptomics dataset), which increases computational resource requirements and risk of overfitting. Thus, as with any machine-learning task, there is a trade-off between interpretability and overfitting/scalability. To help address this problem for at least gene-oriented data (e.g., transcriptomics), it is possible to group genes into pathways (e.g., one network for each cellular function and set of genes), which it is assumed that they capture relevant aspects of cellular and physiological processes underlying disease and normal phenotypes. At this point of the software development, netDx introduced the support to the pathway analysis. If the patients have gene-based data, the user defines pathways as gene sets and a pairwise similarity measure to compute between patient's pathway-specific gene expression profiles, then netDx can build a PSN per cell function and continue with the standard workflow.

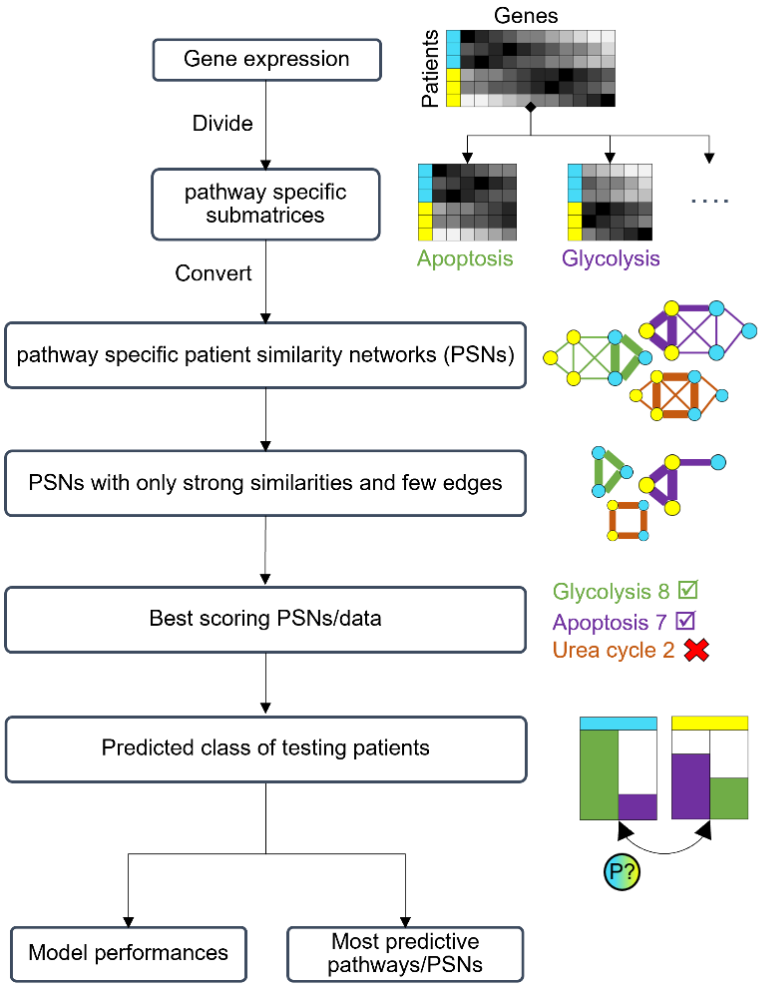


Fig. 6.2. netDx workflow showed as flow diagram when input patient datum is a gene expression dataset, and the user provides pathways to group genes. The dataset is divided in submatrices. Each submatrix contains only the genes included in a pathway. netDx generates a PSN per pathway.

6.1.3 Network sparsification

netDx performs an operation of sparsification for reducing the number of edges in the PSNs, speeding up the downstream workflow and reducing noise. The idea is to remove edges with low weight in the networks such that the only connected patients are the ones with strong competitive similarities. netDx sparsification strategy is based on three parameters: The first is a minimum edge weight to filter out all similarities lower than this "*cutoff*" value. The second parameter is the number of maximum interactions allowed per patient. While the third parameter is an upper bound about the number of edges of the overall network. By default, netDx sparsifies networks by excluding similarities below 0.3 and keeping the top 50

edges per node. However, these are all hyper-parameters that the user can tune to achieve the best classification performances for a specific dataset and set of patients.

6.1.4 GeneMANIA Selection

The sparsified networks in the database describe how patients are similar based on the available data. They are crucial for the classification performances, and they actual represent the main result of netDx. For this reason, the software filters the networks to keep only the most predictive and informative ones. netDx uses GeneMANIA algorithm to rank the PSNs based on their ability to characterize the patient classes.

GeneMANIA is applied on one sparsified PSN at a time; it selects a set of patients belonging to the same class ("+"), solves a constrained regression problem on the network to maximize the values of similarities between the patients (+ +) and minimize those between outsider members (+ -). The ideal network connects all patients of the same class without any connection to the other classes and is enriched positively (+ +). While the network that connects the selected patients to other classes, without connecting any selected patient is enriched negatively. netDx iteratively applies GeneMANIA on the same network changing the starting reference patients in every run. Plus, the iterations are divided in two. In one set of iterations, only the members of a specific class are considered. At the end, a PSN gets a score equal to the number of iterations in which it resulted enriched positively with a specific patient class and netDx gains two databases, each with only the best PSNs for a patient phenotype.

6.1.5 Classification

netDx integrates the testing unknown patients in the PSNs. It computes how much the patients are similar to the training ones and adds them as nodes to the selected best networks. It then runs GeneMANIA on each class-specific database. GeneMANIA selects the training patients of the database's class and combines the PSNs by averaging their similarity scores to produce a single integrated network. Then, it runs a network-based propagation on the integrated network outward from the training patients ("+" nodes) to rank all other nodes in the network by

connectivity (i.e., similarity). This is equivalent to the query: given predictive networks for the class of lung cancer, rank all unknown patients by similarity to lung cancer patients". netDx then assigns testing patients to the class for which they have the highest similarity.

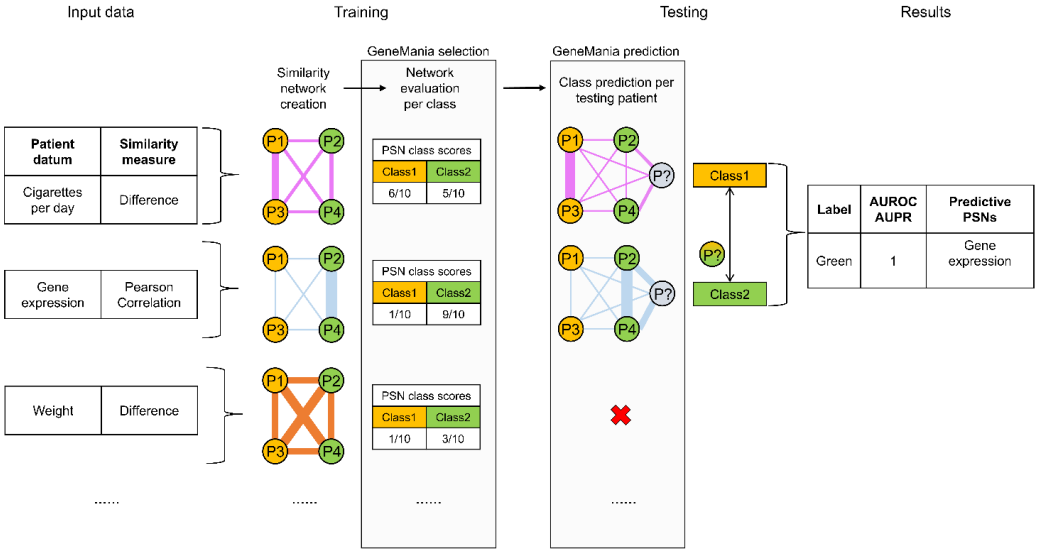


Fig. 6.3. Flow of the data taken in input, converted in PSNs, and elaborated by GeneMANIA. Let us consider the classification of orange versus green class patients. netDx builds a PSN for each datum. It needs a user-defined pairwise similarity to measure how much two patients are similar. netDx suggests the difference for numerical clinical data and Pearson correlation for biological omics. For each class, GeneMANIA scores the PSN based on its topology. A PSN is scored positively based on how much separates the patient classes and based how much one class is cohesive due to strong intraclass similarities. The blue PSN which represents how much patients are similar due to the correlation computed between every two patient's gene expression profiles is predictive of the green class. In fact, the patients P2 and P4 are much more similar than any other pair. While the PSN which measures how much patients are similar based on their weights is not predictive due to the fact that all patients are similar to each other. The PSNs which got a high positive score pass the selection and are used to classify an unknown testing patient. The latter in grey color is integrated in the PSNs and predicted based on its similarities with the training known patients.

6.1.6 Cross-Validation Setting

The method tests its ability to learn predictive patient similarity networks able to classify and characterize the patient classes with a cross-validation approach. It performs at least 10 iterations of the workflow. In every iteration, it randomly divides the patients in the training and testing sets. It creates, sparsifies and selects

the best networks with only the training patients. It then adds and classifies the testing patients.

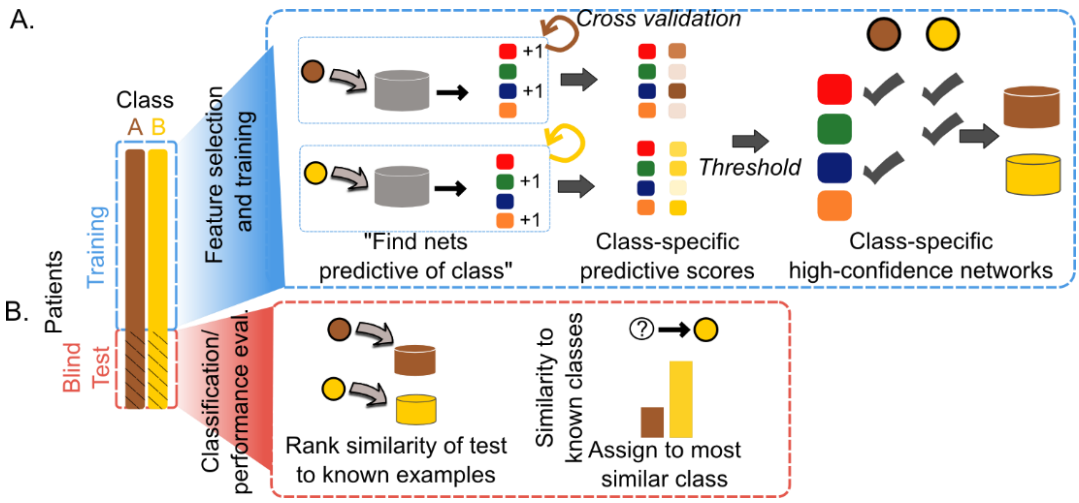


Fig. 6.4. Details of the netDx feature selection and patient classification methods. A. Machine learning is used to identify networks predictive of each patient class. Data are split into training and test samples, and feature selection uses only training samples. Cross validation is used to score how frequently a network is predictive of a given class (e.g., high-risk). This step results in network scores, with higher values indicating more predictive networks. These scores can be thresholded to identify a set of high-confidence networks for each class of interest (yellow and brown cylinders). B. Blind test patients are ranked by similarity to known examples from the training set. For this step, only class-specific feature-selected networks are used. Patients are assigned to the class to which they have highest similarity.

6.1.7 Results

netDx provides the networks used for the classification, their meaning (e.g. cigarette count) and the predicted classes for all testing patients. Finally, it also assesses its classification performances by comparing the predicted classes with the real ones and obtaining the measures as the area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPR), and accuracy. The AUROC compares the true positive rate against the false positive rate, while the AUPR relates the precision against the recall.

An example of netDx application can be found at chapter 8 section “netDx for lymphoma study”

6.2 Pratic

6.2.1 Overview

Somatic mutations are mutations which specifically occur in any cell except in germ cells. They can appear by environmental factors, but they are also provoked by a disease which wants to lead and govern the patient's cells for its own benefit. For their nature, the somatic mutations of a patient are represented in a sparse binary profile (i.e., vector). Out of at least twenty thousand genes, only few of them are indicated as mutated and include a value equal to one, while the rest has zero information. A data type which is difficult to use and process on the contrary of a gene expression profile which has all genes characterized by a quantitative value measuring the single molecule activity in the patient's cell. Further on, a disease can deregulate a cell function for its own benefit manifesting different and multiple mutations. This aspect combined with the sparsity of the datum makes patients described with somatic mutation profiles difficult to compare without ad-hoc strategies.

Patient's somatic mutations represented as ones in a vector (profile) full of zeros are both very few with respect the unaltered genes and rare in the population (i.e., it is difficult that two patients have one mutated gene in common). netDx cannot use standard similarity measures because requiring profiles with continuous numeric values. For example, netDx uses the Pearson correlation as default measure to compute the similarity between patients with gene expression profiles. In this scenario, netDx builds binary similarity networks representing cellular functions, if two patients have a mutation in the same pathway then they have a similarity (i.e., edge) with a weight equal to one in the related PSN, while if they do not share a mutation, they are not connected at all. Grouping genes is the approach that netDx applies to find overlaps and similarities between patient's mutation profiles.

At the same time, a competitor of netDx called NetNorm [65] introduced a novel approach to classify patients with sparse somatic mutation profiles. It applies a network-based propagation algorithm with the patient's mutations to recognise new genes as altered. After the propagation, the patient's profile includes a value equal to one for both the original mutated genes and the inferred ones. It is less sparse,

and it can be more easily compared to the other profiles that are processed in the same manner. In the end, NetNorm classifies the patients with their propagated profiles and a support vector machine.

We have developed and integrated a method in netDx to allow the usage of sparse somatic mutations for classifying patients, it replaced the original binary method of netDx and has been designed to use the network-based propagation as proposed by NetNorm. This method is called Pratic [62] and can be described as follows.

6.2.2 Input Data Preparation

Pratic transforms the binary matrix of somatic mutation patient's profiles (genes x patients such that a gene has 1 if mutated in a patient, while zero otherwise) into a continuous numeric matrix thanks to a propagation performed using a biological network of gene-gene interactions. The information and the values equal to one indicating mutated genes of a patient are mapped into the corresponding nodes of the gene-gene interaction network and propagated. Each node and gene, also if not originally mutated, gets a propagation score which reflects its starting a priori information and its proximity to mutated genes. The score replaces the gene's value in the patient's profile. The patient's propagated profiles are then used to create patient similarity networks representing how patients are similar in pathways.

6.2.3 Pathway-specific Patient Similarity Networks

For creating a PSN, Pratic considers the patient's propagated profiles with only the genes belonging to a specific pathway. It measures how much every pair of patients is similar based on the Weighted Jaccard (WJ) [66]. The latter between two vectors with non-negative entries $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$, is defined as

$$WJ(\mathbf{x}, \mathbf{y}) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)}$$

In other words, WJ evaluates how much two pathway-specific profiles are similar comparing how much the values of the same element (aka gene) are close. Correlation can be high also if two patient's profiles have very different values in the same genes. This property is beneficial when the correlation measures how much two gene expression profiles are similar because it is unlikely that the same

gene is expressed with the same value in two patients. However, the network-based propagation is a normalization and standardization technique. The same gene is likely to have a similar propagation value in two patients if it has the same importance. The Weighted Jaccard is able to capture this aspect.

6.2.4 Network Filtering

Pratic imposes netDx to build a PSN per pathway and, compared to the standard workflow in which netDx creates a PSN per datum (e.g., biological omic), this approach leads to have as many PSNs as the annotated pathways in literature (from 2000 to 50000). For reducing the computational requirements due to so many networks, diminishing the running time and improving the signal-to-noise ratio, Pratic has been designed to perform a filtering of the PSNs based on a non-parametric statistics approach.

For the task, Pratic builds a synthetic PSN which would not allow to characterize the two patient classes in comparison due to strong interclass similarities and compares it to the real similarity networks. More the PSNs are different to the negative model and more they can potentially distinguish the patient classes based on the similarities between individuals and be predictive. The best PSNs are then selected and provided to the standard framework of netDx which proceeds with sparsification, GeneMania selection and prediction of testing patients.

Entering more into the details of the single operations, Pratic discards a PSN if the two classes mix, so if the average of the inter-similarities (between members of different classes) is higher than the average of the intra-similarities of each class. In fact, the network cannot be biologically predictive and characteristic of a patient class if members with different phenotypes are more similar than the ones that share the same. Next, the remained PSNs are assigned to one patient's condition based on their topology. A network is associated to the class in which the patients are the most similar. Pratic obtains two sets of PSNs and in each one the members of one specific class are the most cohesive and similar. As last, the approach requires to keep only the networks that in each set maximize the dissimilarity between the most cohesive class and the opposite group. Pratic applies two filters.

For the first one, it determines the average of the intra-similarities of the cohesive class in all the PSNs and computes the average of averages (aka set's average). It computes how much a single PSN's average is different with respect the set's average. Higher is the difference and more a PSN is considered predictive and characteristic of the cohesive class. This filter is based on the assumption that at least half of the candidate predictive PSNs for a class are false positive. While the remained half of the PSNs with a positive difference, because with an average greater than the one representing the set, are indicating true positive signature pathways. The latter are kept and pass to the second filter. Let us define a node of interest called root, the friends of the root are patients belonging to the same root's class, while strangers are the patients belonging to the opposite class. A stranger can be further on labelled as "rival" or "neutral". It is a rival if the root is more similar to it than to its friends. It is neutral if the root is more similar to its friends. Pratic determines the number of rivals and neutrals for every node in the PSNs and selects the networks satisfying two criteria. A PSN is kept if the members of the cohesive class have more neutrals than rivals. A PSN is kept if the members of the cohesive class have more neutrals than the average of neutrals possessed by the same members in the rest of the other PSNs in the set. The PSNs that remained untouched by the filtering are then passed to netDx for the standard workflow.

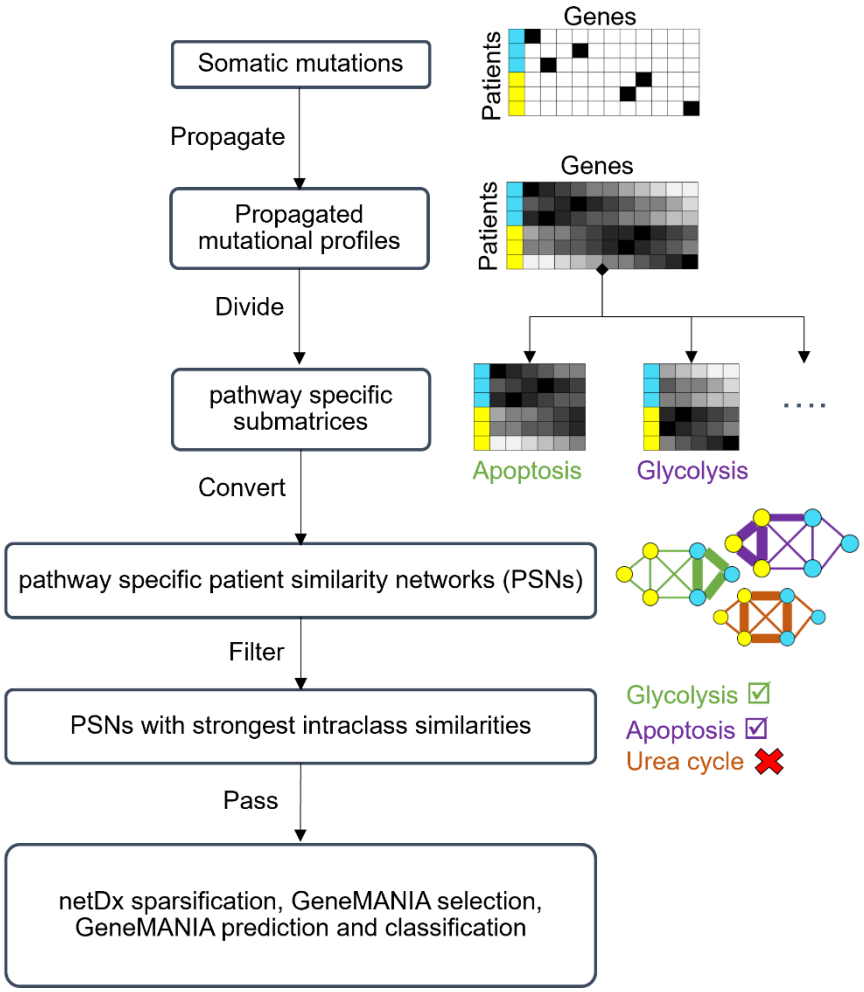


Fig.6.5 Workflow of Pratic that processes patient's somatic mutation profiles for netDx. A patient belongs to one class (yellow or blue) and is described by a somatic mutation profile. A black square indicates a mutated gene in a patient. Mutations between patients are not corresponding, not even the ones between members of the same class. After the propagation, the binary values (1 mutated and 0 non mutated) are replaced with the propagation scores associated to the genes. The matrix of somatic mutation profiles is divided in submatrices. Each submatrix contains only the genes belonging to a specific pathway. Pratic measures with the Weighted Jaccard how much each pair of patients is similar in a pathway and creates the relative PSN. The networks undergo the Pratic selection based on their topology and only the best ones continue the standard workflow of netDx.

6.2.5 Performance Evaluation

We evaluated the performances of Pratic integrated in netDx to classify seven cancer datasets collected from the Cancer Genome Atlas in which patients (ranging from 169 to 430) were described by somatic mutations and divided into two classes based on their survival (Deceased or Alive) at the end of the follow up.

Pan-cancer tumor datasets by NetNorm	**Dimension (genes x patients)**	**Mutated genes per patient**	**Percentage of mutations in the dataset**	**Classes Deceased vs Alive**	**Label**
Lung Adenocarcinoma	13589 x 169	0,02%	1,9%	70 vs 99	Lung Adeno
Ovarian carcinoma	10192 x 363	0,005%	0,5%	172 vs 191	Ovarian
Skin melanoma	17527 x 307	0,04%	2,6%	129 vs 178	Skin
Kidney carcinoma	10608 x 411	0,005%	0,5%	136 vs 275	Kidney
Head and Neck squamous cell Carcinoma	17022 x 388	0.009%	0.92%	140 vs 248	Head Neck
Glioblastoma multiform	14748 x 265	0.009%	0.90%	195 vs 70	Brain

First, we determined the performances of Pratic with the Weighted Jaccard and with the Pearson Correlation. The method for this comparison has been applied without the filtering of the PSNs.

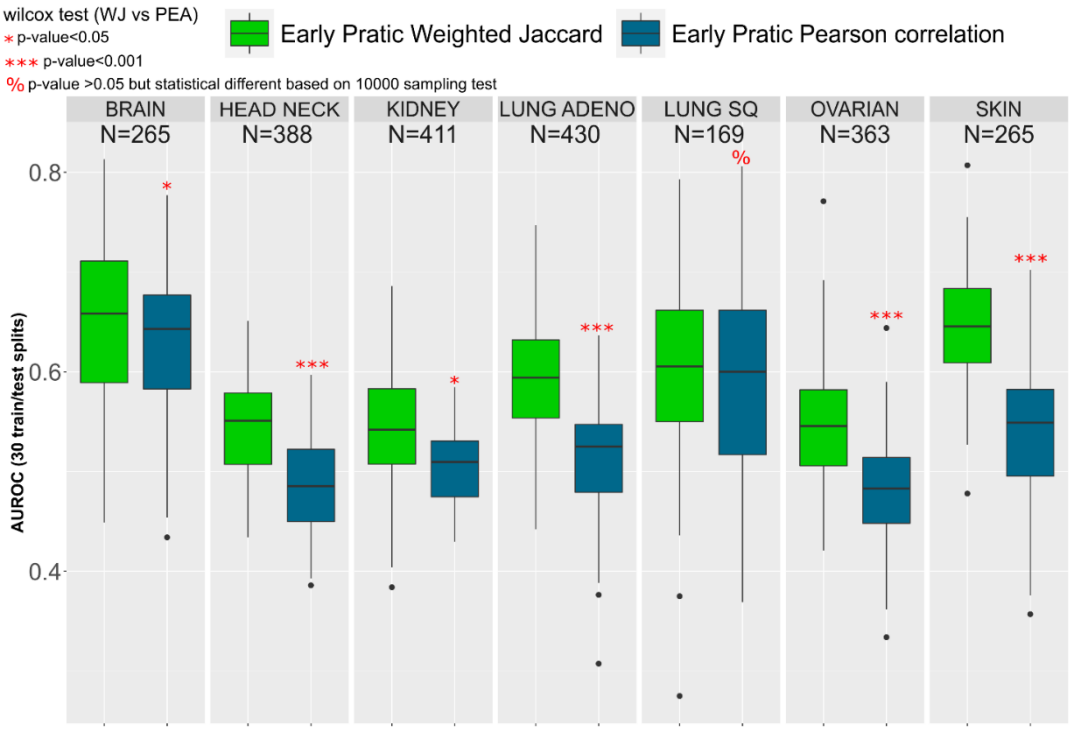


Fig.6.6. Boxplot of AUROC measurements obtained comparing the predictions made by Pratic with the true classes of the patients included in each dataset. In one case, Pratic used the Weighted Jaccard to measure the pairwise patient similarity, while in the second case it used the Pearson correlation. Each panel shows the results for one tumor type, and each boxplot contains the 30 AUROC measurements obtained from 30 runs of cross-validation. Boxplot center indicates median; box bounds indicate 25th and 75th percentile, and whiskers mark 1.5 times the interquartile range. Dots indicate the outliers measurements. Statistical significance is computed using a one-sided Wilcoxon-Mann-Whitney comparing Pratic with Weighted Jaccard and Pearson correlation; in case, there is no statistical significance, a preliminary Bayesian approach is applied to determine which method is better. The label "Early" before Pratic name indicates that the test used the method without its ad-hoc filtering of the similarity networks.

The Weighted Jaccard provided the best classification performances in all the datasets and has been chosen as default patient similarity measure in Pratic.

Second, we compared the performances of Pratic with and without the ad-hoc filtering implemented for improving the signal to noise ratio, the running time and reducing the computational resources.

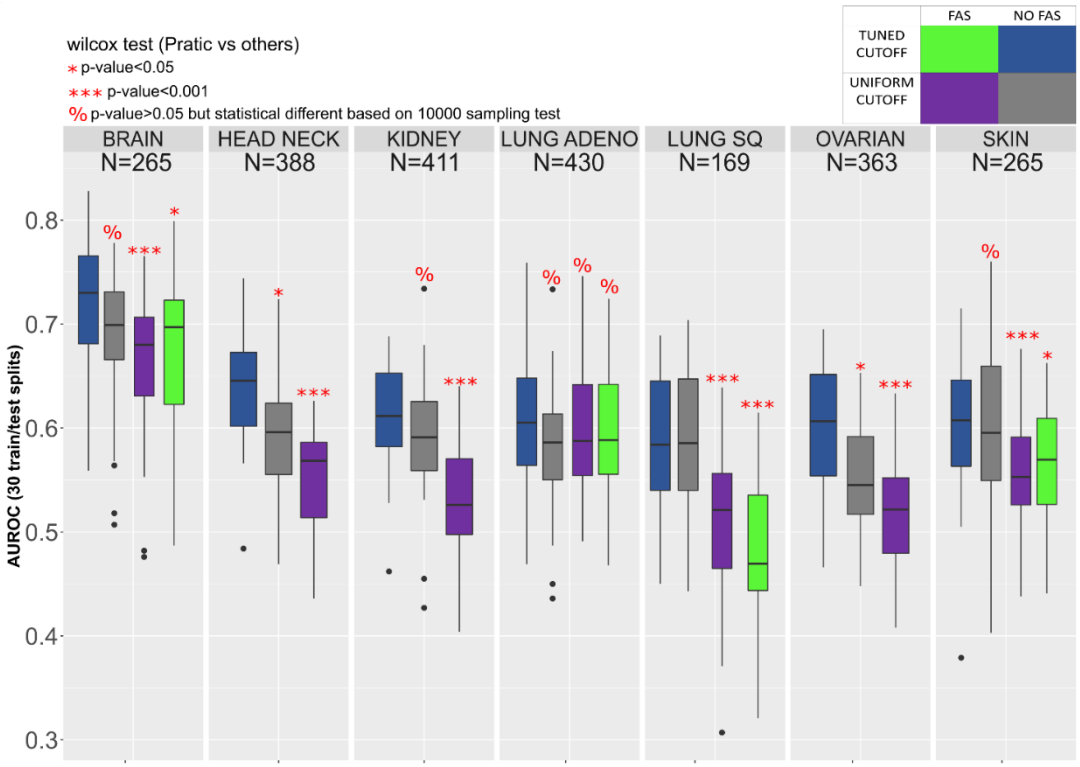


Fig.6.7. Boxplot of AUROC measurements obtained comparing the predictions made by Pratic with the true classes of the patients included in each dataset. Each color of box is associated to a specific Pratic setting. Green indicates Pratic applied with the filtering of the PSNs and with a tuned value for the *cutoff* parameter of netDx in the sparsification step to achieve the best performances. Blue indicates Pratic applied without the filtering of the PSNs. Purple indicates Pratic applied with the filtering and with the standard value of 0.3 for the netDx cutoff. Each panel shows the results for one tumor type, and each boxplot contains the 30 AUROC measurements obtained from 30 runs of cross-validation. Boxplot center indicates median; box bounds indicate 25th and 75th percentile, and whiskers mark 1.5 times the interquartile range. Dots indicate the outlier measurements. Statistical significance is computed using a one-sided Wilcoxon-Mann-Whitney comparing Pratic settings; in case, there is no statistical significance, a preliminary Bayesian approach is applied to determine which method is better.

We found out that the ad-hoc filtering implemented in Pratic gave the worst results. In some case, it was not even allowing the classification of the patients like in the Kidney dataset. The cause may be due to the double filtering. Pratic filters the PSNs based on topological criteria, then passes the best networks to netDx which filters them again based on GeneMANIA criteria. It is likely that the PSNs considered good from Pratic are not considered good from GeneMANIA. This is the reason why we obtained the best results without the ad-hoc filtering and with the tuned value for the *cutoff*.

Finally, we applied Pratic without the ad-hoc filtering to classify the patient classes in each dataset. We applied also the standard netDx and NetNorM. netDx built the binary PSNs for each pathway, while NetNorm used the propagation and then a classification based on random forest. We got the classification performances of all the methods and put it in comparison.

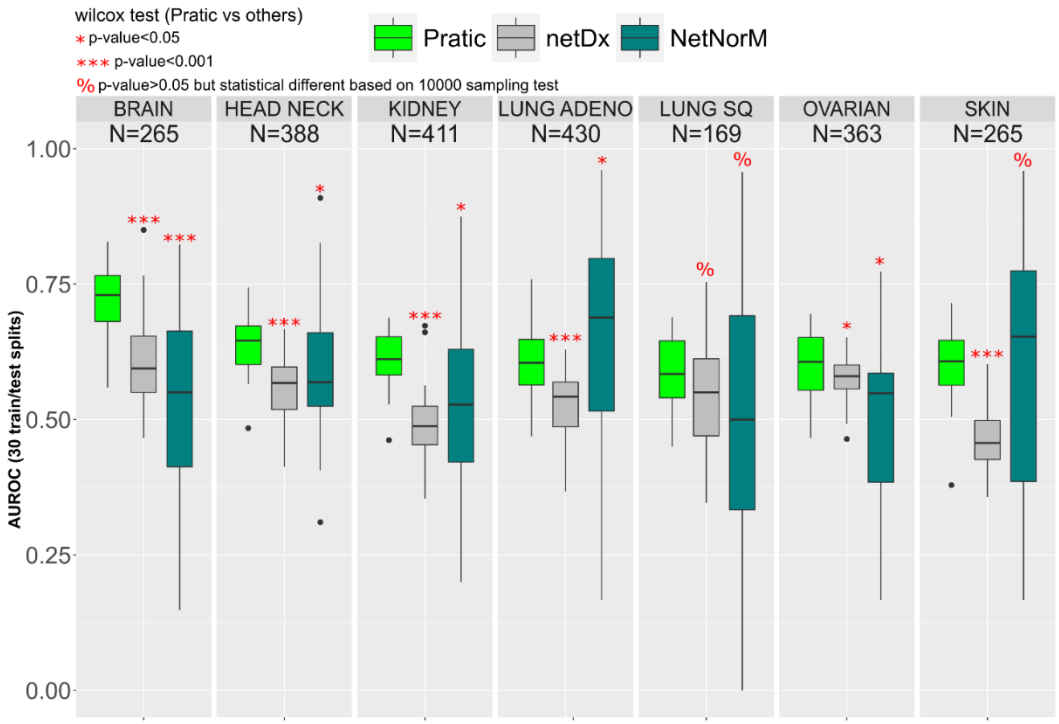


Fig.6.8 Boxplot of AUROC measurements obtained comparing the predictions made by Pratic, netDx and NetNorm with the true classes of the patients included in the cancer datasets. Each panel shows the results for one tumor type, and each boxplot contains the 30 AUROC measurements obtained from 30 runs of cross-validation. Boxplot center indicates median; box bounds indicate 25th and 75th percentile, and whiskers mark 1.5 times the interquartile range. Dots indicate the outlier measurements. Statistical significance is computed using a one-sided Wilcoxon-Mann-Whitney comparing the performances of the different classifiers; in case, there is no statistical significance, a preliminary Bayesian approach is applied to determine which method is better.

Pratic outperformed the competitors in all the datasets, and it has been considered for being integrated in the final version of netDx software.

6.2.6 Pathway Analysis

The classification performances evaluate how much the machine learning algorithm is able to predict correctly the classes of the testing patients and to generalize based on the information learnt with the training patients. However, they do not allow to understand the quality of the resulting biological pathways which are represented

by the best patient similarity networks used by the algorithm to classify. Obtaining valid classification performances without using and providing meaningful cellular functions potentially associated to the patient's diseases and phenotypes would make all the process useless, counter-productive, and dangerous in case of a real application of the findings as biomarkers in the clinical diagnosis. For example, if Pratic would predict the class of deceased lung cancer patients using a generic metabolic pathway as the glycolysis, then the classifier would not be useful to better characterize neither the subjects nor their disease.

We tested the pathways found with Pratic in the classification of Glioblastoma patients. We compared our findings with annotated pathways describing the two phenotypes based on wet-lab studies which have been previously published in the scientific literature.

Precisely, the task has been performed with the following operations. We created a network such that the nodes are the Pratic resulting pathways predictive of a specific class (e.g., deceased Glioblastoma patients) and the literature pathways representing the ground truth of the patients. We then measured the overlap between two pathways with the Jaccard similarity (sum of the shared genes divided by the sum of the size of the two biological processes). We connected the nodes having an overlap with a weighted edge based on their similarity. We performed an unsupervised clustering with the Markov Clustering method (MC) [67] , a fast and scalable algorithm for graphs based on the simulation of a flow similar to the concept behind the network-based propagation that works with both weighted and unweighted graphs. In the paradigm of this algorithm, a cluster is characterized by nodes that are strongly connected with each other and that have few interactions with outsiders. After the detection of the clusters, we arranged the nodes so that pathways belonging to the same cluster were placed close together and we annotated each group. We gave to each cluster a label which summarized the pathways inside based on the WordCloud Adjacent Words method [68]; we selected the words appearing with the highest frequency in the names of the pathways and the words adjacent to them. At the end of these steps, the annotated and clustered enrichment map allowed us to determine how much Pratic findings were dissimilar from the ground truth. We determined the percentage of clusters that shared both

Pratic pathways and literature pathways, we analysed those pathways that appeared isolated in the map indicating possible artifacts of predictive PSNs and we proved the quality of Pratic results.

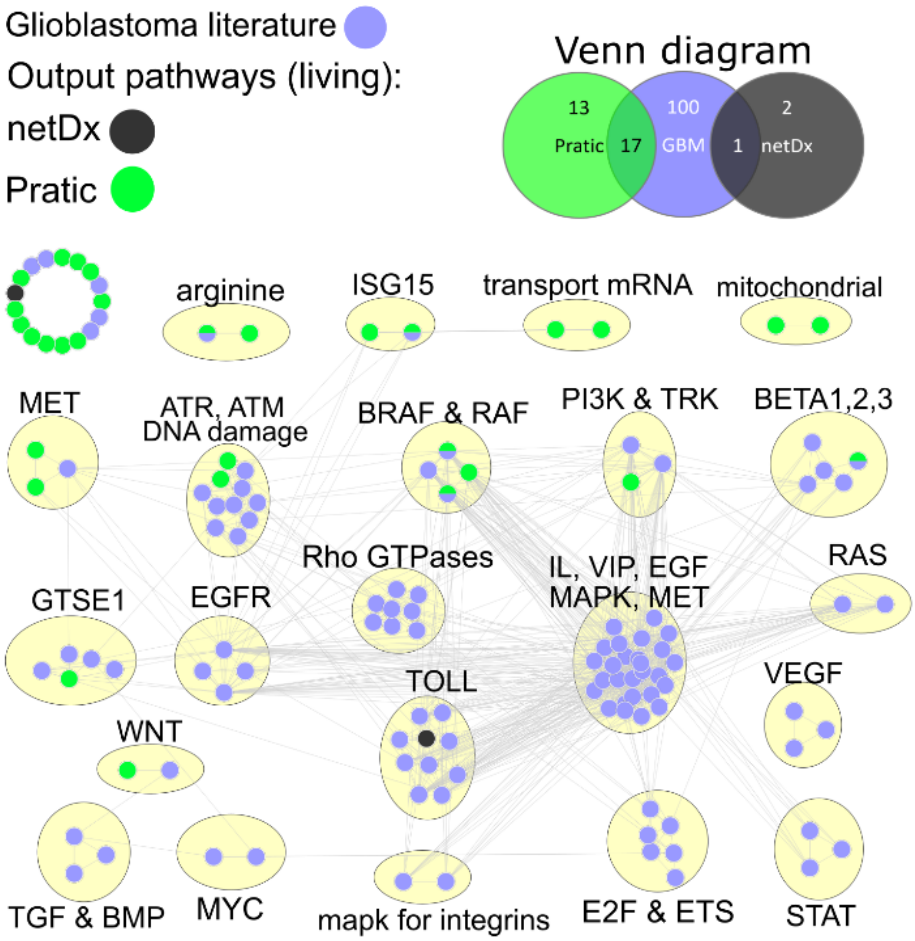


Fig. 6.9 Enrichment map of the pathways that Pratic (bright green) and netDx (black) provided as associated to the best and most predictive PSNs used in predicting the Glioblastoma patient classes. Nodes are pathways, and edges indicate shared genes. The overlap between two pathways is measured with the Jaccard similarity. Nodes are arranged so that highly similar pathways are placed close together. The pathways represented by purple nodes compose the ground truth related to the Glioblastoma; they have been extracted by the literature. The Venn diagram shows how many method-specific pathways are actually related to the disease of the patients.

The enrichment map shows that Pratic detects more disease related pathways than netDx and is a valid classifier.

6.3 Simpati

6.3.1 Overview

Pathway-based patient classification is a supervised learning task which supports the decision-making process of human experts in biomedical applications providing signature pathways associated to a patient class characterized by a specific clinical outcome/phenotype/condition. The task can potentially imply to simulate the human way of thinking in predicting outcomes by pathways, decipher hidden multivariate relationships between the patient's characteristics and their class, and provide more information than a probability value. However, classifiers are rarely integrated into a routine bioinformatics analysis of high-dimensional biological data because they require a nontrivial hyper-parameter tuning, are difficult to interpret and lack in providing new insights. There is the need of new classifiers which can provide novel perspectives about pathways, be easy to apply with different biological omics and produce new data enabling a further analysis of the patients.

We propose Simpati, a pathway-based patient classifier which combines the concepts of network-based propagation, patient similarity network, cohesive subgroup detection and pathway enrichment. It exploits a propagation algorithm to classify both dense, sparse, and non-homogenous data. It organizes patient's biological features in pathways represented by patient similarity networks for being interpretable, handling missing data and preserving the patient privacy. A network represents patients as nodes and a novel similarity determines how much every pair act co-ordinately in a pathway.

Simpati detects signature biological processes based on how much the topological properties of the related networks separate the patient classes. In this step, it includes a novel cohesive subgroup detection algorithm to handle patients not showing the same pathway activity as the other class members. An unknown patient is classified based on how much is similar with known ones. Simpati outperforms state-of-art classifiers on five cancer datasets described with two biological omics, classifies well sparse data, identifies more relevant pathways associated to the patient's diseases than the competitors and has the lowest computational requirements between pathway-based classifiers.

Simpati can serve as generic-purpose pathway-based classifier of patient classes. It learns signature pathways that are enriched in a class with respect the reference. It tests the pathways based on how much they can correctly predict the class of unknown patients. It detects signature pathways divided into up or downinvolved based on how much the members of one class are similar. Upinvolved if the activity of the molecules in a pathway is high and similar between the patients considering the signaling cascades derived from the activity of also all the other molecules. The signature pathways are then related to the patients in study with the integration of the Disgnet and Human Protein Atlas databases. Together with the results of a unique pathway analysis, Simpati provides in a vectorial and graphical format the patient similarity networks associated to the biological processes. This allows to find the patient which is core or outlier with respect the other members due to specific molecules and pathways. We provide an R implementation which starts Simpati with one function, a GUI interface for the navigation of the patient's propagated profiles and a function which offers an ad-hoc visualization of patient similarity networks. The software is available at: https://github.com/LucaGiudice/Simpati

6.3.2 Problem

High-throughput biological data provide valuable information to clinicians for the prognosis and treatment response of patients. They offer quantitative and qualitative evidences to biomedical scientists for developing a study or confirming wet-lab results [69–71]. Pathway-based analysis is a technique for mining these data. It provides an intuitive and comprehensive understanding of the molecular mechanisms related to the patients [72,73]. The pathway space is more robust to noise than the single feature level, summarizes the information of multiple patient's features into the pathway activity (inhibited or activated), reduces the model complexity and maintains predictive accuracy in the face of uncontrolled variation [74–77]. These motivations boosted the development of enrichment tools for the pathway analysis but not of machine learning algorithms. The latter are neither considered in reviews [42,56,78,79] nor in bioinformatics best practises [35–37]. Fabris et al. [38] detailed the drawbacks of a supervised classification approach. It lacks a formal statistical basis, is computationally expensive, includes a not trivial

hyper-parameter setting and does not well handle neither imbalances classes nor structured feature types as the biological pathways.

Few attempts have been made in this direction. In 2010, Pang et al. [80] proposed a bivariate node-splitting random forest integrating pathways for the survival analysis of cancer patients in microarray studies. In 2018, Hao et al. [81] proposed the first generic-purpose pathway-based deep neural network for the prediction of Glioblastoma patients. The method builds a network model by leveraging prior biological knowledge of pathway databases and predicts considering hierarchical nonlinear relationships between the biological processes and the patient classes in comparison. However, the method requires a non-trivial tuning of hyper-parameters which are difficult to interpret for bioinformaticians or computational biologists without a background in deep learning, demands high computational resources, does not provide a graphical representation to explain why specific pathways have been selected and includes in the results only the classification performances.

In the same year, Pai et al. officially introduces the emerging patient similarity network (PSN) paradigm [82] for the precision medicine. In a PSN, each node is an individual patient and an edge between two patients corresponds to pairwise similarity for a given datum (gender, height, gene expression …). The paradigm brings many advantages. Analysing the similarities to gather new information is conceptually intuitive. A PSN can lead to the identification of patient subgroups or the prediction of a patient's class. Similarity networks can represent any datum, naturally handle missing and heterogenous data, have a history of successes in gene and protein function prediction [64,83,84] and can preserve the patient privacy by being shared in place of the sensitive raw information (topic which is growing in concern) [85–87].

Pai and Giudice et al. propose netDx [61,62] as patient classifier based on the PSN paradigm. Any available datum (e.g., age, gender, gene expression, …) is converted into a PSN. The method proceeds by performing GeneMania on the similarity networks. GeneMANIA [64] scores each input PSN based on how well it classifies an input set of patients known to be in the same class (i.e., training set). A linear combination of the best PSNs is used to create a composite network on which the

unknown patients (i.e., testing patients) are classified based on their similarity with the training ones. Despite researchers' efforts for standing up to the challenge, the method does not accept data in matrix format, requires to define multiple functions in order to set up the model, does not provide a graphical representation of the PSNs used to predict, does not give access to the data processed during the workflow, depends by the quality of the user-selected similarity measure, requires multiple hyper-parameters to manually tune, demands high computational resources and includes in the results only the classification performances together with the pathway names.

There is the need of new classifiers able to get the benefits of both the methodologies: classification and enrichment. As defined by Fabris et al., from the enrichment side the new method should be computationally light, easy to understand, provide more information than the probability value and not requiring assumptions to satisfy. While, from the classification side, it should be non-parametric, interpretable, consider multivariate interactions between features and patient classes, not requiring hyper-parameters difficult to set by the final user and cope well with both high-class imbalance and structured feature types.

We want to stand up to the challenge by proposing a pathway-based classifier called Simpati. Our main contributions with this method include: 1) the combination of graph-theory concepts as network propagation, cohesive subgroup detection and graph topology with machine learning to handle different biological omics, unbalanced patient classes and outliers, 2) a novel patient similarity measure adapt for any biological data type, 3) a novel concept of enriched pathways, 4) a user-friendly implementation, 5) an integrated prioritization system for the biomarker selection, 6) an integrated literature-based enrichment of the pathways, 7) explorable results which can lead to further findings, 8) two visualization tools for the patient analysis

6.3.3 Summary

Simpati considers the patient's biological profiles (e.g., genes per patients) divided into classes based on a clinical information (e.g., cases versus controls). It prepares the profiles singularly applying guilty-by-association approach to determine how

much each biological feature (e.g., gene) is associated and involved with the other ones and so to the overall profile. Higher is the guilty score and more the feature is involved in the patient's biology. Simpati proceeds by building a pathway-specific patient similarity network (psPSN). It determines how much each pair of patients is similarly involved in the pathway. If the members of one class are more similar (i.e., stronger intra-similarities) than the opposite patients and the two classes are not similar (i.e., weak inter-similarities), then Simpati recognizes the psPSN as signature. If the classes are likely to contain outlier patients (i.e., patients not showing the same pathway activity as the rest of the class), then Simpati performs a filtering to keep only the most representative members of each class and re-test the psPSN for being signature. Unknown patients are classified in the best pathways based on their similarities with known patients and on how much they fit in the representative subgroups of the classes (i.e., more they are similar to the representatives of a class and more they fit). As results, Simpati provides the classes of the unknown patients, the tested statistically significant signature pathways divided into up and down involved (new pathway activity paradigm based on similarity of guilty scores), the patient similarity networks in vectorial and graphical format to allow further analysis of the subjects in studies using graph-theory methods, the guilty scores associated to the biological single-level features and all the data produced during the workflow.

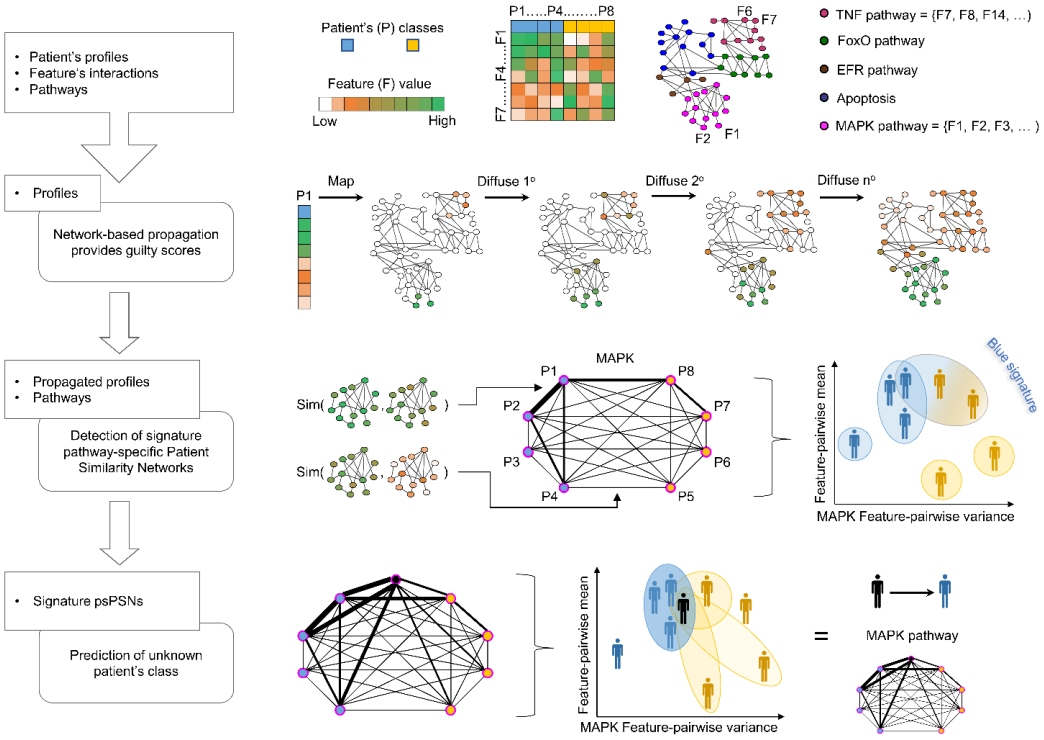


Fig.6.10. Workflow of Simpati. Patient profiles are divided in two classes and are described by biological features. A feature-feature interaction network together with pathways are further input data required by the software (e.g., gene-gene interaction network and KEGG pathways). All profiles are individually propagated over the network. The profile's values are replaced by scores that reflect the feature's starting information and interactions. Simpati proceeds by creating a patient similarity network for each pathway (psPSN). The pairwise similarity evaluates how much two patients have a similar pathway activity. It evaluates how much the features between two patients are close and high in term of propagation values. Two patients that act on a pathway with the same features and same expression values get the maximum similarity. In the figure, more an interaction is thick and more the two corresponding patients are similar. The psPSN is recognised as signature if one class is cohesive, one is sparse, and the two classes are not similar. In case of signature pathway, an unknown patient is classified based on how much is like the other patients.

6.3.4 Method

Input Data Acquisition

Simpati works with patient's biological profiles (e.g. gene expression profiles), the classes of the patients (e.g. cases and controls), a list of pathways and an interaction network (e.g. gene-gene interaction network). Simpati is designed to handle multiple biological omics but requires that the type of biological feature (e.g. gene) describing the patients is the same one that composes the pathways and the network which models how the features interact or are associated. In this study, we tested Simpati in the classification of Early versus Late cancer stage patients. In fact,

identifying the cancer mechanisms which drive the tumor from early to late stages is challenging [88–90] but it can improve the early cancer diagnosis, lead to develop more precise therapeutic strategies and increase the survival rates [91]. A late-stage cancer spreads to nearby lymph nodes and other organs, the survival rate decreases due to the necessity of more advanced and risky treatment strategies. While, early localized stages are easier to treat and have better survival rates [90,92–94]. For setting up this biological and pathway-based classification challenge, we collected data about Liver hepatocellular carcinoma (LIHC), Stomach adenocarcinoma (STAD), Kidney renal clear cell carcinoma (KIRC), Bladder Urothelial Carcinoma (BLCA), Lung squamous cell carcinoma (LUSC) and Esophageal carcinoma (ESCA) cancers from The Cancer Genome Atlas (TCGA) using the R packages curatedTCGAData [95] and TCGAutils [96]. We kept only the patients having RNA sequencing (RNAseq) data, somatic mutation data and the histological type as clinical information. We added a new information based on the pathological stage attribute. We applied a binarization and labelled the stage I and II in Early, while the stage III and IV in Late based on the tumor/node/metastasis (TNM) system [97–100]. We proceeded with preparing the biological omics. We followed the workflow defined by Law et al. [101] for the RNAseq. Genes not expressed at a biologically meaningful level have been filtered out to increase the reliability of the mean-variance relationship. We removed the differences between samples due to the depth of sequencing and normalised the data using the trimmed mean of M-values (TMM) [102] method. While somatic mutation data have been converted into a binary data type, where a value equal to one indicates a mutated gene in a patient and zero otherwise. We ended up with two biological omics for five datasets with 14 LIHC (7 Early, 7 Late), 21 STAD (8 Early, 13 Late), 37 KIRC (24 Early, 13 Late), 45 BLCA (8 Early, 37 Late), 75 LUSC (60 Early, 13 Late) and 152 ESCA (91 Early, 61 Late) patients. The first four datasets to simulate wet-lab routine studies and the last two to have more precise classification performances [103]. We then collected the pathways and created the biological interaction network. We retrieved the biological processes from the major databases MSigDB [46], GO [104] and Kegg [45], while we used Biogrid [105] to model the biological feature's

interactions. A node represents a gene, and the edges are experimental and manually curated gene-gene interactions (GGi) (564,325 interactions and 26,433 genes).

Formally, given a set of features $FEA = \{F_1, F_2, F_3, \dots, F_A\}$ with $A \in \mathbb{N}$ and of patient's profiles $PAT = \{P_1, P_2, P_3, \dots, P_B\}$ with $B \in \mathbb{N}$ where each element is a vector of feature's values, Simpati requires the concatenation of the patient's profiles in a matrix $M: AxB = (m_{a,b})$ where $m_{a,b}$ is equal to the value of F_a in patient P_b. The set of pathways $PATH = \{PH_1, PH_2, PH_3 \dots, PH_C\}$ where each element is defined by a finite set not exclusive of features (e.g., $PH_1 = \{F_4. F_5, F_6\}$). The biological network is defined as an undirected unweighted graph $GGi = (V, E)$ where $V = \{v_1, v_2, \dots, v_N\}$ is a set of vertices (aka nodes) representing features $f: FEA \rightarrow V$ and $E = \{(v_i, v_j) | v_i, v_j \in V\}$ is a set of pairs of vertices indicating edges (aka interactions, associations or relations) between features.

The adjacency matrix of GGi is the symmetric matrix $W: N \times N = (w_{i,j})$ defined by:

$$w_{i,j} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Finally, the patient's classes are two vectors containing the indexes of the profiles belonging to them:

$\overline{EARLY} = (1,2,3, \dots, D)$ and $\overline{LATE} = (D+1, \dots, B)$ with $D <= B$.

Input Data Preparation

Simpati starts with protecting the privacy of the patients and enhancing the advantage of using the patient similarity network paradigm in the workflow. It converts the patient's original information in anonymous labels. It then gives to the user the possibility to share data and results with or without the map. When the patient privacy is preserved, Simpati transforms the patient's biological profiles using a network-based propagation algorithm. Each feature gets a new value based on its a priori information (e.g., expression or mutation value) and by its associations with all the other features. In other words, it gets a new value based on how much is found "guilty" of being involved in the overall patient's biology and disease. This is based on three assumptions: the a priori information measures the

strength of the link between the feature and the patient (e.g. the expression measures the gene importance and activity), the feature's associations indicate shared molecular or phenotypic characteristics (e.g. interacting genes have similar cellular functions) [25] and a disease is rarely a consequence of an abnormality in a single feature but reflects the perturbations and signaling cascades provoked by multiple dysfunctional molecules [49].

In our application, Simpati maps the a priori values of the genes to their corresponding nodes in the GGi network. It propagates the values through the interactions using the propagation algorithm. Each node, even the one without value, gets a score which reflects its starting information and the amount given and received from its neighbours. The amount shared between nodes depends by the propagation type and the network topology. Simpati uses the random walk with restart (RWR) algorithm and the row-normalized version of the network. The RWR is a state-of-the-art network-based propagation algorithm [106] and flexible standardization technique [25] which has been successfully applied for the disease characterization [107] and the prioritization of multiple disease-associated biological features as genes [106], pathways [108], miRNAs [109,110], lncRNA [111], proteins [27] and somatic mutations [112]. While, the row normalization guarantees that a node gives the same amount of information equally to all its neighbours independently by their degree [113,114]. This allows to not favour specific nodes against others.

Simpati uses the propagation to get the same continuous numeric datum from any input profile, to include in a single feature's value also the perturbations provoked by the disease-specific activities of all the other patient's molecules and to boost the signal-to-noise ratio [25]. For example, a poorly expressed gene gets a high score if close to strongly expressed genes and a non-mutated gene gets a high score if close to a mutated one. As consequences, this step allows to handle different biological omics (e.g. dense gene expression data, sparse somatic mutation data) [115–118], to use a novel ad-hoc similarity measure and to not let the interpretation and meaning of Simpati's results depending by user-defined parameters.

For each profile $b \in \{1, \dots, B\}$, we define the set of its features represented as vertices with a priori information $SV_b = \{V_i \in V \mid \forall i \in \{1, \dots, N\}\, s.t.\ \exists\, F_i \rightarrow V_i\ and\ m_{i,b}\,! = 0\}$. The RWR algorithm measures the importance of each node v to SV_b. RWR mimics a walker that moves from a current node to a randomly selected adjacent node or goes back to source nodes with a back-probability $\gamma \in (0, 1)$. RWR is described as follows:

$$P_b^{t+1} = (1-\gamma)W' P_b^t + \gamma P_b^0 \quad (2)$$

where P^t is a $N \times 1$ probability vector at time point t of which the i_{th} element represents the probability of the walker being at node $v_i \in V$, P^0 is the $N \times 1$ initial probability vector and defined as follows:

$$P_b^0 = \{\tfrac{1}{|SV|}\ if\ v_i \in SV, 0\ otherwise\} \quad (3)$$

W' is the transition matrix of the graph GGi and $W'_{i,j}$ denotes a probability with which a walker at v_i moves to v_j. Formally, $W'_{i,j}$ is defined based on the row normalization:

$$W'_{i,j} = \frac{W_{i,i}}{\sum_i W_{i,j}} \quad (4)$$

The propagated profile P_b^{t+1} replaces the original profile P_b^0 in the matrix M.

Trending Matching Similarity Measure

Simpati works with patient's propagated biological profiles. The feature has a score which measures how much is "guilty" of being associated and involved in the patient's disease or generic clinical condition driving its biology with specific alterations and disfunctions. Higher the score and more the feature is involved. Plus, the propagation score is meaningful also between profiles. Lower the propagation score varies between patients and more the feature is assuming the same role. This point from a pathway perspective is important. Two patients may strongly involve one biological process but may act on it from different directions. For example, two patients may have high scores for the EGFR pathway genes (n=79) but have very different values for few ones as EGFR, JAK, IL-6 and GAB1. This may due to the fact that, the disease of one patient is acting on the pathway using exclusively EGFR

and JAK [119], while the other is using IL-6 and GAB1 [120,121]. By literature, the direction of regulation of a pathway is disease specific and a biomarker.

We wanted to capture both the aspects of the propagation score in developing the similarity measure to estimate how much two patients were similar, so we designed a novel pairwise similarity measure called Trending Matching (TM) similarity (0 lowest, 1 highest). It is the weighted sum of two components: the mean and the variation of the propagation scores of the features belonging to a pathway. The first component measures how much the same feature is strongly or poorly involved in the patients. While the second component measures how much the same feature is similarly involved. For example, two patients described by a gene with high but different propagation scores (e.g., 1 and 0.7) are less similar than the pair which has lower but more close values (e.g., 0.8 and 0.7). The components are first determined for each single gene and then are summarized to represent the pathway. More the genes are strongly guilty, more the genes are similarly guilty and more the patients are considered similar in involving the pathway. This also prevents that, one outlier patient with genes strongly associated to the pathway can have a high similarity with another patient when they act on the process differently.

The trending matching similarity measures can be defined as follows; given a pathway $PH_u = \{ F_a \mid a \in \{1, \dots, A\}\}$ and two patient's profiles P_b and P_k, the similarity $TM_u(P_b, P_k)$ is the sum of three components:

$$WJ_u(P_b, P_k) = \frac{\sum_a \min(m_{a,b}, m_{a,k})}{\sum_a max(m_{a,b}, m_{a,k})} \quad (5)$$

$$MG_u(P_b, P_k) = \frac{(\sum_a (m_{a,b} + m_{a,k})/2)}{|PH_u|} \quad (6)$$

$$DIFF_u(P_b, P_k) = 1 - |WJ_u(P_b, P_k) - MG_u(P_b, P_k)| \quad (7)$$

$$TM_u(P_b, P_k) = WJ_u(P_b, P_k) + MG_u(P_b, P_k) + DIFF_u(P_b, P_k) \quad (8)$$

$$TM_u^-(P_b, P_k) = WJ_u(P_b, P_k) + (1 - MG_u(P_b, P_k)) + DIFF_u(P_b, P_k) \quad (9)$$

The TM similarity is designed in two variants. TM_u is designed to capture what we will call upinvolved psPSNs, higher is the second component and higher is the similarity between two patients. The most cohesive class has higher propagation

scores for the same genes than the opposite class. TM_u^- is designed to capture downinvolved psPSNs, lower is the second component and highest is the similarity between two patients. In the next sections and paragraphs, we will detail the operations using only TM_u but the method performs every task with both the variants.

Patient Similarity Networks

Simpati aims to predict the class of an unknown patient comparing its propagated biological profile to the ones of the patients who are composing the classes of interest. For accomplishing the task, Simpati simulates a physician's decision process applied to solve the diagnosis and prognosis of a new individual. Creation of a mental database of known patients linked by their similarity (e.g. Lung cancer patients and healthy controls), selection of the features in which patients of the same class are similar between each other but dissimilar from others (e.g. EGFR biomarker with overexpression in Lung cancer patients with respect healthy controls) and assessment of the clinical outcome of the new individual based on its similarities with the database ones.

Doing a parallelism, the mental database is a set of pathway-specific patient similarity networks. For each biological process, Simpati represents the patients as vertices of a network and weighs their interactions based on the Trending Matching similarity measure. Formally, the similarity network is defined as an undirected weighted graph $PSN_u = (PV, PE)$ composed by a set of vertices $PV = \{pv_1, pv_2, \dots, pv_B\}$ representing the patient's profiles and a set of weighted edges $PE = \{(pv_n, pv_m) | pv_n, pv_m \in PV\}$ with $f': PE -> \mathbb{R}$ such that $f'(pv_n, pv_m) = TM_{u\in\{1,\dots,C\}}(P_n, P_m)$ representing how much the pair of patients is similar in a specific pathway PH_u. The adjacency matrix corresponding to $PSN_u(PV, PE)$ is $W'': BxB = (w''_{n,m})$ where $w''_{n,m} = TM_u(P_n, P_m)$.

Simpati proceeds by selecting the pathways recognised as signature because represented by a PSN dividing the two classes while characterizing one. The members of one class must be more similar (i.e., stronger intra-similarities) than the opposite patients and the two classes not similar (i.e. weak inter-similarities). For this task, we developed a ranking system which evaluates a psPSN from 0 to 10

(the power of a PSN). Higher is the power and more a class is stronger than the opposite one and less the classes are similar (i.e., mixed together due to strong inter-similarities). First, we obtain three distributions based on the values of the similarities in the psPSN. The distribution of the intra-similarities possessed by the members of one class, the distribution of the intra-similarities of the opposite patients and the inter-similarities between the members of the two classes. For each distribution, we compute a low and high percentile (e.g., 0.4 and 0.6). Then, we check if the distribution of intra-similarities of one class has the low percentile greater than the high percentiles of the other two distributions. In case the condition is satisfied, we decrease the low percentile, increase the high percentile, and compare again. For example, power 7 is satisfied when the 20 percentile of the intra-similarities of one class is higher than the 80 percentile of the other distributions. The power 9 when the 15 percentile of one class is higher than the 85 percentile of the other distributions. When, a psPSN has at least power 1 is considered signature of the most cohesive patient class.

When the psPSN is built with the TM_u similarity and has a power greater than 1 is considered signature and upinvolved because the members of the most cohesive class are similar due to higher feature's guiltless than the patients of the opposite class. On the contrary, the psPSN built with the TM_u^- similarity is considered downinvolved because the most cohesive class has the lowest feature's propagation values. Biologically speaking, the two classes are acting on the pathway differently, the members of one class are cohesive because their shared clinical condition is requiring and leading a precise alteration of the pathway, while the opposite class shows an heterogenous behaviour and a less need of acting on that cell function. We designed to capture this topological pattern and we do not require that the weak class must be cohesive following the study of Marquand et al. who reported that, assuming that both the classes in comparison are well defined precludes the inference of true diagnostic labels [122]. A clinical population may be composed of subjects belonging to different subtypes of the same disease or the heterogeneity may appear as result of misdiagnosis and comorbidities.

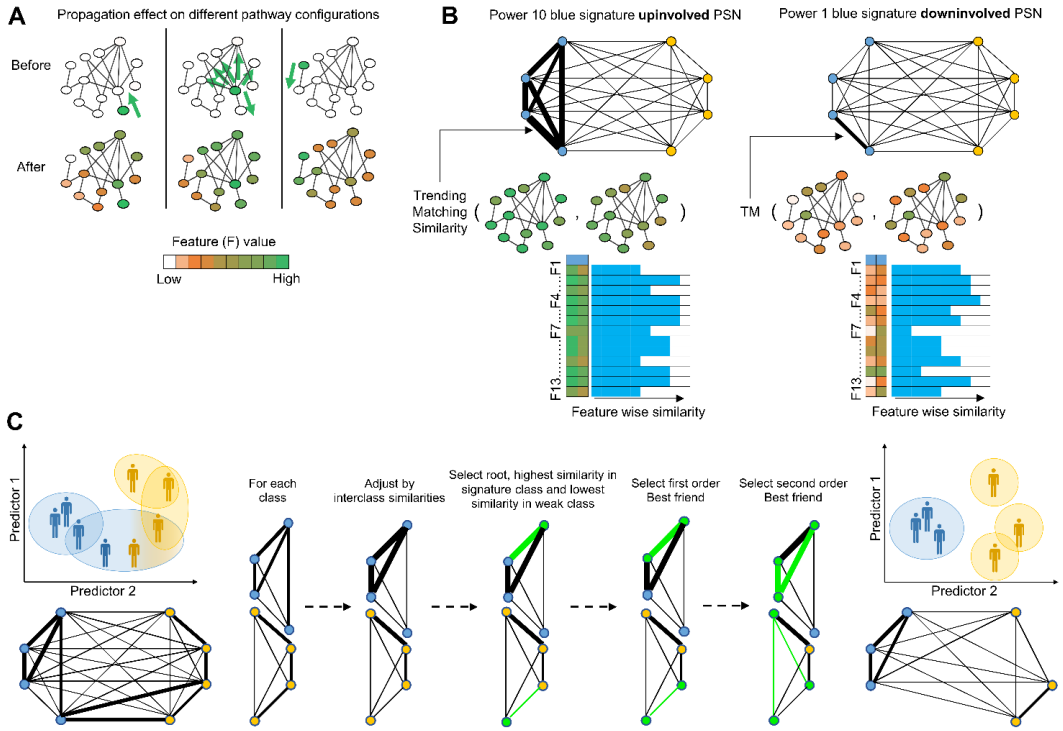
Best Friends Connector Algorithm

Simpati creates the database of psPSNs and then it aims to find the signatures. On the contrary of the starting situation in which a patient is described by the set of its biological feature's values, now the patient is described by its similarities with both the members of its same class and the non-members. This is extraordinary with respect supervised enrichment (e.g., differential expression analysis and gene set enrichment analysis tools) and machine learning tools which do not normally neither expect nor assume that a patient can relate to the individuals of the opposite class. However, the patient similarity network paradigm intrinsically introduces the presence of outliers as it could be likely to have them in the study of a real clinical cohort [122] or disease-specific class [123,124]. The meaning behind is that the outlier patient shows a pathway activity which is different from the one exhibited by the rest of the members.

Simpati tackles this aspect assuming that if a psPSN is not signature, but it still includes a class more cohesive than the opposite one, then this psPSN is likely to contain outlier patients with low intra-similarities and high inter-similarities. If this case appears, Simpati performs a filtering to get rid of possible outliers which are misleading the topological analysis of the network.

We introduce the Best Friend Connector algorithm (BFC) for identifying the most representative and cohesive subgroup in each class, for removing members that are not similar to the majority and for maximizing the psPSN signature power (i.e., the grade of separability between the classes). At first, it determines which class is the strongest, then it performs the selection. For the strongest class, it finds the subgroup with the strongest intraclass, and weakest interclass similarities. For the weakest class, the subgroup with the weakest interclass similarities. The patients not selected in the subgroups are considered outliers and removed from the psPSN. Simpati keeps count of in how many pathways a patient has been considered outlier and it provides this information as result of the workflow to allow a further analysis of the a priori input data. The psPSN is tested for being signature and then recomposed as it was originally.

The algorithm exploits the concept of first order best friend (1BF), second order best friend (2BF) and outsiders. A patient is a 1BF of another member called root when their similarity is in the root's best ones. A patient is a 2BF when it is 1BF of one root's 1BF and it has the root as 1BF. An outsider is a patient that does not belong to the class which is undergoing the analysis. The algorithm performs the following operations for each class individually. It adjusts the weights of the intraclass connections. Precisely, it increases the similarity of two patients when



both have a weak similarity with outsiders, while decreases in the opposite case. Iteratively, it considers one patient as root, it assesses the average of the intraclass similarities of the subgroup composed by his 1BFs and 2BFs. When each patient has been considered as root, the algorithm retrieves the set of best friends who got the highest average. This guarantees of avoiding selecting multiple strong subgroups identified by different roots which would not represent the starting class uniquely.

Fig.6.11. This figure includes graphical details about the phase of data preparation, creation of psPSNs and filtering of outliers based on the Best Friends Connector algorithm. The frame A shows how the result of the propagation changes due to the position of the patient's molecule having a priori information. The propagation models how the pathway is specifically deregulated depending by which molecule is altered by the patient's disease or generically active in a sample's control profile. In the frame B, the Trending Matching similarity captures this aspect of the propagation.

Two patients are considered strongly similar if each pathway's feature is involved in the same way because it means that also the pathway is perturbated and deregulated similarly by the disease of the patients. Plus, the frame illustrates the difference between up and downinvolved pathway. In the first case, higher are the propagation scores and higher is the similarity between patients, while in the second case is the opposite. Frame C shows the steps which compose the Best Friends Connector algorithm. For each class, it adjusts the similarity between every pair of patients based on their connections with the members of the opposite class, then it selects first and second order best friends to compose the subgroup around a root. When all the patients have been considered as root once, the subgroup of best friends with the strongest interclass connectivity is chosen as best one.

For explaining the pseudocode of the algorithm, we introduce the following mathematical details: Given four points in the Euclidian space $Q(x1, y1), E(x2, y2), R(x3, y3), T(x4, y4)$, a vector of continuous values $\bar{V}$, two indexes of patient's profiles $b, k \in \{1, \dots, B\}$ and a value $th \in [1, \dots, 100]$, we define the formula of the quadrilateral area, the formula of the percentile, the average of the interclass similarities in two scenarios with respect b and k, and the first and second order best friend with respect b.

The quadrilateral area:

$$QA(Q, E, R, T) = (1/2) \cdot \{(x1y2 + x2y3 + x3y4 + x4y1) - (x2y1 + x3y2 + x4y3 + x1y4)\}. \quad (10)$$

The percentile formula:

$$percentile\_value(\bar{V}, th) = \{\bar{V}(sup(\frac{th}{100}) * |\bar{V}|) \quad (11)$$

The average of the interclass similarities:

if $b, k \in \{EARLY\}$ then

$$r' = 1 - \sum_{i \in \{LATE\}} \frac{w''_{i,b}}{|LATE|} \quad r'' = 1 - \sum_{i \in \{LATE\}} \frac{w''_{i,k}}{|LATE|} \quad (12)$$

while if $b, k \in \{LATE\}$ then

$$r' = 1 - \sum_{i \in \{EARLY\}} \frac{w''_{i,b}}{|EARLY|} \quad r'' = 1 - \sum_{i \in \{EARLY\}} \frac{w''_{i,k}}{|EARLY|} \quad (13)$$

and the best friends with respect b:

$$1BF_b = \{k \mid k \in \{1, \dots, B\}\ and\ (k, b \in \{EARLY\}\ or\ k, b \in \{LATE\}) AND\ TM_u(P_b, P_k) >= percentile\_value([w''_{1,b}, w''_{2,b}, \dots, w''_{B,b}], th)\} \quad (14)$$

$$2BF_b = \{k \mid k \in \{1, \dots, B\} AND\ P_k \notin 1BF_b\ AND\ (k, b \in \{EARLY\}\ OR\ k, b \in \{LATE\})\ AND\ TM_u(P_h, P_k) >= percentile\_value([w''_{1,h}, w''_{2,h}, \dots, w''_{B,h}], th)\}\ AND\ P_h \in 1BF_b\} \quad (15)$$

$$Max_b = \frac{\sum_{k \in \{BFS\}} w''_{k,b}}{|BFS|}\ with\ BFS = \{1BFb\ U\ 2BFb\} \quad (16)$$

The Best Friends Connector algorithm is then shown in form of pseudocode in Fig.6.12 frame A.

A

BFC(PSN_u,th=5){
For each b,k∈{1,…,B} such that b,k∈{EARLY}
Determine r' and r''; Q=[0,0] E=[0,$w''_{b,k}$] R=[0,r'] T=[$w''_{b,k}$, r''];
$w''_{b,k}$ = QA(P,Q,T,Z)
For each b,k∈{1,…,B} such that b,k∈{LATE}
Determine r' and r''; Q=[0,0] E=[0,$w''_{b,k}$] R=[0,r'] T=[$w''_{b,k}$, r''];
$w''_{b,k}$ = QA(P,Q,T,Z)
For each b∈{1,…,B}
Determine $1BF_b$ and $2BF_b$, Max=0
For each b∈{EARLY}
Determine BFS and Max_b
If Max_b>Max then Max=Max_b and PV''=BFS
Max=0
For each b∈{LATE}
Determine BFS and Max_b
If Max_b>Max then Max=Max_b and PV'''=BFS
PV' ={PV'' U PV'''}
Return PSN'_u(PV',PE')=BFC(PSN_u,x) with PV'⊆PV and a set of edges PE'⊆PE each one of which is incident with vertices from PV' only
}

B

For each u ∈ {1, … , C}
For each x ∈ T = $(5 * n)_{n=1}^{95}$
M1'=M2'=M1''=M2''=0 and th=95
Determine $1BF_z$ assuming z ∈ EARLY
$M1' = \frac{\sum_{k \in \{1BF_z\}} W'''_{k'z}}{|1BF_z|}$
$PSN'_u(PV', PE') = BFC(PSN_u, x)$ with PV' ⊆ PV and a set of edges PE' ⊆ PE each one of which is incident with vertices from PV' only
If $pv_z \in PV'$ then M2' = x
Determine $1BF_z$ assuming z ∈ LATE
$M1'' = \frac{\sum_{k \in \{1BF_z\}} W'''_{k'z}}{|1BF_z|}$
$PSN'_u(PV', PE') = BFC(PSN_u, x)$ with PV' ⊆ PV and a set of edges PE' ⊆ PE each one of which is incident with vertices from PV' only
If $pv_z \in PV'$ then M2'' = x
If M2' > M2'' AND M1' > M1'' then
Add EARLY to the LABELS set
elseIf M2'' > M2' AND M1'' > M1' then
Add LATE to the LABELS set
Consensus_class = $\text{argmax}_a |a \in LABELS|$

Fig.6.12. Figure of pseudocode relative to the two algorithms used in the Simpati implementation. The frame A shows the Best Friends Connector algorithm that is applied to a patient similarity network in form of adjacency matrix to filter outlier patients and detect the most cohesive subgroup. The frame B shows the step of class prediction related to an unknown patient performed at the end of the Simpati workflow.

Classification

Once Simpati created the database of signature psPSNs, it uses them to classify an unknown patient. This for understanding the quality of the selected pathways in characterizing and distinguish the classes in comparison. Simpati performs the

operation continuing to follow the physician's decision process. The unknown patient is compared to the ones already annotated in the mental database and assigned to the same class of who is most similar to. However, the only strength of similarity could be misleading. The unknown patient could have the strongest similarity with outlier members of the class. Therefore, we designed Simpati to consider also how much the unknown patient fits in the class.

The method prepares the unknown patient's profile. The profile is replaced with its propagated version, compared to the patients in every signature pathway and added as new node in the corresponding psPSNs. Then, Simpati associates the profile to one of the classes based on the results of two approaches. For the first, it determines the average of the highest values of similarity that the patient has inside each class. The patient would be associated to the class with which has the highest average. While for the second approach, Simpati pretends that the patient belongs to one class and measures how much is far from being considered an outlier. The patient would be associated to the class in which is considered less outlier with respect the other members. In details about this step, the patient is simulated to belong to one class and the BFC algorithm is performed iteratively. At each run, the algorithm decreases the size of the subgroup of patients which retrieves. It stops when the patient does not belong to the best subgroup. Higher is the number of iterations and more the patient is considered having a stronger similarity with the class representatives than the other members which are more likely to be outliers. Simpati uses the iteration number as distance measure from the "outlier" status. Due to this, the patient would be candidate to be associated to the class in which survived the highest number of iterations.

Simpati associates the patient to the class that has been predicted by both the approaches. In case the results are not concordant, then Simpati does not make the prediction and the pathway together with its PSN are removed from the downstream operations. This step is performed for all signature psPSNs, then Simpati performs the consensus prediction. The patient's definitive class is the one to which has been most frequently assigned.

Formally this would be, a new patient's profiles P_z such that $z \notin EARLY\ and\ z \notin LATE$ is added as node in each patient similarity network. Let us define the new $PSN_u(PV, PE)$ composed by the set of nodes $PV = \{pv_1, pv_2, \ldots, pv_B, pv_z\}$ representing the patient's profiles and the set of weighted edges $PE = \{(pv_n, pv_m) | pv_n, pv_m \in PV\}\ with\ f' : E \rightarrow \mathbb{R}\ s.t.\ f'(pv_n, pv_m) = TMu(P_n, P_m)$. The class of the new profile is found by a topological analysis of all the psPSNs shown in form of pseudocode in the figure 6.13 frame B.

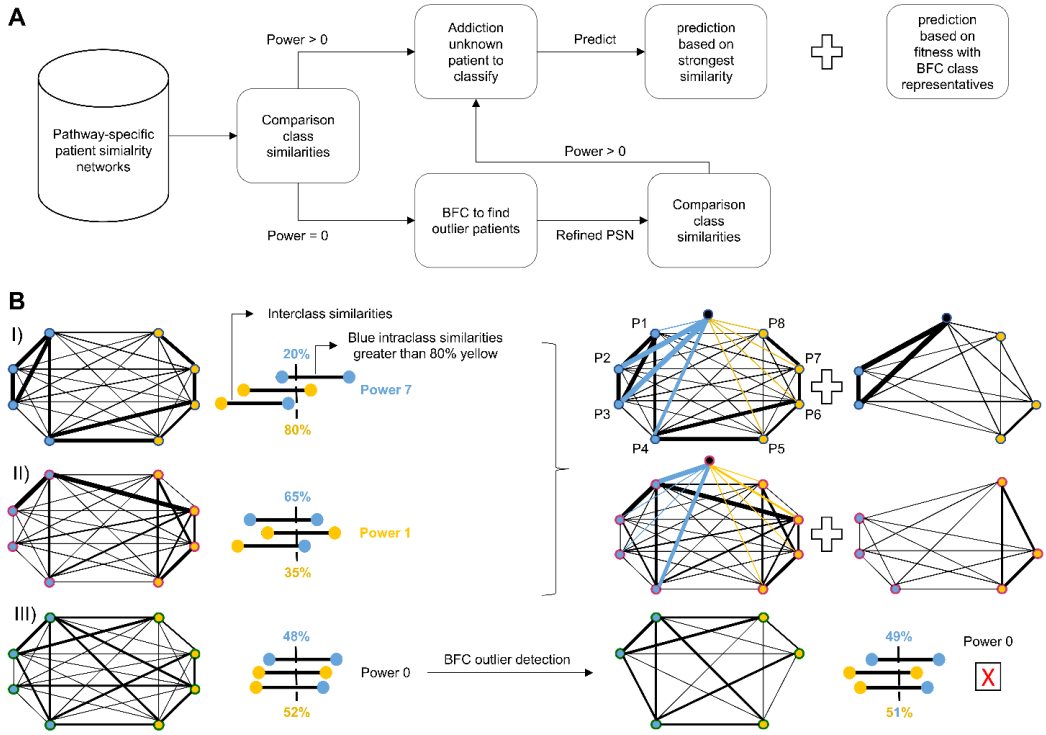


Fig.6.13. This figure shows the journey of a psPSN. In the frame A with a flow diagram, while in the frame B with different graphical scenarios. First, the psPSNs are created and compose a database. Then, Simpati determines the power of the networks. If a psPSN has a significative power is used for the classification, otherwise the method performs the BFC algorithm trying to detect outlier patients. If there is a cohesive subgroup in the psPSN and the removal of outliers significantly increase the power of the network, then the latter is reintroduced in the main workflow as signature pathway. The networks with sufficient power are used to classify an unknown patient. Simpati measures how much the patient is similar to the members of the classes and how much is distant from being an outlier. For this last step, Simpati runs the BFC algorithm iteratively getting each time a cohesive subgroup stronger than the one of previous iterations, more times the unknown patient is found in the best subgroup and less is considered outlier.

Results

Simpati provides the signature pathways used in the classification step, the corresponding PSNs in vectorial format and reports statistics to allow further analysis and considerations. For example, the average of the intra and inter similarities for recognising the most cohesive class, the psPSN power to get the pathways which more separate the classes in comparison and a probability value (p.value). The latter is assessed testing the psPSN to retrieve an equal or higher power than the original one when patients are permutated between classes. This information allows to filter out pathways which have been detected as signature due to random.

Simpati also includes two tools for the visualization of the data produced with the workflow, one tool is an internal function able to produce a compact representation of a psPSN, while one is a graphical user interface (GUI) for the exploration of a patient's propagated biological profile. The function provides a compact representation of a psPSN by reducing the patients which are visible as nodes. This is necessary to allow the user to understand how patients are similar between each other. In fact, more the number of nodes increases and more becomes difficult to follow the edges of the network and so how the patients are similar between each other. To do so, Simpati groups up patients of the same class that are considered similar, chooses a patient to represent each subgroup and filters the original network by keeping only the chosen (i.e., the representative of each subgroup). First, it determines how much every pair of patients are similar in the network. It uses the measure $WJ_u(P_b, P_k)$ which is applied between two patient's profiles composed by their similarity values in the psPSN. It gets a psPSN of the psPSN (aka psPSN^2). Simpati proceeds and iteratively performs the BFC algorithm on the psPSN^2. In this case, the BFC algorithm is not applied to filter outliers but to detect multiple cohesive subgroups. At every iteration, the BFC algorithm detects and removes the most cohesive subgroup composed by the twenty percent of the original members of one class. When the BFC cannot be applied anymore due to the small size of the class, Simpati filters the original psPSN and keeps only the class members which have not been grouped in the last iteration together with the root of each class-specific detected cohesive subgroup. Further about the aesthetic aspects, the size of

a node is used to indicate how much the relative patient is similar inside its own class compared to how much is similar with the outsiders. This is assessed with the difference between two PageRank [125] centrality scores, one is measured only with the patient connected by similarity to the members of its class and one with only the outsiders. Higher the difference, higher the size of the node, more central the patient is in its class and less similar to the outsiders. While, the position of the nodes in the plot of the psPSN is determined using the Fruchterman & Reingold's force-directed layout [24]. The network becomes compact, easy to analyse and still representative of how all patients are similar. Thanks to the plot, the user can understand if the psPSN is correctly a signature, see how the similarities are distributed, identify which patient is crucial for the connectivity of its own class and which is instead behaving as outlier.

Complementary to the visualization of the psPSN, we also provide an R shiny graphical user interface (GUI) to allow the exploration of the propagation effect over a patient's profile. This enables the user to understand how the values of the patient's biological features changed and for which reason. For example, a gene with low expression value that has been removed from the Limma analysis due to the function "filterByExpr" in the differential expression analysis, it may get a high propagation score and the user may get interested in understanding the reason. We believe that this can be another useful instrument to make the method and the data more accessible.

Enrichment

The power, the p.value, the distribution of the similarities are all technical information regarding a psPSN that allow to understand how patients and classes are structured. However, they offer a limited utility in prioritizing the best pathways because they are not related to any biological background of the patients. For this reason, in case the patient's features are genes, we designed Simpati to perform a query in the Disgnet [126] and Human Protein Atlas [127] (HPA). DisGeNET is a database which provides open access to annotated disease associations with genes and variants. While HPA is a unique world-leading effort to map all the human proteins in cells, tissues, and organs in the human body using antibody-based imaging, mass spectrometry-based proteomics, transcriptomics, and systems

biology. Simpati requires the semantic type of the patient's disease (e.g., Neoplastic Process, Congenital Abnormality, etc..) and key words (e.g., TCGA-KIRC: Kidney, Renal, Carcinoma). Then, it gets which published articles have associated the genes of a signature pathway to the semantic type and to the key words. Plus, in case of cancer patients, this integration allows to get which genes are favourable to be prognostic, while in case of non-cancer disease which genes are associated to the user-defined tissue. As indicated by Lin et al. [128] this operation allows to prioritize the signature pathways based on their associations with the patient's clinical outcome and to understand better the validity of Simpati results.

Cross-Validation Setting

Simpati ability to classify the classes in comparison is tested with a leave one out cross validation (LOO-CV). Given a dataset of patient's biological profiles and the classes associated to them, Simpati iteratively performs the following operations: one patient is considered unknown and compose the testing set, while the remaining patients are considered known and used as training set. The latter is used to build the psPSNs, to find the signature pathways and as ground truth in the classification step. While, the testing patient has the biological profile which class must be predicted. In the end, the predicted classes of the testing patients collected from all the iterations are compared to their real ones for determining the classification performances. Simpati is designed to value its ability to predict based on two measures following netDx design [62]. The first one called AUC-ROC is the area under the curve where the x-axis is the false positive rate (FPR) and the y-axis is true positive rate (TPR). While the second one called AUC-PR is the area under the curve where the x-axis is the recall and y-axis is the precision [129].

6.3.5 Comparisons

We tested Simpati performances to classify patients of five different TCGA cancer types described by two biological omics: transcriptomics with gene expression data and genomics with somatic mutations. The classes assigned to the patients were Early or Late based on their cancer stage. We increased the challenge including the performances of the current published generic-purpose pathway-based classifiers: netDx [62] and PASNet [81]. netDx creates a database of predictive PSNs

associated to pathways for each class, builds a consensus similarity network and applies GeneMANIA (state-of-art gene function prediction algorithm) for the prediction of the testing patients. netDx tests its performances with a 10-fold cross-validation which in each run includes another cross-validation for the selection of the most predictive PSN. While, PASNet incorporates biological pathways in a Deep Neural Network. The neural network is composed by a gene layer (an input layer), a pathway layer, a hidden layer that represents hierarchical relationships among biological pathways and an output layer that corresponds to the patient classes. PASNet tests its performances with a stratified 5-fold cross-validation repeated 10 times. The two competitors either support or use the classification evaluation based on the area under the receiver operating characteristic curve and the area under precision-recall curve measures and they differ from canonical supervised machine learning algorithms. For these reasons, we performed the comparison using each method based on how it has been designed and following the vignettes provided by the authors.

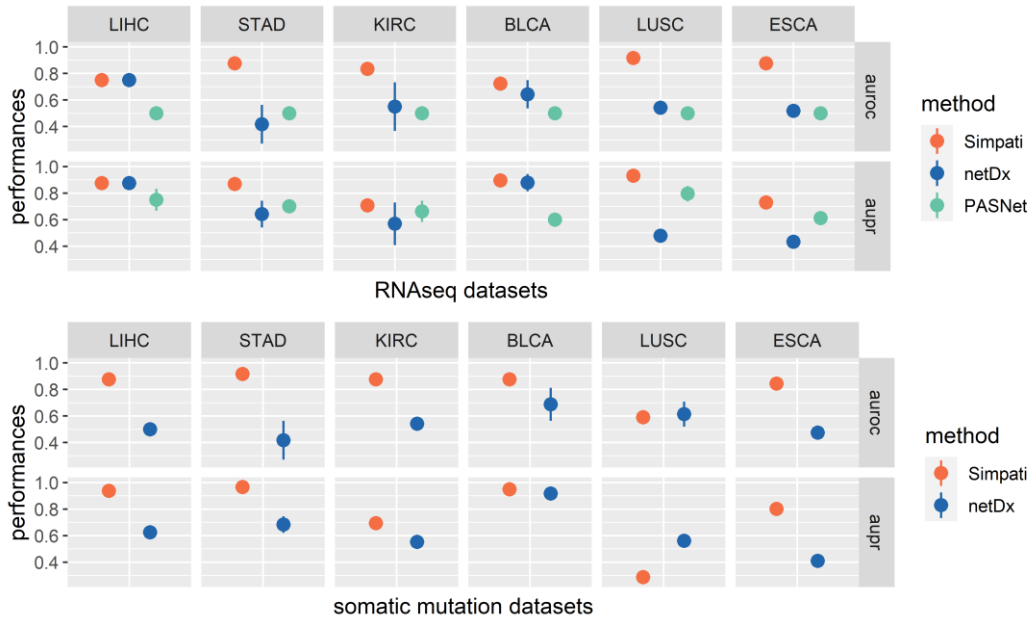


Fig.6.14. Comparison of the classification performances between the pathway-based classifiers Line plot of median (dot) classification performances with error bars (line). X-axis indicates the datasets. Y-axis indicates the value of area under the roc/pr curve. The same plot is presented twice, one including the performances when the methods classify the RNAseq data, while one the somatic mutations. PASNet does not have performances with somatic mutations because it does not handle sparse biological data. The plot shows that Simpati performs better than the competitors in all the

datasets except for LUSC with somatic mutation.

Simpati performs better than the competitors for both the measures with all the omics. Simpati also proves to be more reliable in each dataset with a standard error equal to zero due to its leave one out cross-validation approach. While the performances of the competitors highlight common classification issues; their performances vary a lot probably due to the number of patients, the size of the classes in comparison and the ability of the classifier to naturally handle multiple omics and data types.

Simpati and netDx provide pathways and related PSNs as result of the analysis. However, the methods use different techniques for the pathway selection. netDx selects a pathway if the corresponding PSN allows GeneMANIA to correctly predict the classes of the training and testing patients. While Simpati selects a pathway if the corresponding PSN topologically separates the classes in comparison. The resulting pathways and related PSNs should help to characterize the patient classes, explain why they have been used to predict and they should increase the interpretability of the model. We compared the topology of the PSNs selected by the two contenders based on their power.

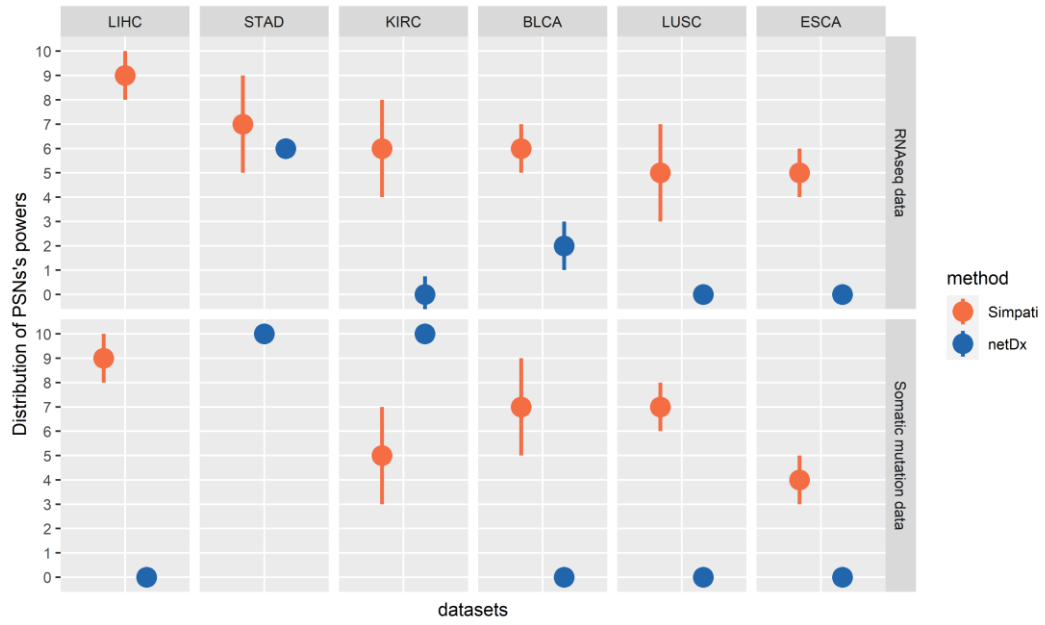


Fig.6.15. Comparison between the topology of the PSNs retrieved as result by netDx and Simpati with the TCGA datasets. The topology of the PSNs is measured with their power. Each frame of the image is dedicated to the pathways selected for the classification of a specific biological omic. The

Y-axis indicates the power of the PSNs retrieved by a specific method. The X-axis indicates the datasets. Specifically, the dot indicates the median of the power of the PSNs resulted by applying a specific method with a specific dataset, while the line ranges based on the standard deviation of the same PSNs' powers. Simpati selects better PSNs except in STAD patients described with somatic mutation profiles.

Simpati provides more pathways with high power than netDx in all the datasets except one. This is probably due to how the selection is done. Simpati discerns PSNs based on their topology and then performs the classification. While netDx evaluates a pathway based on the mere ability of the GeneMANIA algorithm to use its PSN for classifying. This makes the difference in terms of interpretability of the model. From the final user perspective, Simpati's psPSNs together with their visual representation make easier to understand why they have been selected for the classification and can be perceived as more trustable.

The patient similarity network paradigm used by Simpati and netDx brings many advantages both in the feature selection, in the classification phase and in the overall interpretability of the software. However, these pros come with a price which is the software scalability already introduced as challenge by Pai et al. [82]. A PSN is a complete graph that the methods build with all the patients and for every pathway. This means that an increment in the number of patients and in the number of annotated pathways lead the methods to require more computational resources. netDx and Simpati faced this point with different approaches. netDx is implemented in R and Java, uses the disk to save temporary files and applies a sparsification of the PSNs to decrease the number of edges and so the amount of information associated to them. While Simpati is implemented completely in R, natively support parallel computing and handles all the data of the workflow as sparse matrices or vectors. To understand which software handles better this issue, we captured the ram usage and the running time that each method required to classify the TCGA datasets with the same hardware setting (AMD Ryzen Threadripper 3970X 32-Core Processor, 251 Gigabyte System memory and Linux ubuntu-1804-slurm 5.4.0-72-generic). Simpati resulted to be more efficient in both the running time and the ram used.

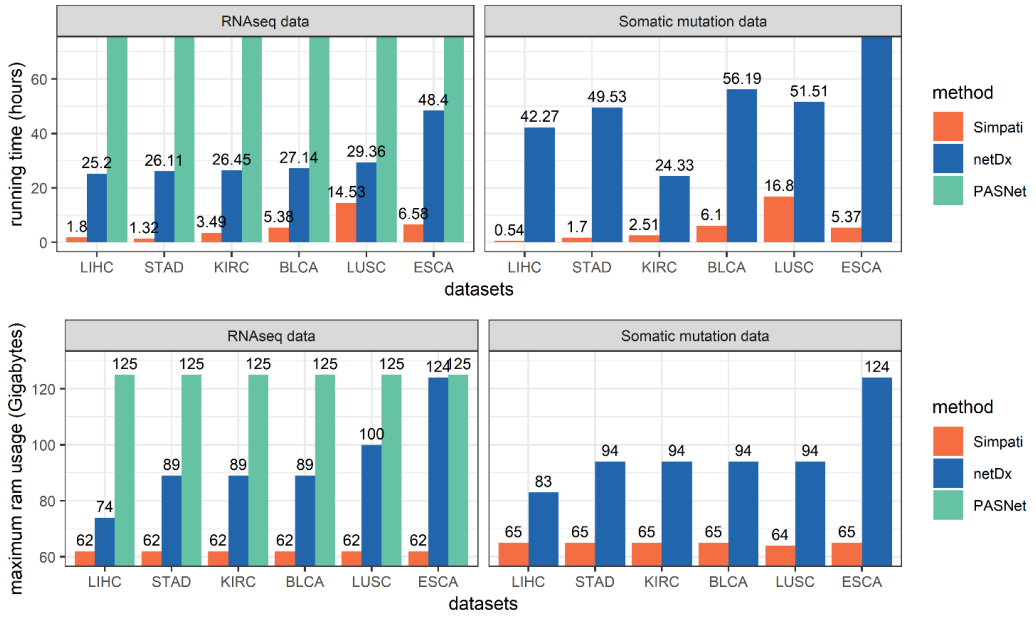


Fig.6.16. Barplot shows the comparison between the computational resources used by Simpati, netDx and PASNet to classify the TCGA datasets. The measures used for this comparison include the running time in hours and the memory ram in the maximum amount needed by the software in Gigabyte. In fact, the maximum amount is the real obstacle to the correct execution of the software. The X-axis indicates the datasets. The Y-axis indicates the measure. PASNet with the RNAseq data has a running time which exceeds the three days (72 hours). The plot shows how Simpati outperforms the competitors in time and memory for all the datasets.

Pathway-based classifiers aim to classify correctly unknown patients using the biological pathway information. This means that the prediction of a patient's class passes through the selection of pathways which due to method-specific criteria are considered useful for the task. In a cross-validation setting, the final classification performances indicate how much the classifier is reliable and better than a random predictor. However, they do not represent a measure of how much the pathways are biologically significant. A classifier as Simpati can provide further details about how it used the pathways, why it selected them and the biological interpretation under the filtering criteria but still, this information does not allow to understand if pathways are biologically meaningful. For this reason, we designed Simpati to integrate an enrichment step and we performed this operation also to the results of the other competitors. Precisely, we kept only the resulted significant pathways having at least one publication associated to each of the key words defined per

dataset and having at least the 90% of the genes associated to the patient's specific cancer. Then, we compared the numbers of pathways satisfying these constraints.

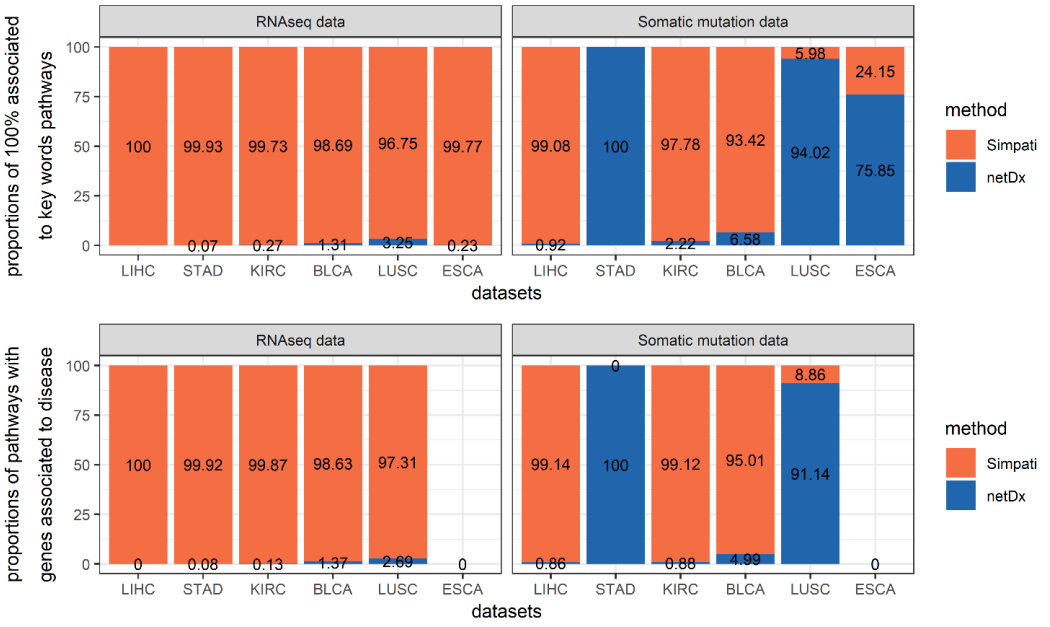


Fig.6.17. 100% stacked bar chart shows the comparison between the enrichment statistics obtained by querying the resulting pathways of the different classifiers in Disgnet and the Human Protein Atlas with respect the patient's cancer types. Each frame compares the methods based on how their pathways are qualified with a specific measure. The X-axis indicates the datasets in comparison. A bar is divided based on the number of pathways obtained by the two methods and satisfying the criteria indicated by the Y-axis. For example, netDx did not selected any pathway for the classification that satisfied the criteria in the classification of the LIHC patients with RNAseq profiles, while for the somatic mutation profiles Simpati has selected 99% more pathways.

This analysis highlights that Simpati is both able to select, use and provide biologically significant pathways directly associated to the patients that it is classifying and that performs better than the competitors. netDx retrieves always much fewer pathways biologically associated to the tumor of the patient's profiles than Simpati. We have been unable to include PASNet due to its lack in providing the pathways that it considers significant and predictive of the patients classified.

In this last paragraph, we compare the Simpati signature pathways obtained from the classification of KIRC patients with literature findings. We considered the key pathways identified by Pang et al. [130] and Cui et al. [131].

Late stage KIRC regulates the cell adhesion to escape from the immune system. It disturbs the ATP supplement, making cells to adopt anaerobic respiration, producing an ATP deficit and hypoxia. It provokes an immune response from the host organism, regulates the PI3K-Akt signalling pathway, promotes anabolism and inhibits catabolism.

Simpati has been able to find key pathways differentiating the pathological stages of the patients. Specifically, it confirmed the relationship between the late cancer stage and both the upregulation of the cell adhesion, the response of the immune system and the downregulation of the catabolism. For this comparison, we also included the results of the standard gene set enrichment analysis (GSEA) performed thanks to the GAGE method [132]. The two methods found similar key pathways. However, GSEA provided the statistics mean, while Simpati provided the patient similarity network.

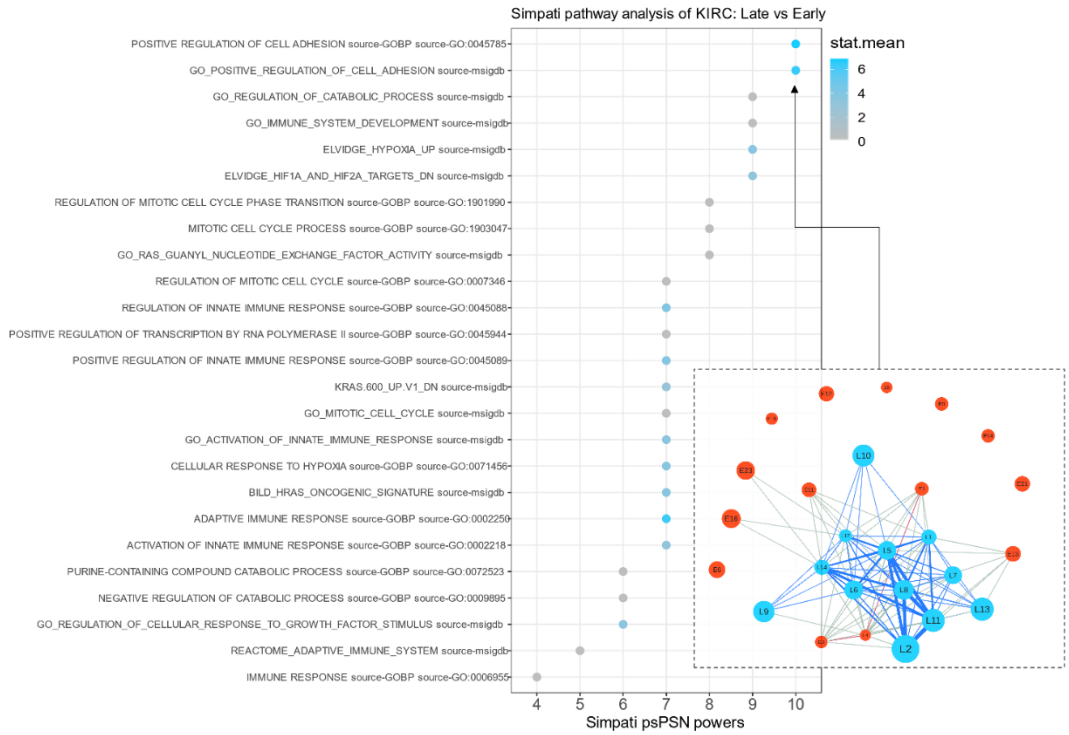


Fig.6.18. Dotplot showing the signature pathways detected by Simpati and the standard gene set enrichment analysis for the comparison of the KIRC patients between Late and Early stage. In the y-axis, there are the signature pathways and psPSNs resulting by performing Simpati. The x-axis indicates the power of the similarity networks. The legend describes how much the standard GSEA finds the same pathways altered between the two classes. The pathways with zero statistics mean (stat.mean) are not significant for the GSEA. Finally, in the bottom right cornet, the plot of the

psPSN related to the positive regulation of the cell adhesion pathway. Blue nodes represent Late KIRC patients, while red nodes represent Early patients. The blue edges are the Late interclass similarities and the grey are the interclass.

6.3.6 Conclusions

We propose the pathway-based classifier called Simpati. The method can be applied to different omics, proved to obtain quality classification performances and detect signature pathways. In other words, it identifies biological processes that distinguish and uniquely characterize the clinical classes of the patients in comparison.

On top of the technical conclusions, we want to suggest Simpati as tool for computational biologists and bioinformaticians that want to get unique insights about their patients or samples. We designed Simpati to simulate a physician's decision process applied to solve the diagnosis and prognosis of a new individual. As a physician, our software processes, stores and learns information related to the patients. All the data used during the classification are then made available for allowing further analysis.

Simpati associates to each single biological feature (e.g. gene, protein, mutation, ...) a propagation score which reflects the overall biology of the patient. A high score indicates that the feature is strongly involved in the patient's biology, while a low score the opposite. The scores can be explored in two ways. They can be considered as values of a standard high-dimensional matrix with patients at the columns and features at the row. They can be visually taken into account in an ad-hoc graphical-user interface. The matrix format allows any statistical analysis with clusterProfiler [133], while the GUI permits to understand how much specific biological features of interest are important without any programming and statistical knowledge. The information retrieved by analysing the propagation scores can be combined to the results obtained from a differential expression (DE) analysis. For example, a DE list can be filtered to keep only the genes that have a high score in order to reduce the false positive or can be expanded integrating those genes that are DE in term of propagation values.

Simpati models pathways as patient similarity networks. In a psPSN, patients are connected to others based on how much their biology similarly regulate a specific pathway. Like in a social network in which people are connected to others based on their hobbies and how they practice them (e.g., the place, the effort, the time). More two patients involve and regulate similarly a pathway (e.g., with the same genes and with the same expression values) and more they are strongly connected. In case a pathway is found significant and of interest, it can be explored in two ways. The adjacency matrix or the graphical representation of the related psPSN. The matrix format allows any network analysis with NetworkToolbox [134], while the plot permits to have intuitions about how much the patient classes separate and to identify patients that are central or tend to be outlier. The information retrieved by analysing the topology of the psPSNs can be used to verify the clinical information associated to the patients, identify subclasses, and can be combined to the results obtained from a clustering analysis or a non-negative matrix factorization (NMF) [135]. For example, the subclasses of patients that have been identified with an unsupervised technique can be checked in the psPSNs to find in which pathways are mostly similar.

Simpati finds signature pathways to characterize and distinguish the patient classes in comparison. The pathways must satisfy a constraint. The members of one class must be more similar than the opposite patients. Then, it uses the similarities to predict the class of new patients. In this sense, Simpati can be combined to a standard gene set enrichment analysis because detects pathways that satisfy a criterium not taken into account by other tools.

Simpati does not assume that the patient classes are well defined, and it considers the possibility that members of the same class may regulate the same biological process differently. When this is likely to happen with a pathway, Simpati identifies the patients that most represent their own class, uses only them to check the signature condition and the remaining members are considered outliers. For this reason, Simpati is suitable to real case scenarios which often include either patients or samples associated to clinical outcomes due to a priori information by wet lab scientists. The latter check the expected sample classes using a principal component analysis or clustering. However, both the methods are not designed to detect

differences at the level of pathways or single biological features which could reveal unique biological aspects of a sample and differentiate it from the rest of its class. For example, in a knock-out study, samples are labelled as knocked-out based on the experiment but this does not necessarily imply that each member of the clinical population shows changes in gene expression levels against the control group. Standard gene set enrichment analysis tools are ineffective due the possible low variation between the classes and possible knock-out samples not showing any change.

6.3.7 Discussion

Generic purpose pathway-based classifiers propose themselves as powerful tools for classifying patients and providing biologically meaningful results in form of pathways. The first one has been introduced in the 2010 but, at the time of writing, there are very few software available. they remain very few. This is due to the many challenges that must be faced to produce high-quality and functioning software they inherit. We tried to report and detail all the issues related to the development of this kind of machine learning algorithm. At the same time, we tried to build a software that could have been considered a future example for other researchers. Thanks to the combination of new and popular strategies, Simpati proves that is possible to both obtain satisfying results and tackle common issues in the pathway-based classification.

The preparation of the patient's biological profiles with a transformation technique as the network propagation allows to get the same kind of data and information before the classification. This allows the researchers to develop a workflow which is flexible, consistent, and involving less hyper-parameters. As a matter of fact, we developed the Trending Matching similarity to capture a specific relationship between the patient's propagated profiles and the scored genes. On the contrary, netDx suggests using the Pearson correlation as default measure directly on the raw profiles but the authors did not provide a biological interpretation of what kind of patient similarity leads to catch and leave the user to the uncontrolled intrinsic disadvantages of the measure [136]. For example, the correlation is biased and leads to incorrect inferences when considers genes that have not been perturbed (e.g.,

environmentally or genetically) in order to cause a meaningful change in expression level [137].

The selection of the predictive pathways including criteria that can be explained from a biological point of view allows the classifier to not drift away from the patient's biology. For example, both PASNet [81] and netDx [61] select the pathways which perform the best in predicting the training patients. This approach is undeniable well suited for securing the ability to predict an unknown patient. However, it may lead to select pathways which are useful for the algorithm of prediction and meaningless for the patient's biology. On the contrary, Simpati performs a first selection based on criteria that can be biologically interpretated and then it analyses the pathways for the classification. As we proved with our results, a biological selection does not necessarily negatively affect the final classification performances but it indeed changes positively the resulting pathways produced by the classifier.

The analysis of outlier training patients because their biological features are not showing the same activity in a pathway as the rest of the class increases the granularity of the cell process description that the classifier handles and provides as result. This brings multiple advantages. It makes the classifier less sensitive to how much the data have been cleaned, how the patient classes have been defined, and it allows to give a hint about subclasses of patients which are using different pathways at the net of one shared clinical status.

At the same time, it is worth to also mention the price of such strategies. The propagation leads the classifier to require also the network of interactions or associations between the features of the biological omic. The selection of pathways based on criteria which are biologically explainable is not trivial and may makes the classification inconclusive due to no pathway passing the filter. The analysis of the outliers requires either parameters to set up which may be not correct for all the applications or hyper-parameters to determine.

7. CONCLUSIONS

Patient similarity network: each node is an individual patient and an edge between two patients corresponds to pairwise similarity for a given characteristic. In this paradigm, each input patient datum (e.g., age, sex, mutation status) can be represented as a network of pairwise patient similarities. Each network is a "view" of patients that can be integrated with all the other views to identify patient subgroups and predict a condition, phenotype, or generic outcome. In case, the prediction is correct, the network and the reason why patients are connected by similarity can be defined as marker correlating with the patient's phenotypes in study.

As a simple example of the concept, the class of lung cancer patients is compared to healthy individuals. Representing the smoking frequency as a patient similarity network could likely show a correlation with the patient's conditions. Patients who are frequent smokers would be tightly connected to each other, while others would be connected in an heterogenous manner. This network is potentially highly predictive of lung cancer status, as disease cases would be likely similar. If a new patient is not a smoker, it would be more similar to healthy controls and a classifier would predict it as such. In case the network is proved predictive, the smoking frequency is a marker and correlates with the patient's conditions.

This thesis presents three patient classifiers based on the patient similarity network paradigm. Patient data are encoded as patient similarity networks (e.g., age, gender, pathways, metabolite quantities, gene expression levels). To train the model, patients are partitioned into two groups: a training set and a testing set. The training set provides PSNs to describe each patient class in study and enables the model to learn predictive networks. A selection is used to identify the smallest set of the most predictive PSNs, which can help understand how the prediction is being made and can speed up prediction. A recommender system predicts the class of the testing patients based on their similarities to the training ones in the best PSNs. The independent testing set allows to measure the classification performance of the model. A common concern in model-building is overfitting, when the model learns weights based on the bias in the training sample and does not generalize to the wider

population. Cross-validation is used to estimate model generalizability; here, a portion of the training sample is held out from the learning process and is used to evaluate fitting error on the held-out testing set, and this is repeated many times. There are several measures for evaluating a model's performance: these include the balance between the specificity and sensitivity of the model (area under the Receiver Operator Characteristic curve or AUROC; area under the Precision-Recall curve or AUPR). The ideal classification algorithm is accurate, generalizable, provides a prediction in a reasonable time frame for clinical decision-making and is interpretable so that it can be understood by a clinician.

The three proposed classifiers similarly test a patient's characteristic based on how the patients are similar due to it. If the similarity network proves itself of being a marker because enabling the correct recognition and prediction of testing unknown patients, then the characteristic is considered correlating with the patient classes. On the opposite side, the differences between the classifiers lie on which data are tested and how the patient similarity networks are processed.

netDx focuses on determining which generic datum and information (e.g., age, gender, transcriptomics, genomics, etc ...) creates a patient similarity network that GeneMANIA, an external recommender system, is able to use to predict correctly the testing patients. Pratic focuses on determining which biological pathway creates a patient similarity network that GeneMANIA is able to use to predict based on patient's somatic mutations. Simpati focuses on determining which biological pathway creates a patient similarity network that topologically shows a specific pattern (the members of one class are cohesive and similar, while the non-members are sparse) and predicts testing patients based on their similarity with the training ones.

netDx adopts an operation of sparsification to remove similarities in the networks and uses GeneMANIA to selects the best PSNs. Pratic tried to add a filtering step to netDx workflow for removing networks not showing a topology able to distinguish the two patient classes, but the approach proved to be incompatible with GeneMANIA criteria of selection and classification. Simpati evolves the Pratic

topological selection of the PSNs and does not use a sparsification operation that costs hyper-parameters and information.

netDx requires a similarity measure per datum, includes an operation of integration of multiple networks representing different information and serves as generic classifier. Simpati does not require a similarity measure from the user, includes a standardization of the single biological omic which considers the interconnected nature of the described molecules and is designed to simulate a pathway analysis.

In overall, a patient similarity network framework allows to build classifiers that are accurate, generalizable, able to integrate heterogeneous data, and naturally handle missing information. This paradigm also provides excellent model interpretability and additionally, may be better suited to protect patient privacy than most established machine learning methods.

Thanks to the development of these classifiers, patient similarity networks proved to naturally handle heterogeneous data, as any data type can be converted into a similarity network by defining a similarity measure; once converted, all data is represented in the same manner, as a network that can be directly input into analysis methods. Further on, they natively manage missing information, as a patient missing in one network may be in another and could still be used. They can be easily represented and conceptually intuitive to understand and interpret; in this sense, decision boundaries which have been considered by PSN-based methods can be visually evident. They allow to simulate the clinical diagnosis, which often involves a physician relating a patient to a mental database of similar patients they have seen. They can be feature engineered for further improving the model interpretability. For instance, creating networks at the level of biological pathways helps to identify cellular processes that may be causal mechanisms for a given patient phenotypic class.

However, at the time of writing, the PSN paradigm is still novel; two methods have used this framework for patient clustering [83,84] and other two, netDx with Pratic and Simpati, for supervised classification. When compared to other clustering and classification approaches, these methods demonstrate superior performance.

Having already described the PSN-based classifiers, two clustering methods have been reported to date. The first identified subgroups of type 2 diabetes patients using 73 clinical variables obtained from electronic medical records of ~11,000 patients [84]. Networks were generated using singular value decomposition and cosine similarity, the latter being a popular similarity metric in text mining applications. Using medical records and genotype data on the same individuals, the authors demonstrated that identified patient clusters were enriched for different comorbidities and biological pathways. In Similarity Network Fusion (SNF), a patient similarity network is generated from each input data type; for continuous valued measures, similarity is based on Euclidean distance followed by exponential scaling [83]. The set of networks is then fused by iteratively increasing the weights of edges that are concordant among different layers and decreasing the weights of those that are only present in some but not all layers. Spectral clustering is then applied to "cut" the final network into highly interconnected clusters. SNF performance was benchmarked against naïve integrative clustering and a method based on joint latent variable models. Patient subgroups were identified in five tumours by integrating mRNA expression, DNA methylation and miRNA expression [83,138]. SNF significantly outperformed the other approaches in identifying clinically distinct clusters in all cases, and demonstrated consistent fast algorithm run times regardless of the number of genes included in the input data **57**. Since its development, SNF has been used in various applications, including subtyping medulloblastoma patients from DNA methylation and gene expression, and clustering pancreatic ductal adenocarcinoma tumours from RNA, DNA methylation and miRNA expression [139,140].

With the development of digital healthcare systems and high throughput laboratory technologies, a variety of patient-specific outcomes such as diagnoses, treatment records, biochemical tests, genetic information are electronically available [141]. The ability to automatically build a patient similarity network that is able to correlate an outcome/condition/phenotype with a patient's characteristic either clinical or biological, without incurring into additional efforts from physicians can improve the efficiency of our health care systems and be beneficial for both patients and hospitals. Precisely, one field that PSN based classifiers would contribute to is

the precision medicine. According to the Precision Medicine Initiative, precision medicine is "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person" [142]. With the right patient similarity network, physicians could retrieve a cohort of similar patients for a target patient in care, make medical comparisons and, thereafter, make personalized treatment plan effectively.

However, especially for the age of the these classifiers, many challenges must be solved to reach their full potential. First, methods must be improved to: 1) handle large data sizes (e.g. thousands of genomes); 2) identify the most relevant features for prediction including non-linear interactions between features (e.g. pathways); 3) automate ways to generally improve signal-to-noise ratio; 4) automate ways to characterize patient heterogeneity, like disease subtypes [143–145]; 5) make the best use of complementary omics which may have complex relationships (e.g. gene expression is modulated by genetic variants in a tissue-specific manner) [146].

At the same time, it is exciting to envision a doctor’s clinic of the future, similar to the one described by Friend and Ideker [147], that uses network-based approaches for clinical decision-making. A clinical researcher would identify patients to include in model training and select which types of clinical and genomic data to include. The training of the model would run on centralized high performance computing systems, and results could then be interactively visualized in a web-based interface. Following completion of a research study, similarity networks correlated to a phenotype could be uploaded to a repository such as NDEx [148] for sharing with the research community. Eventually, as the technology matures and as classifiers are validated, it would evolve to be useful to practicing physicians for use with their patients. This would require the development of additional reporting tools tailored for use in clinical decision-making. These would include a summary report card of overall confidence in the predictor as well as classification accuracy for a given patient, graphical summaries of relevant features used, and alerts about specific patient details that would affect result interpretation (e.g. ethnicity, lifestyle, genetic variants). It would also include links to relevant medical literature associating specific features with the disorder, to provide the clinician with information on prior knowledge to aid in decision-making. It would also provide

the history of success rate for specific treatment choices for this condition in the health system, which would improve with data collection over time. Algorithms like SNF, netDx and Simpati advance several ideas to achieve this goal. They permit the usage of several omics for patient subtyping or classification to directly answer specific clinical questions. Simpati identify biological pathways which alteration is predictive of patient outcome, integrate publish databases, provide visualization tools. It is closing the gap between current machine learning algorithms and the ideal methods that will appear in the future for finding correlations between genotype and phenotype.

8. PUBLICATIONS

miR-669c

Intracerebral overexpression of miR-669c is protective in mouse ischemic stroke model by targeting MyD88 and inducing alternative microglial/macrophage activation [71]

We investigated non-coding micro RNAs (miRNAs) (type of non-coding genomic element which regulates gene activity similar to lncRNAs) and their effects on cell type specific biomarker genes of Stroke. After a canonical bioinformatics analysis involving differential analysis, pathway analysis and integration of mRNA with miRNA data, I had to prioritize the miRNA-target interactions to study and verify the most important ones in-vitro. The problem was about: one miRNA can regulate multiple genes and it does not exist a one-size-fits-all methodology to prioritize or rank miRNA-target interactions. We solved this issue developing two approaches. About the first one, it retrieves the anticorrelated miRNA-targets, profiles the associations based on gene set enrichment analysis, identifies the miRNAs targeting genes which modulate important disease-association pathways (guided by the biologist or clinician) and ranks them performing a multi-criteria decision analysis (MCDA). While about the second technique, it exploits graph-theory. It uses the differentially expressed miRNAs and genes retrieved by the comparison in study (stroke vs control) to create a network composed of miRNA-targets interactions and gene-gene interactions. Then it analyzes the topology of the network to find which target is the most crucial to hold the disease specific connectivity and allow the altered signaling cascade to strongly deregulate the cell functions. Based on these two approaches, we discovered miRNA-targets interactions associated to the disease and we verified in wet lab.

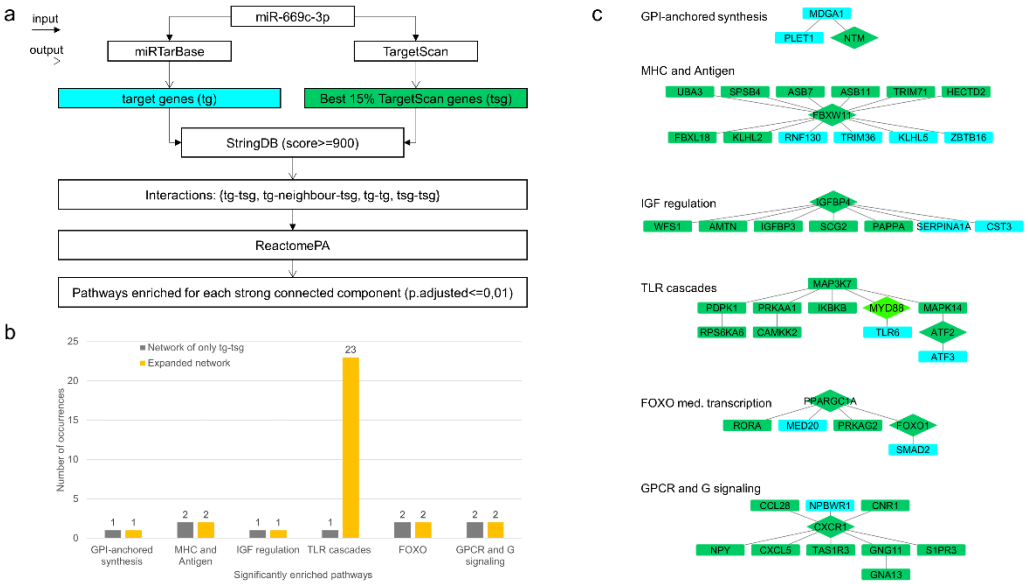


Fig.8.1 Neuroinflammation-related network analysis revealed MyD88 one of the most relevant predicted targets for miR-669c-3p. Panel a represents the workflow adapted in order to detect new genes significantly associated with miR-669c-3p. MiR-669c-3p was used as an input to query miRTarBase and TargetScan databases. Both experimentally validated targets from miRTarBase and TargetScan genes associated with miR-669c-3p were subsequently evaluated in STRING interaction database. Then a network was created, composed by the interactions between aforementioned targets depicted as nodes, as well as the interactions between TargetScan and miRTarBase targets including only those with high confidence (score ≥ 900). Finally, a pathway enrichment analysis of each connected component of the network was applied using the R package ReactomePA and only pathways enriched with an adjusted p value < 0.01 were retrieved. Panel b shows the results of pathway enrichment analysis prior (gray bars) and after (orange bars) the addition of interactions between TargetScan and miRTarBase targets in the network construction according to the workflow. When the network includes within-group interactions, only the connected component containing MyD88 (TargetScan target) and TLR6 (miRTarBase target) significantly increases the number of enriched neuroinflammatory pathways (23 Toll-like-receptor pathways). Panel c features the network obtained with adapted workflow. Diamond-shape green nodes are TargetScan targets directly connected to miRTarBase targets represented as rectangle-shape blue nodes. Rectangle-shape green nodes are the TargetScan targets connected to blue nodes due to the within-group interactions (green-green). Each subnetwork is one of the connected components enriched in pathways associated with neuroinflammation.

The analysis and results can be replicated with the code provide in the following repository: https://github.com/LucaGiudice/Network-analysis-miR-669c

Denove Assembly for lncRNAs

Long Non-Coding RNAs as Molecular Signatures for Canine B-Cell Lymphoma Characterization [70]

We studied non-coding RNA in Diffuse large B-cell lymphoma (DLBCL), marginal zone lymphoma (MZL) and follicular lymphoma (FL) in dogs. Here the problem was about: dog genome was not well annotated. Therefore, aligning and assembling transcripts, creating a unique transcriptome, comparing the data with the reference genome was leading to detect neither accurate nor novel lncRNAs. We faced this issue developing an R pipeline which performs a transcriptome assembly combining multiple algorithms to uncover novel lncRNAs and delineate genome-wide expression of unannotated and annotated lncRNAs. The development of the pipeline required a lot of tuning of hyperparameters which have been done in dialogue with dog cancer experts. Plus, the pipeline also included a software for detecting functional modules created with lncRNA-genes interactions. In conclusion, our study provided an in-depth analysis of the lncRNAs transcriptome in canine B-cell lymphoma subtypes. LncRNAs, quantified by our pipeline, separated normal from pathological samples and uncovered previously unidentified differences between DLBCL and MZL. We also found clusters with different prognostic outcomes within the DLBCL histotype.

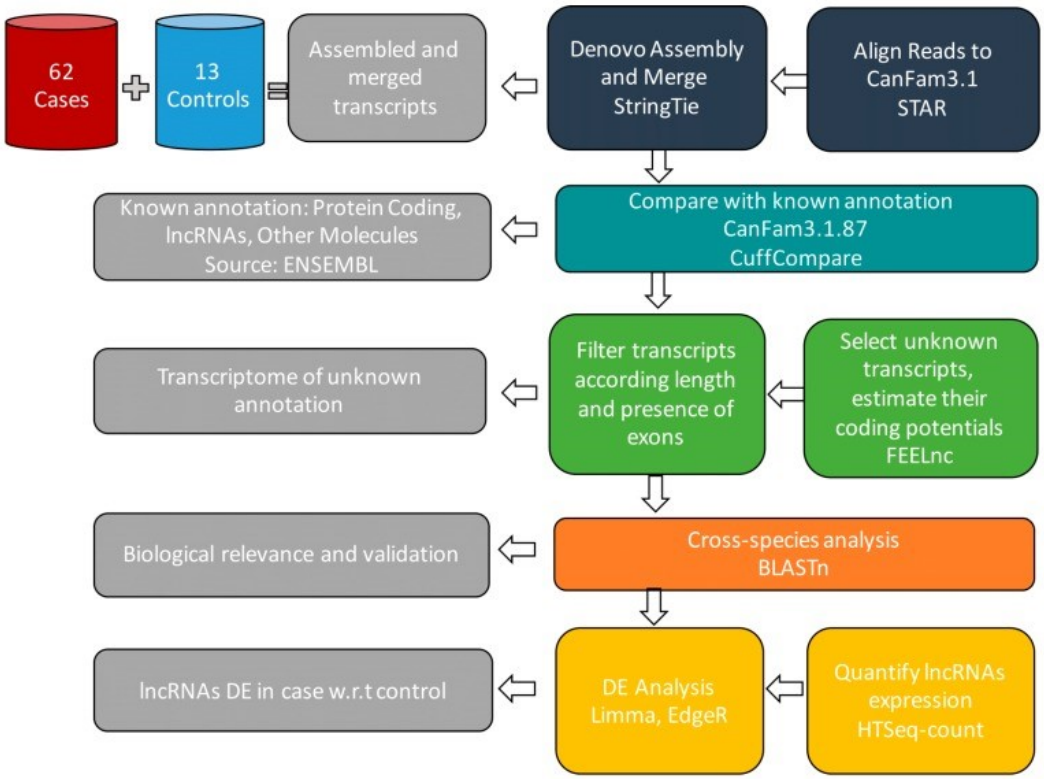


Fig.8.2 Diagram of analysis from the high-throughput single-molecule raw quantitative information of transcripts to differentially expressed molecules and pathways. High-throughput reads (units read by the sequencing machine about the patient's DNA sequence) are first pre-processed in order to assess the quality of the sequences. Reads are aligned using the canine reference genome (CanFam3.1.87), assembled in de novo mode to find novel and annotated transcripts, and then merged. The consensus patient's transcriptome is compared to CanFam3.1.87. The comparison step detects all the protein coding genes, annotated lncRNAs, and other transcripts. The annotated transcripts are then filtered out maintaining only the ones flagged as unknown and considered as potential novel lncRNAs. The resulting transcriptome is converted into a sequence file and analysed to estimate the coding potential of transcripts. This step produced a list of all the transcripts associated to their corresponding coding potential and the cutoff that separates coding from non-coding transcripts. Finally, transcripts are quantified and analysed by differential expression and pathway analysis.

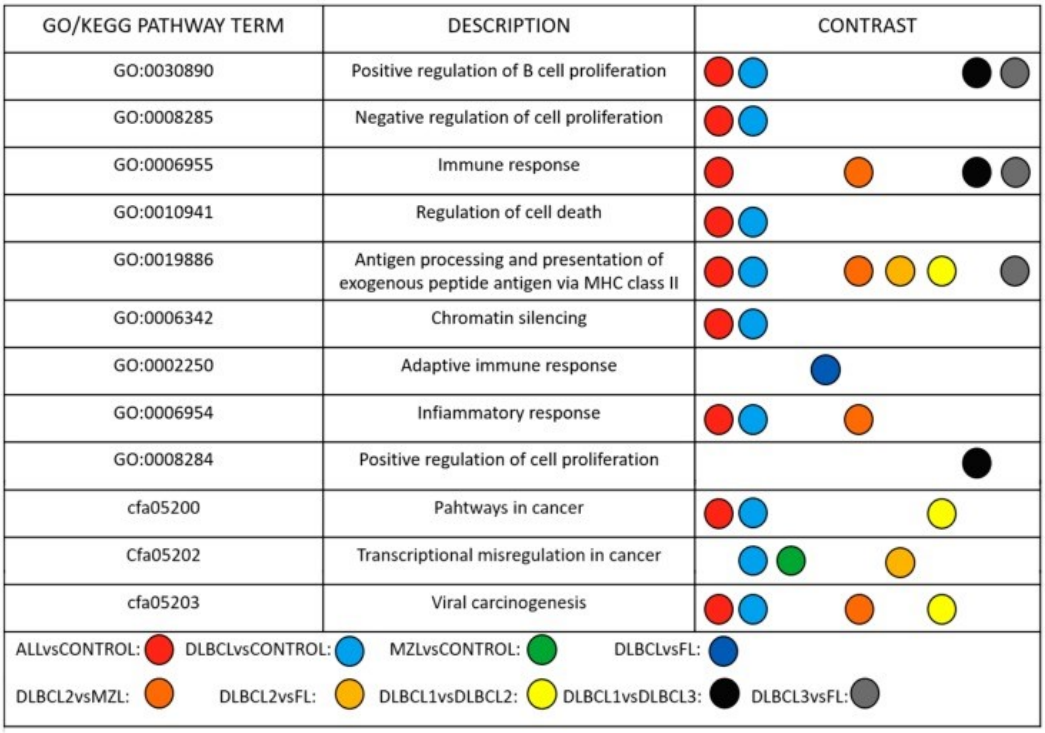


Fig.8.3 Potential functions and putative pathways of deregulated lncRNAs in the comparisons of interest. Each colored circle represents a comparison and, if present in one row, shows the significant enrichment of the modulated lncRNAs for that comparison for the specific GO term or KEGG pathway.

netDx for lymphoma study

Integrated analysis of transcriptome, methylome and copy number aberrations data of marginal zone lymphoma and follicular lymphoma in dog [69]

We characterized marginal zone lymphoma (MZL) and follicular lymphoma (FL) in dogs based on three different omics. Here the problem was about: how to integrate three omics for obtaining interpretable results which could help to characterize the classes in study. We solved this task using a netDx which exploits the patient similarity network (PSN) paradigm. A PSN models pathway-specific pairwise similarities as edges between nodes representing patients (i.e. the dogs in study). Defined a problem as classifying two patient classes (i.e., MZL vs FL) the method takes all the available data and converts them into PSNs. The networks describing the patient's similarities in the same pathway by different omics are merged with a network integration algorithm. The method tests the ability of the

consensus network to classify the patients, so it runs a label propagation algorithm to predict the class of unknown patients which have been previously excluded in a cross-validation loop. The software returns the PSNs representing pathways which best allowed to discriminate and classify the two classes. We used this feature to perform the analog of a gene set enrichment analysis. A total of 8 pathways have been found predictive of the FL group and the majority were related to cell metabolism and gap junction's dysregulation. In MZL, a total of 60 pathways were found predictive and both TP53 pathway with the related FoxM1 network and the TLR associated TICAM1-dependent IRF3 and IRF7 activation pathway were identified. Furthermore, since the software tests its ability to distinguish the classes in comparison, unexpectedly, the union of three omics still confirmed a poor distinction between MZL and DLBCL (AUROC = 0.64), whereas FL and DLBCL resulted clearly separated (AUROC = 0.83).

The analysis and results can be replicated with the code provide in the following repository: https://github.com/LucaGiudice/Multiomics-netDx-classification

netDx with Pratic

netDx: Software for building interpretable patient classifiers by multi-'omic data integration using patient similarity networks. [62]

We developed the software side that exploits a gene-gene interaction network to handle sparse somatic mutation data of patients. The context has been the following one: Patient classification has widespread biomedical and clinical applications, including treatment of diseases and prevention in precision medicine. However, standard classification methods do not natively support highly sparse data, such as genome-wide mutations. For instance, correlation can often not be computed for sparse vectors (too few overlaps). Methods, such as “network-based stratification (NBS)”, have been recently developed that use prior knowledge of gene interactions, combined with network propagation to address the sparse data challenge. We have extended this idea to additionally use pathway information and explore its use in patient classification. The approach is composed by two steps. In the first step, a binary matrix of patient-gene mutations is smoothed using known gene-gene interactions. Gene-level mutations are overlaid on a gene interaction

network and patient mutations are smoothed using label propagation. The smoothed patient profiles are used to create patient similarity networks (each one representing a specific biological process). These will be used as features to classify, and they undergo a filtering step based on a nonparametric statistics approach. This starts building a synthetic PSN which would not allow to classify the patients and then compares the real similarity networks to it. More the PSNs are different to the negative model and more they have a potential to discriminate the different patient classes based on their topology. The best PSNs are selected and provided to the supervised learning algorithm called netDx. Using somatic mutations from seven cancer types from the Cancer Genome Atlas (N ranging from 169 to 430), we found that representing pathways as PSNs which similarities are computed between smoothed profiles improves classification of patient survival.

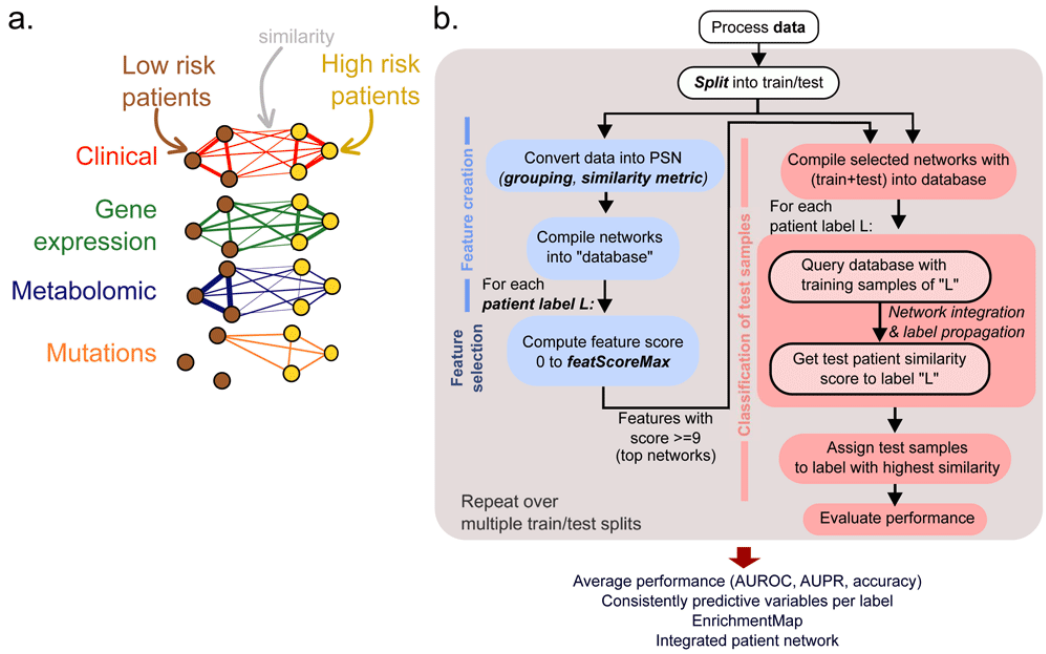


Fig.8.4. netDx workflow. (a) Conceptual visualization of patient similarity networks. Nodes are patients and edge weights measure pairwise similarity. The example shows a two-class problem (high and low risk patients), with four features shown as patient similarity networks: similarity for clinical (red), gene expression (green), metabolomic (blue) and mutation (orange) data. (b) Conceptual workflow for netDx predictor. Samples are split into train and test samples, and training samples are subjected to feature selection (blue flow). Feature selection uses regularized regression to integrate networks, such that networks with non-zero regression weights have their feature score increased. This process is repeated with different subsamples of training data, for a user-provided maximum of times . This process is repeated for each patient label. Features passing a user-specified threshold are used to classify held-out samples. Test patients are classified by highest similarity.

Patient networks that combine training and test patients are then integrated; only networks from features passing selection are used for this step. Label propagation is used to compute similarity of each test patient to training samples from each label; a given patient is assigned to the class with highest similarity. Average model performance is computed by running this whole process over several train/test splits. Features with consistent high scores can be used to classify an independent validation set.

Piezo1

"Time window" effect of Yoda1-evoked Piezo1 channel activity during mouse skeletal muscle differentiation [149]

Mechanosensitive Piezo1 ion channels emerged recently as important contributors to various vital functions including modulation of the blood supply to skeletal muscles. The specific Piezo1 channel agonist Yoda1 was shown to regulate the tone of blood vessels similarly to physical exercise. However, the direct role of Piezo1 channels in muscle function has been little studied so far. For annotating our marker in muscle cells, we collected the transcriptome data from Dell'Orso et al. [150] describing satellite cells and primary myoblasts isolated from homeostatic or regenerating muscles by single-cell RNA- sequencing. We replicated the analysis described in the section "Data processing and clustering" to get the single cell populations identified by the authors. Next, we analysed Piezo1 frequency and expression in the dataset. For each cell type in study, we determined the number of cells expressing a gene, assessed its rank with respect to the counts of all other genes (i.e., rank 1 for the lowest abundant gene) and normalized the ranks by the number of genes in analysis (i.e. ratio of 0 for the cell-type specific lowest abundant gene, while 1 for the gene expressed in the highest number of cells). We repeated the same approach also with the gene normalized level of expression (e.g., ratio of 1 for the most expressed gene in a cell type). Our analysis showed that Piezo1 was more frequently found in muscular stem cell close to quiescence cells (MuSc cQ) compared to 96.54% of all the other genes present in the same cell type, in muscular skeletal stem cell early activation cells (MuSc eA) more than the 92.28% and in cells positive to myogenin (MyoG) more than 77.38%, confirming an abundant expression of Piezo1 in muscle progenitors and myocytes in vivo conditions. Apart from the frequency, Piezo1 revealed to be also strongly expressed. In MuSc cQ cell

type, Piezo1 has an expression level higher than the 98.10% of other genes present in the same cells, in MuSc eA cells higher than than the 94.35% and in MyoG positive myocytes 79.03%. This confirmed the importance of our marker in muscle cell types.

LErNet

LErNet: characterization of lncRNAs via context-aware network expansion and enrichment analysis [151]

LErNet is a method able to define and predict the function of lncRNAs based on the coding genes that are near them and compose what we call the lncRNA genomic context. In details, the method requires the set of functionally unknown lncRNAs, and a set of protein-coding genes associated to the phenotype of interest (the same which is assumed to be related with the lncRNAs in study). LErNet starts computing the genomic context of the lncRNAs; coding genes within the range of 10Kb from each lncRNA are kept and the other ones are discarded. Next, the method creates an undirected sparse graph; nodes are genes and an edge between two nodes is a physical interaction between them which appears in the STRING database. Once the network is created, the method begins an iterative expansion procedure. First, it extracts the neighbours of the genes in the graph using STRING. Then, it adds to the graph the minimum number of neighbours which maximize the connectivity of the network. This operation makes the network dense, and it allows to improve the function information of the starting genes composing the context which are often very few (it depends by the phenotype in study). After the expansion procedure, the method does a pathway analysis of the genes in the graph for detecting the biological pathways associated to the seed lncRNAs. This approach simulates a real biological lncRNA feature. In fact, lncRNAs are biologically associated to their neighbouring coding genes by regulating their expression. The results showed that LErNet outperforms previously published enrichment methods in literature and to be robust in case of only few genomic context elements.

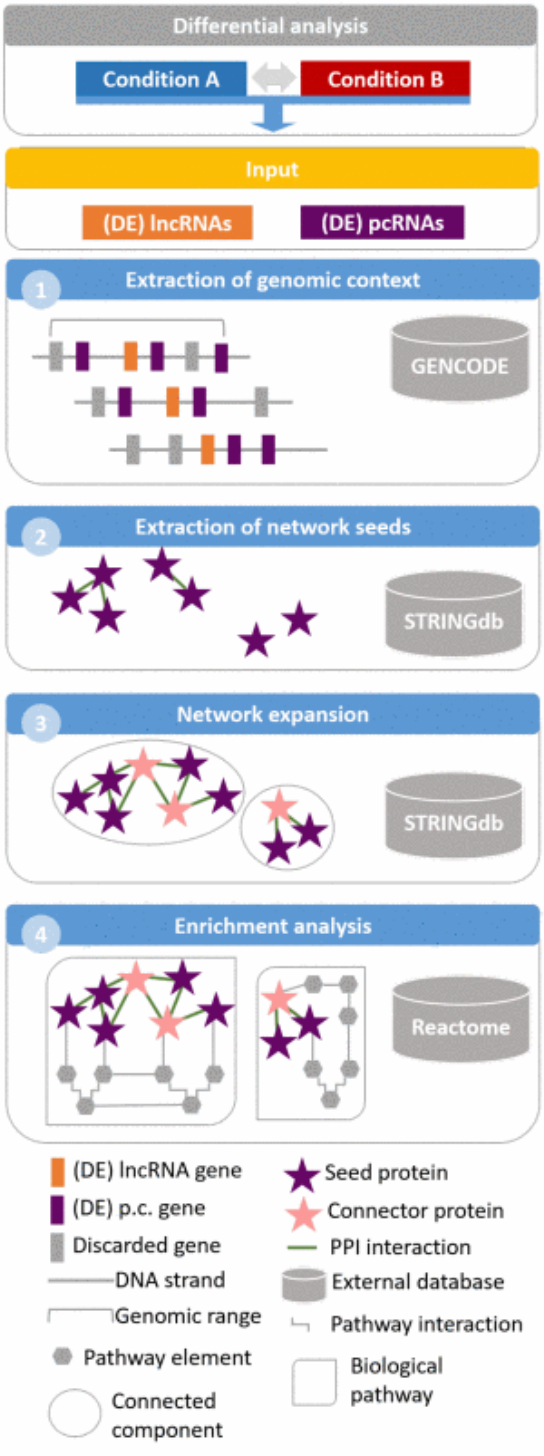


Fig. 8.5. Workflow of the software

The method is available with the code provide in the following repository: https://github.com/InfOmics/LErNet

Esearch3D

Esearch3D: Propagating gene expression in chromatin networks to illuminate active enhancers

Signals pertaining to transcriptional activation are transferred from enhancers to genes in the form of transcription factors, cofactors, and various transcriptional machineries such as RNA Pol II. How and where this information is transmitted to and from is central for decoding the regulatory landscape of any gene and identifying enhancers. Esearch3D is an unsupervised algorithm to predict enhancers. It reverses engineering the flow of information and identifies regulatory enhancers using solely gene expression and 3D genomic data.

The information about how genes, enhancers and other non-coding molecules are connected inside a cell is obtained from chromosome conformation capture (3C) technology. However, the latter does not provide indirect connections making the detection of relevant enhancers complicated due to the interconnected nature of the molecules and does not include a quantitative information about the activity of an enhancer.

We implemented Esearch3D to solve the question: “given a gene expression profile associated to a phenotype of interest, which enhancer has participated in the regulation of the expressed genes? and how much has participated?”. The method creates a chromatin interaction network (CIN) and then exploits the network-based propagation to find active enhancers.

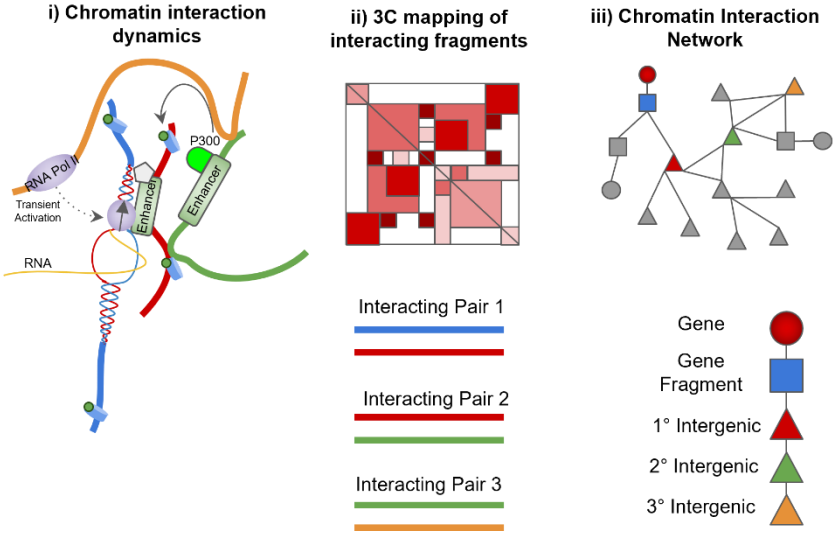


Fig.8.6 Schematic of converting chromatin dynamics into networks. i) Enhancers are localised and can influence the regulation of promoters (regions of interaction with a specific gene represented with a blue square node) directly and indirectly. ii) 3C capture chromatin interactions but only in a pairwise manner. iii) Representation of chromatin fragments as nodes in a network and their interactions as edges preserves indirect associations between interacting chromatin regions.

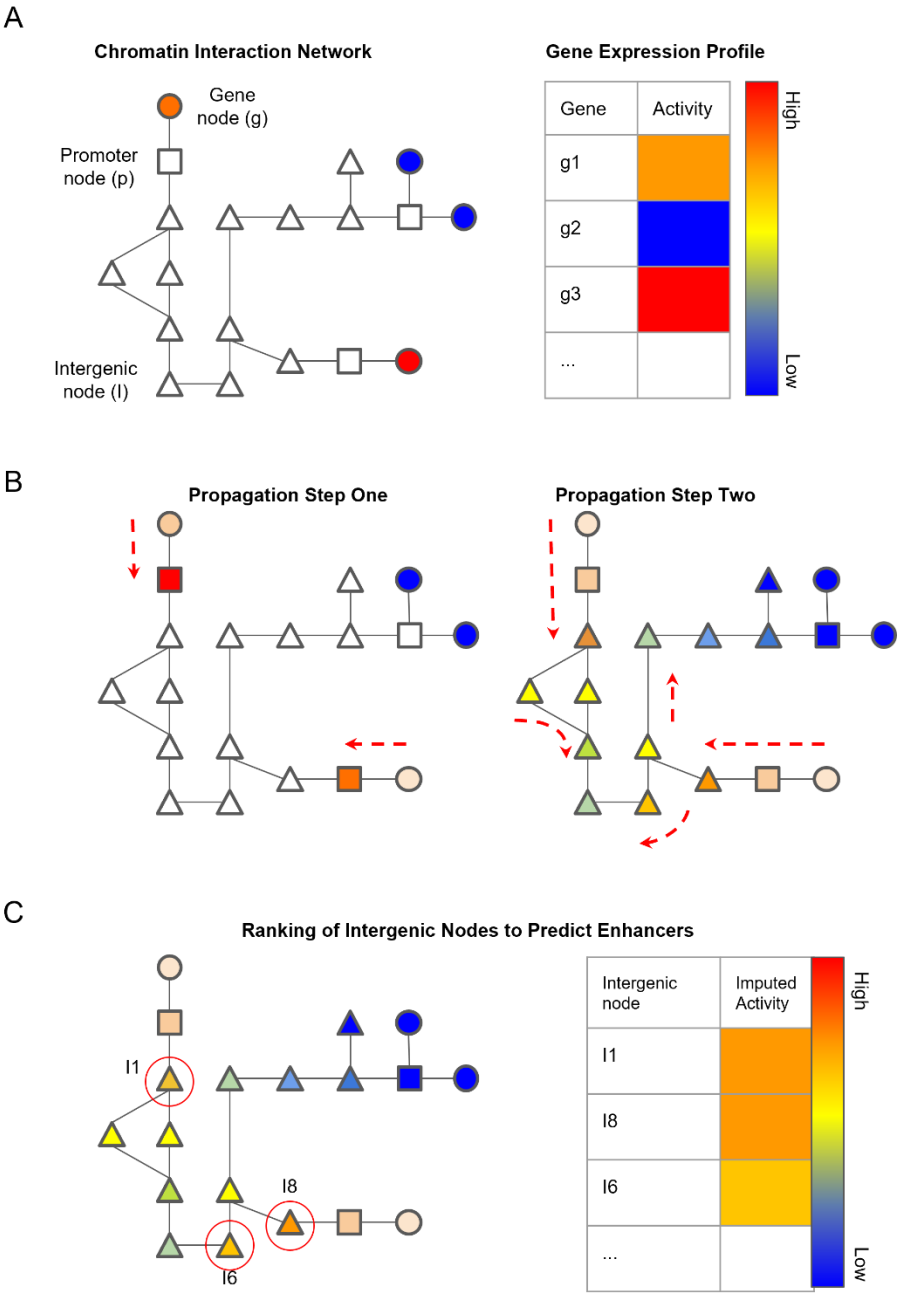


Fig.8.7 *Schematic diagram of the network propagation used to impute activity values at intergenic nodes. A) Genes are mapped to nodes representing genic chromatin fragments. Each gene has an associated gene activity value determined by RNA-seq data. B) Gene activity is propagated from*

gene nodes to genic chromatin nodes in propagation step one. Activity scores are then imputed in intergenic chromatin nodes by propagating the scores from genic chromatin nodes. C) Ranking of non-genic nodes by the imputed activity score to identify high confidence enhancer nodes.

The method is under submission and available with the code provide in the following repository: https://github.com/LucaGiudice/Esearch3D

Co-LCNEC

Lung combined large-cell neuroendocrine carcinomas (Co-LCNEC): more than a hybrid neoplasm

We studied the molecular pathogenesis of combined neuroendocrine and non-neuroendocrine cancers using the bulk RNA sequencing data of patients affected by Combined Large Cell Neuroendocrine Carcinomas (CoLCNEC). We integrated patients from different studies performing normalization techniques. We combined clustering approaches to identify 7 histological cancer subtypes (24 AC, 40 ADC, 35 LCNEC, 8 Control, 21 SCLC, 33 SQC, 21 TC) and annotated each subtype by differential expression and pathway analysis. The latter has been used also to explain three new subtypes called CL4, 7 and 9.

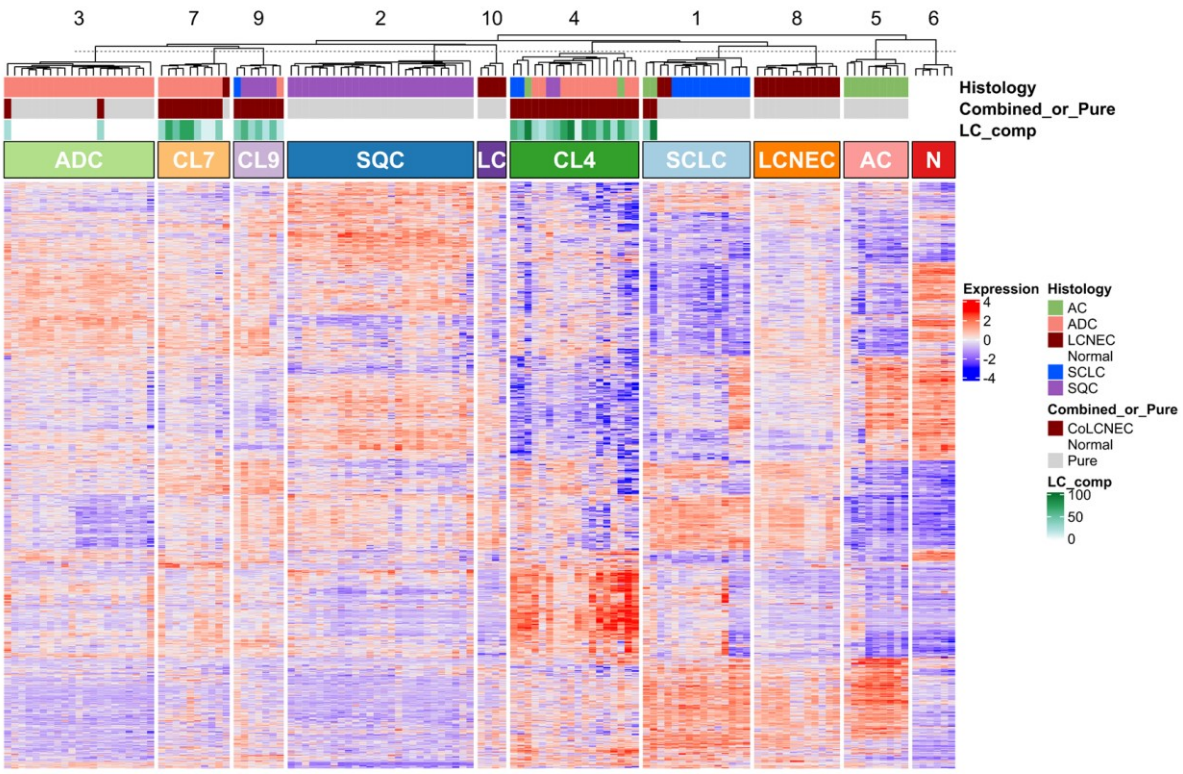


Fig.8.8 Heatmap of the clusters differentiated as the single molecular level. Rows are genes, columns are patients. A cell contains the expression value of a gene in a patient.

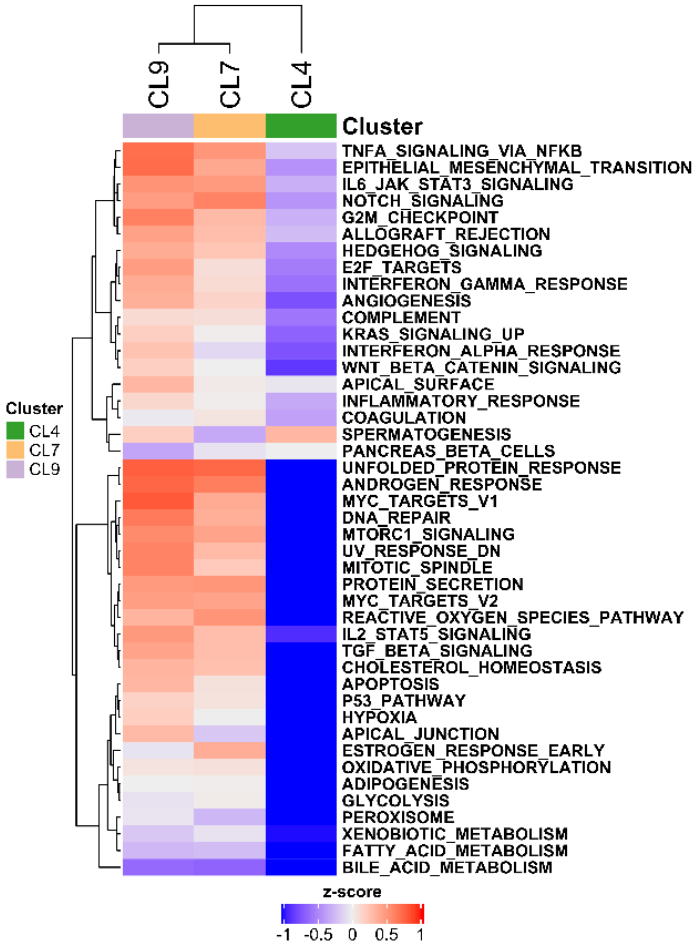


Fig.8.9 Heatmap of the unknown disease subtypes and clusters differentiated as the pathway level. Rows are pathways, columns are clusters of patients. A cell contains the level of alteration of a pathway in a patient group.

The study is in submission and the code to replicate the analysis is provided in the following repository: https://github.com/LucaGiudice/Supplementary-coLCNEC.

9. BIBLIOGRAPHY

[1] Klipp E, Liebermeister W, Wierling C, Kowald A. Systems Biology: A Textbook, 2nd Edition. 2016.

[2] Ramon y Cajal S, Segura M, Huemmer S. Interplay Between ncRNAs and Cellular Communication: A Proposal for Understanding Cell-Specific Signaling Pathways. Frontiers in Genetics 2019;10. https://doi.org/10.3389/fgene.2019.00281.

[3] Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biology 2017;18:83. https://doi.org/10.1186/s13059-017-1215-1.

[4] Horgan RP, Kenny LC. 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. The Obstetrician & Gynaecologist 2011;13:189–95. https://doi.org/10.1576/toag.13.3.189.27672.

[5] Shi X-J, Wei Y, Ji B. Systems Biology of Gastric Cancer: Perspectives on the Omics-Based Diagnosis and Treatment. Frontiers in Molecular Biosciences 2020;7:203. https://doi.org/10.3389/fmolb.2020.00203.

[6] Braun R. Systems analysis of high-throughput data. Adv Exp Med Biol 2014;844:153–87. https://doi.org/10.1007/978-1-4939-2095-2_8.

[7] Perkins AD, Tanentzapf G. An Ongoing Role for Structural Sarcomeric Components in Maintaining Drosophila melanogaster Muscle Function and Structure. PLOS ONE 2014;9:e99362. https://doi.org/10.1371/journal.pone.0099362.

[8] Tkacik G, Bialek W. Cell biology: Networks, regulation, pathways. ArXiv:07124385 [q-Bio] 2007.

[9] Zhang B, Tian Y, Zhang Z. Network Biology in Medicine and Beyond. Circ Cardiovasc Genet 2014;7:536–47. https://doi.org/10.1161/CIRCGENETICS.113.000123.

[10] Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 2009;37:D412-416. https://doi.org/10.1093/nar/gkn760.

[11] Biological Pathways Fact Sheet. GenomeGov n.d. https://www.genome.gov/about-genomics/fact-sheets/Biological-Pathways-Fact-Sheet (accessed December 3, 2021).

[12] Ochsner SA, Abraham D, Martin K, Ding W, McOwiti A, Kankanamge W, et al. The Signaling Pathways Project, an integrated 'omics knowledgebase for mammalian cellular signaling pathways. Sci Data 2019;6:252. https://doi.org/10.1038/s41597-019-0193-4.

[13] p53 Pathway Antibodies n.d. https://rockland-inc.com/p53-pathway.aspx (accessed December 3, 2021).

[14] Mostafavi S, Goldenberg A, Morris Q. Labeling Nodes Using Three Degrees of Propagation. PloS One 2012;7:e51947. https://doi.org/10.1371/journal.pone.0051947.

[15] Mason O, Verwoerd M. Graph theory and networks in Biology. IET Syst Biol 2007;1:89–119. https://doi.org/10.1049/iet-syb:20060038.

[16] Koutrouli M, Karatzas E, Paez-Espino D, Pavlopoulos GA. A Guide to Conquer the Biological Network Era Using Graph Theory. Front Bioeng Biotechnol 2020;8:34. https://doi.org/10.3389/fbioe.2020.00034.

[17] Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, et al. Using graph theory to analyze biological networks. BioData Min 2011;4:10. https://doi.org/10.1186/1756-0381-4-10.
[18] Frank KA. Identifying cohesive subgroups. Social Networks 1995;17:27–56. https://doi.org/10.1016/0378-8733(94)00247-8.
[19] Freeman LC. Segregation in Social Networks. Sociological Methods & Research 1978;6:411–29. https://doi.org/10.1177/004912417800600401.
[20] Wasserman S, Faust K, Urbana-Champaign) S (University of IW. Social Network Analysis: Methods and Applications. Cambridge University Press; 1994.
[21] Seidman S, Foster B. A graph-theoretic generalization of the clique concept*. Journal of Mathematical Sociology 1978;6:139–54. https://doi.org/10.1080/0022250X.1978.9989883.
[22] Mokken RJ. Cliques, clubs and clans 1977. https://doi.org/10.1007/BF00139635.
[23] Roistacher RC. A Review of Mathematical Methods in Sociometry. Sociological Methods & Research 1974;3:123–71. https://doi.org/10.1177/004912417400300201.
[24] Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. Software: Practice and Experience 1991;21:1129–64. https://doi.org/10.1002/spe.4380211102.
[25] Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nature Reviews Genetics 2017;18:551–62. https://doi.org/10.1038/nrg.2017.38.
[26] Chen Y, Liu L. Chen, Y, and Liu, L. Modern methods for delivery of drugs across the blood-brain barrier. Adv Drug Deliv Rev 64: 640-665. Advanced Drug Delivery Reviews 2011;64:640–65. https://doi.org/10.1016/j.addr.2011.11.010.
[27] Le D-H. A novel method for identifying disease associated protein complexes based on functional similarity protein complex networks. Algorithms Mol Biol 2015;10:14. https://doi.org/10.1186/s13015-015-0044-6.
[28] Navlakha S, Kingsford C. The power of protein interaction networks for associating genes with diseases. Bioinformatics 2010;26:1057–63. https://doi.org/10.1093/bioinformatics/btq076.
[29] Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machine learning in bioinformatics. Briefings in Bioinformatics 2006;7:86–112. https://doi.org/10.1093/bib/bbk007.
[30] Auslander N, Gussow AB, Koonin EV. Incorporating Machine Learning into Established Bioinformatics Frameworks. International Journal of Molecular Sciences 2021;22:2903. https://doi.org/10.3390/ijms22062903.
[31] Liu H, Motoda H. Feature Selection for Knowledge Discovery and Data Mining. USA: Kluwer Academic Publishers; 1998.
[32] Narendra P, Fukunaga K. A Branch and Bound Algorithm for Feature Subset Selection. IEEE Transactions on Computers 1977. https://doi.org/10.1109/TC.1977.1674939.
[33] Pudil P, Novovičová J, Kittler J. Floating search methods in feature selection. Pattern Recognition Letters 1994;15:1119–25. https://doi.org/10.1016/0167-8655(94)90127-9.

[34] Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence 1997;97:273–324. https://doi.org/10.1016/S0004-3702(97)00043-X.
[35] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol 2016;17:13. https://doi.org/10.1186/s13059-016-0881-8.
[36] Koch CM, Chiu SF, Akbarpour M, Bharat A, Ridge KM, Bartom ET, et al. A Beginner's Guide to Analysis of RNA Sequencing Data. Am J Respir Cell Mol Biol 2018;59:145–57. https://doi.org/10.1165/rcmb.2017-0430TR.
[37] Geraci F, Saha I, Bianchini M. Editorial: RNA-Seq Analysis: Methods, Applications and Challenges. Front Genet 2020;11. https://doi.org/10.3389/fgene.2020.00220.
[38] Fabris F, Palmer D, de Magalhães JP, Freitas AA. Comparing enrichment analysis and machine learning for identifying gene properties that discriminate between gene classes. Brief Bioinform 2020;21:803–14. https://doi.org/10.1093/bib/bbz028.
[39] Differential expression analysis for sequence count data | Genome Biology | Full Text n.d. https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r106 (accessed December 5, 2021).
[40] McDermaid A, Monier B, Zhao J, Liu B, Ma Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. Brief Bioinform 2019;20:2044–54. https://doi.org/10.1093/bib/bby067.
[41] Maleki F, Ovens K, Hogan DJ, Kusalik AJ. Gene Set Analysis: Challenges, Opportunities, and Future Research. Front Genet 2020;11:654. https://doi.org/10.3389/fgene.2020.00654.
[42] Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol 2012;8:e1002375. https://doi.org/10.1371/journal.pcbi.1002375.
[43] Das S, McClain CJ, Rai SN. Fifteen Years of Gene Set Analysis for High-Throughput Genomic Data: A Review of Statistical Approaches and Future Challenges. Entropy 2020;22:427. https://doi.org/10.3390/e22040427.
[44] The Gene Ontology resource: enriching a GOld mine - PubMed n.d. https://pubmed.ncbi.nlm.nih.gov/33290552/ (accessed May 31, 2021).
[45] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30. https://doi.org/10.1093/nar/28.1.27.
[46] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50. https://doi.org/10.1073/pnas.0506580102.
[47] Tamayo P, Steinhardt G, Liberzon A, Mesirov JP. The limitations of simple gene set enrichment analysis assuming gene independence. Stat Methods Med Res 2016;25:472–87. https://doi.org/10.1177/0962280212460441.
[48] Kim S-Y, Volsky DJ. PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics 2005;6:144. https://doi.org/10.1186/1471-2105-6-144.
[49] Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 2007;23:980–7. https://doi.org/10.1093/bioinformatics/btm051.

[50] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44–57. https://doi.org/10.1038/nprot.2008.211.
[51] Vêncio RZ, Koide T, Gomes SL, de B Pereira CA. BayGO: Bayesian analysis of ontology term enrichment in microarray data. BMC Bioinformatics 2006;7:86. https://doi.org/10.1186/1471-2105-7-86.
[52] Zhang S, Cao J, Kong YM, Scheuermann RH. GO-Bayes: Gene Ontology-based overrepresentation analysis using a Bayesian approach. Bioinformatics 2010;26:905–11. https://doi.org/10.1093/bioinformatics/btq059.
[53] Bauer S, Gagneur J, Robinson PN. GOing Bayesian: model-based gene set analysis of genome-scale data. Nucleic Acids Res 2010;38:3523–32. https://doi.org/10.1093/nar/gkq045.
[54] Sass S, Buettner F, Mueller NS, Theis FJ. A modular framework for gene set analysis integrating multilevel omics data. Nucleic Acids Res 2013;41:9622–33. https://doi.org/10.1093/nar/gkt752.
[55] Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician 2016;70:129–33. https://doi.org/10.1080/00031305.2016.1154108.
[56] Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009;37:1–13. https://doi.org/10.1093/nar/gkn923.
[57] Freitas AA, Wieser DC, Apweiler R. On the importance of comprehensible classification models for protein function prediction. IEEE/ACM Trans Comput Biol Bioinform 2010;7:172–82. https://doi.org/10.1109/TCBB.2008.47.
[58] Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-Generation Machine Learning for Biological Networks. Cell 2018;173:1581–92. https://doi.org/10.1016/j.cell.2018.05.015.
[59] Alonso-Betanzos A, Bolón-Canedo V. Big-Data Analysis, Cluster Analysis, and Machine-Learning Approaches. Adv Exp Med Biol 2018;1065:607–26. https://doi.org/10.1007/978-3-319-77932-4_37.
[60] Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications 2017;73:220–39. https://doi.org/10.1016/j.eswa.2016.12.035.
[61] Pai S, Hui S, Isserlin R, Shah MA, Kaka H, Bader GD. netDx: interpretable patient classification using integrated patient similarity networks. Mol Syst Biol 2019;15. https://doi.org/10.15252/msb.20188497.
[62] Pai S, Weber P, Isserlin R, Kaka H, Hui S, Shah MA, et al. netDx: Software for building interpretable patient classifiers by multi-'omic data integration using patient similarity networks. F1000Res 2021;9:1239. https://doi.org/10.12688/f1000research.26429.2.
[63] Giudice L. Simpati: patient classifier identifies signature pathways based on similarity networks for the disease prediction. 2021. https://doi.org/10.1101/2021.09.23.461100.
[64] Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting

gene function. Genome Biology 2008;9:S4. https://doi.org/10.1186/gb-2008-9-s1-s4.
[65] Morvan ML, Zinovyev A, Vert J-P. NetNorM: Capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. PLOS Computational Biology 2017;13:e1005573. https://doi.org/10.1371/journal.pcbi.1005573.
[66] Ioffe S. Improved Consistent Sampling, Weighted Minhash and L1 Sketching. 2010 IEEE International Conference on Data Mining, 2010, p. 246–55. https://doi.org/10.1109/ICDM.2010.80.
[67] Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 2002;30:1575–84. https://doi.org/10.1093/nar/30.7.1575.
[68] Hearst MA, Pedersen E, Patil L, Lee E, Laskowski P, Franconeri S. An Evaluation of Semantically Grouped Word Cloud Designs. IEEE Transactions on Visualization and Computer Graphics 2020;26:2748–61. https://doi.org/10.1109/TVCG.2019.2904683.
[69] Giannuzzi D, Giudice L, Marconato L, Ferraresso S, Giugno R, Bertoni F, et al. Integrated analysis of transcriptome, methylome and copy number aberrations data of marginal zone lymphoma and follicular lymphoma in dog. Vet Comp Oncol 2020;18:645–55. https://doi.org/10.1111/vco.12588.
[70] Giudice L, Cascione L, Ferraresso S, Marconato L, Giannuzzi D, Napoli S, et al. Long Non-Coding RNAs as Molecular Signatures for Canine B-Cell Lymphoma Characterization. Noncoding RNA 2019;5. https://doi.org/10.3390/ncrna5030047.
[71] Kolosowska N, Gotkiewicz M, Dhungana H, Giudice L, Giugno R, Box D, et al. Intracerebral overexpression of miR-669c is protective in mouse ischemic stroke model by targeting MyD88 and inducing alternative microglial/macrophage activation. J Neuroinflammation 2020;17:194. https://doi.org/10.1186/s12974-020-01870-w.
[72] Jin L, Zuo X-Y, Su W-Y, Zhao X-L, Yuan M-Q, Han L-Z, et al. Pathway-based analysis tools for complex diseases: a review. Genomics Proteomics Bioinformatics 2014;12:210–20. https://doi.org/10.1016/j.gpb.2014.10.002.
[73] Drier Y, Sheffer M, Domany E. Pathway-based personalized analysis of cancer. PNAS 2013;110:6388–93. https://doi.org/10.1073/pnas.1219651110.
[74] Segura-Lepe MP, Keun HC, Ebbels TMD. Predictive modelling using pathway scores: robustness and significance of pathway collections. BMC Bioinformatics 2019;20:543. https://doi.org/10.1186/s12859-019-3163-0.
[75] Raghavan N, Amaratunga D, Cabrera J, Nie A, Qin J, McMillian M. On methods for gene function scoring as a means of facilitating the interpretation of microarray results. J Comput Biol 2006;13:798–809. https://doi.org/10.1089/cmb.2006.13.798.
[76] Lee E, Chuang H-Y, Kim J-W, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Comput Biol 2008;4:e1000217. https://doi.org/10.1371/journal.pcbi.1000217.
[77] Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006;439:353–7. https://doi.org/10.1038/nature04296.

[78] Mitrea C, Taghavi Z, Bokanizad B, Hanoudi S, Tagett R, Donato M, et al. Methods and approaches in the topology-based analysis of biological pathways. Front Physiol 2013;4:278. https://doi.org/10.3389/fphys.2013.00278.
[79] Xie C, Jauhari S, Mora A. Popularity and performance of bioinformatics software: the case of gene set analysis. BMC Bioinformatics 2021;22:191. https://doi.org/10.1186/s12859-021-04124-5.
[80] Pang H, Datta D, Zhao H. Pathway analysis using random forests with bivariate node-split for survival outcomes. Bioinformatics 2010;26:250–8. https://doi.org/10.1093/bioinformatics/btp640.
[81] Hao J, Kim Y, Kim T-K, Kang M. PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data. BMC Bioinformatics 2018;19:510. https://doi.org/10.1186/s12859-018-2500-z.
[82] Pai S, Bader GD. Patient Similarity Networks for Precision Medicine. J Mol Biol 2018;430:2924–38. https://doi.org/10.1016/j.jmb.2018.05.037.
[83] Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nature Methods 2014;11:333–7. https://doi.org/10.1038/nmeth.2810.
[84] Li L, Cheng W-Y, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Science Translational Medicine 2015;7:311ra174-311ra174. https://doi.org/10.1126/scitranslmed.aaa9364.
[85] McGuire AL, Fisher R, Cusenza P, Hudson K, Rothstein MA, McGraw D, et al. Confidentiality, privacy, and security of genetic and genomic test information in electronic health records: points to consider. Genetics in Medicine 2008;10:495–9. https://doi.org/10.1097/GIM.0b013e31817a8aaa.
[86] Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying Personal Genomes by Surname Inference. Science 2013;339:321–4. https://doi.org/10.1126/science.1229566.
[87] Chen F, Wang S, Jiang X, Ding S, Lu Y, Kim J, et al. PRINCESS: Privacy-protecting Rare disease International Network Collaboration via Encryption through Software guard extensionS. Bioinformatics 2017;33:871–8. https://doi.org/10.1093/bioinformatics/btw758.
[88] Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. Nature 2018;553:446–54. https://doi.org/10.1038/nature25183.
[89] Lu C, Bera K, Wang X, Prasanna P, Xu J, Janowczyk A, et al. A prognostic model for overall survival of patients with early-stage non-small cell lung cancer: a multicentre, retrospective study. Lancet Digit Health 2020;2:e594–606. https://doi.org/10.1016/s2589-7500(20)30225-9.
[90] Shridhar V, Lee J, Pandita A, Iturria S, Avula R, Staub J, et al. Genetic analysis of early- versus late-stage ovarian tumors. Cancer Res 2001;61:5895–904.
[91] Promoting Cancer Early Diagnosis n.d. https://www.who.int/activities/promoting-cancer-early-diagnosis (accessed May 31, 2021).
[92] Hanna TP, King WD, Thibodeau S, Jalink M, Paulin GA, Harvey-Jones E, et al. Mortality due to cancer treatment delay: systematic review and meta-analysis. BMJ 2020;371:m4087. https://doi.org/10.1136/bmj.m4087.

[93] Raphael MJ, Biagi JJ, Kong W, Mates M, Booth CM, Mackillop WJ. The relationship between time to initiation of adjuvant chemotherapy and survival in breast cancer: a systematic review and meta-analysis. Breast Cancer Res Treat 2016;160:17–28. https://doi.org/10.1007/s10549-016-3960-3.
[94] Russell B, Liedberg F, Khan MS, Nair R, Thurairaja R, Malde S, et al. A Systematic Review and Meta-analysis of Delay in Radical Cystectomy and the Effect on Survival in Bladder Cancer Patients. Eur Urol Oncol 2020;3:239–49. https://doi.org/10.1016/j.euo.2019.09.008.
[95] Multiomic Integration of Public Oncology Databases in Bioconductor - PubMed n.d. https://pubmed.ncbi.nlm.nih.gov/33119407/ (accessed May 31, 2021).
[96] Ramos M, Schiffer L, Davis S, Waldron L. TCGAutils: TCGA utility functions for data management. TCGAutils: TCGA Utility Functions for Data Management 2021.
[97] Rosen RD, Sapra A. TNM Classification. StatPearls, Treasure Island (FL): StatPearls Publishing; 2021.
[98] Barclay ME, Abel GA, Greenberg DC, Rous B, Lyratzopoulos G. Socio-demographic variation in stage at diagnosis of breast, bladder, colon, endometrial, lung, melanoma, prostate, rectal, renal and ovarian cancer in England and its population impact. British Journal of Cancer 2021;124:1320–9. https://doi.org/10.1038/s41416-021-01279-z.
[99] Hu Z-D, Zhou Z-R, Qian S. How to analyze tumor stage data in clinical research. J Thorac Dis 2015;7:566–75. https://doi.org/10.3978/j.issn.2072-1439.2015.04.09.
[100] McCormack V, Aggarwal A. Early cancer diagnosis: reaching targets across whole populations amidst setbacks. British Journal of Cancer 2021;124:1181–2. https://doi.org/10.1038/s41416-021-01276-2.
[101] Law CW, Alhamdoosh M, Su S, Dong X, Tian L, Smyth GK, et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000Res 2016;5. https://doi.org/10.12688/f1000research.9005.3.
[102] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010;11:R25. https://doi.org/10.1186/gb-2010-11-3-r25.
[103] Beleites C, Neugebauer U, Bocklitz T, Krafft C, Popp J. Sample size planning for classification models. Analytica Chimica Acta 2013;760:25–33. https://doi.org/10.1016/j.aca.2012.11.007.
[104] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–9. https://doi.org/10.1038/75556.
[105] Oughtred R, Rust J, Chang C, Breitkreutz B-J, Stark C, Willems A, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci 2021;30:187–200. https://doi.org/10.1002/pro.3978.
[106] Le D-H. Random walk with restart: A powerful network propagation algorithm in Bioinformatics field. 2017 4th NAFOSTED Conference on Information and Computer Science, 2017, p. 242–7. https://doi.org/10.1109/NAFOSTED.2017.8108071.

[107] Ideker T, Sharan R. Protein networks in disease. Genome Res 2008;18:644–52. https://doi.org/10.1101/gr.071852.107.
[108] Vandin F, Clay P, Upfal E, Raphael BJ. Discovery of mutated subnetworks associated with clinical data in cancer. Pac Symp Biocomput 2012:55–66.
[109] Le D-H. Network-based ranking methods for prediction of novel disease associated microRNAs. Comput Biol Chem 2015;58:139–48. https://doi.org/10.1016/j.compbiolchem.2015.07.003.
[110] Shi H, Xu J, Zhang G, Xu L, Li C, Wang L, et al. Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes. BMC Syst Biol 2013;7:101. https://doi.org/10.1186/1752-0509-7-101.
[111] Sun J, Shi H, Wang Z, Zhang C, Liu L, Wang L, et al. Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. Mol Biosyst 2014;10:2074–81. https://doi.org/10.1039/c3mb70608g.
[112] Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods 2013;10:1108–15. https://doi.org/10.1038/nmeth.2651.
[113] Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. ArXiv:160902907 [Cs, Stat] 2017.
[114] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph Attention Networks. ArXiv:171010903 [Cs, Stat] 2018.
[115] Di Nanni N, Bersanelli M, Milanesi L, Mosca E. Network Diffusion Promotes the Integrative Analysis of Multiple Omics. Front Genet 2020;11. https://doi.org/10.3389/fgene.2020.00106.
[116] Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front Genet 2017;8. https://doi.org/10.3389/fgene.2017.00084.
[117] Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics 2016;17:S15. https://doi.org/10.1186/s12859-015-0857-9.
[118] Pak M, Jeong D, Moon JH, Ann H, Hur B, Lee S, et al. Network Propagation for the Analysis of Multi-omics Data. In: Yoon B-J, Qian X, editors. Recent Advances in Biological Network Analysis: Comparative Network Analysis and Network Module Detection, Cham: Springer International Publishing; 2021, p. 185–217. https://doi.org/10.1007/978-3-030-57173-3_9.
[119] Andl CD, Mizushima T, Oyama K, Bowser M, Nakagawa H, Rustgi AK. EGFR-induced cell migration is mediated predominantly by the JAK-STAT pathway in primary esophageal keratinocytes. Am J Physiol Gastrointest Liver Physiol 2004;287:G1227-1237. https://doi.org/10.1152/ajpgi.00253.2004.
[120] Badache A, Hynes NE. Interleukin 6 inhibits proliferation and, in cooperation with an epidermal growth factor receptor autocrine loop, increases migration of T47D breast cancer cells. Cancer Res 2001;61:383–91.
[121] Takahashi-Tezuka M, Yoshida Y, Fukada T, Ohtani T, Yamanaka Y, Nishida K, et al. Gab1 acts as an adapter molecule linking the cytokine receptor gp130 to ERK mitogen-activated protein kinase. Mol Cell Biol 1998;18:4109–17. https://doi.org/10.1128/MCB.18.7.4109.

[122] Marquand AF, Rezek I, Buitelaar J, Beckmann CF. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. Biological Psychiatry 2016;80:552–61. https://doi.org/10.1016/j.biopsych.2015.12.023.
[123] Kirk R. Personalized medicine and tumour heterogeneity. Nat Rev Clin Oncol 2012;9:250–250. https://doi.org/10.1038/nrclinonc.2012.46.
[124] Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 2012;366:883–92. https://doi.org/10.1056/NEJMoa1113205.
[125] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 1998;30:107–17. https://doi.org/10.1016/S0169-7552(98)00110-X.
[126] Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res 2017;45:D833–9. https://doi.org/10.1093/nar/gkw943.
[127] Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science 2017;357. https://doi.org/10.1126/science.aan2507.
[128] Lin L, Yang T, Fang L, Yang J, Yang F, Zhao J. Gene gravity-like algorithm for disease gene prediction based on phenotype-specific network. BMC Systems Biology 2017;11:121. https://doi.org/10.1186/s12918-017-0519-9.
[129] Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform 2005;38:404–15. https://doi.org/10.1016/j.jbi.2005.02.008.
[130] Pang S, Sun Y, Wu L, Yang L, Zhao Y-L, Wang Z, et al. Reconstruction of kidney renal clear cell carcinoma evolution across pathological stages. Sci Rep 2018;8:3339. https://doi.org/10.1038/s41598-018-20321-4.
[131] Cui H, Shan H, Miao MZ, Jiang Z, Meng Y, Chen R, et al. Identification of the key genes and pathways involved in the tumorigenesis and prognosis of kidney renal clear cell carcinoma. Sci Rep 2020;10:4271. https://doi.org/10.1038/s41598-020-61162-4.
[132] Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics 2009;10:161. https://doi.org/10.1186/1471-2105-10-161.
[133] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation 2021:100141. https://doi.org/10.1016/j.xinn.2021.100141.
[134] Christensen A P. NetworkToolbox: Methods and Measures for Brain, Cognitive, and Psychometric Network Analysis in R. The R Journal 2019;10:422. https://doi.org/10.32614/RJ-2018-065.
[135] Frigyesi A, Höglund M. Non-Negative Matrix Factorization for the Analysis of Complex Gene Expression Data: Identification of Clinically Relevant Tumor Subtypes. Cancer Inform 2008;6:275–92.
[136] Saccenti E, Hendriks MHWB, Smilde AK. Corruption of the Pearson correlation coefficient by measurement error and its estimation, bias, and

correction under different error models. Sci Rep 2020;10:438. https://doi.org/10.1038/s41598-019-57247-4.
[137] Powers S, DeJongh M, Best AA, Tintle NL. Cautions about the reliability of pairwise gene correlations based on expression data. Front Microbiol 2015;6. https://doi.org/10.3389/fmicb.2015.00650.
[138] McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, M. Mastrogianakis G, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455:1061–8. https://doi.org/10.1038/nature07385.
[139] Cavalli FMG, Remke M, Rampasek L, Peacock J, Shih DJH, Luu B, et al. Intertumoral Heterogeneity within Medulloblastoma Subgroups. Cancer Cell 2017;31:737-754.e6. https://doi.org/10.1016/j.ccell.2017.05.005.
[140] Cancer Genome Atlas Research Network. Electronic address: andrew_aguirre@dfci.harvard.edu, Cancer Genome Atlas Research Network. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. Cancer Cell 2017;32:185-203.e13. https://doi.org/10.1016/j.ccell.2017.07.007.
[141] Kohane IS. HEALTH CARE POLICY. Ten things we have to do to achieve precision medicine. Science 2015;349:37–8. https://doi.org/10.1126/science.aab1328.
[142] Garrido P, Aldaz A, Vera R, Calleja MA, de Álava E, Martín M, et al. Proposal for the creation of a national strategy for precision medicine in cancer: a position statement of SEOM, SEAP, and SEFH. Clin Transl Oncol 2018;20:443–7. https://doi.org/10.1007/s12094-017-1740-0.
[143] Witt H, Mack SC, Ryzhova M, Bender S, Sill M, Isserlin R, et al. Delineation of two clinically and molecularly distinct subgroups of posterior fossa ependymoma. Cancer Cell 2011;20:143–57. https://doi.org/10.1016/j.ccr.2011.07.007.
[144] Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61–70. https://doi.org/10.1038/nature11412.
[145] Mack SC, Witt H, Piro RM, Gu L, Zuyderduyn S, Stütz AM, et al. Epigenomic alterations define lethal CIMP-positive ependymomas of infancy. Nature 2014;506:445–50. https://doi.org/10.1038/nature13108.
[146] Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature 2017;550:204–13. https://doi.org/10.1038/nature24277.
[147] Friend SH, Ideker T. Point: Are we prepared for the future doctor visit? Nat Biotechnol 2011;29:215–8. https://doi.org/10.1038/nbt.1794.
[148] Pratt D, Chen J, Welker D, Rivas R, Pillich R, Rynkov V, et al. NDEx, the Network Data Exchange. Cell Syst 2015;1:302–5. https://doi.org/10.1016/j.cels.2015.10.001.
[149] Bosutti A, Giniatullin A, Odnoshivkina Y, Giudice L, Malm T, Sciancalepore M, et al. “Time window” effect of Yoda1-evoked Piezo1 channel activity during mouse skeletal muscle differentiation. Acta Physiologica n.d.;n/a:e13702. https://doi.org/10.1111/apha.13702.
[150] Dell’Orso S, Juan AH, Ko K-D, Naz F, Perovanovic J, Gutierrez-Cruz G, et al. Single cell analysis of adult mouse skeletal muscle stem cells in

homeostatic and regenerative conditions. Development 2019;146:dev174177. https://doi.org/10.1242/dev.174177.

[151] Bonnici V, Caligola S, Fiorini G, Giudice L, Giugno R. LErNet: characterization of lncRNAs via context-aware network expansion and enrichment analysis. 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2019, p. 1–8. https://doi.org/10.1109/CIBCB.2019.8791487.