UNIVERSITA' DEGLI STUDI DI SALERNO

DOTTORATO IN INFORMATICA E INGEGNERIA
DELL'INFORMAZIONE

CURRICULUM INFORMATICA

COORDINATORE: Ch.mo. Prof. Alfredo De Santis

Ciclo XV N.S.

Multi-View Learning and Data Integration for *omics* Data

Relatori
Ch.mo. Prof. Roberto Tagliaferri
Ch.mo. Prof. Dario Greco

Candidato
Angela Serra
Matr. 8888100002

ANNO ACCADEMICO 2015/2016

How to reach a goal?
Without haste but without rest
Goethe

Acknowledgements

During my PhD in the NeuRoNeLab I met wonderful friends whom I warmly thank: Luca, Paola, Massimo, thanks for the constant support, the countless discussions and above all for the entertainment of the past three years.

A big thank you goes to the the special people I met in Helsinki: they never let me feel alone! Many thanks to Vittorio, Veer, Marit and Pia.

A special thank goes to my family who always support me.

I really do want to thank my mentors, Prof. Roberto Tagliaferri and Prof. Dario Greco, for the patient, the constant support and for the valuable advice of the last three years.

Finally, the biggest thanks goes to Michele, without him I would not have come so far. ♥♥♥

Abstract

In recent years, the advancement of high-throughput technologies, combined with the constant decrease of the data-storage costs, has led to the production of large amounts of data from different experiments that characterise the same entities of interest. This information may relate to specific aspects of a phenotypic entity (e.g. Gene expression), or can include the comprehensive and parallel measurement of multiple molecular events (e.g., DNA modifications, RNA transcription and protein translation) in the same samples.

Exploiting such complex and rich data is needed in the frame of systems biology for building global models able to explain complex phenotypes. For example, the use of genome-wide data in cancer research, for the identification of groups of patients with similar molecular characteristics, has become a standard approach for applications in therapy-response, prognosis-prediction, and drug-development.ÂăMoreover, the integration of gene expression data regarding cell treatment by drugs, and information regarding chemical structure of the drugs allowed scientist to perform more accurate drug repositioning tasks.

Unfortunately, there is a big gap between the amount of information and the knowledge in which it is translated. Moreover, there is a huge need of computational methods able to integrate and analyse data to fill this gap.

Current researches in this area are following two different integrative methods: one uses the complementary information of different measurements for the

study of complex phenotypes on the same samples (multi-view learning); the other tends to infer knowledge about the phenotype of interest by integrating and comparing the experiments relating to it with respect to those of different phenotypes already known through comparative methods (meta-analysis). Meta-analysis can be thought as an integrative study of previous results, usually performed aggregating the summary statistics from different studies. Due to its nature, meta-analysis usually involves homogeneous data. On the other hand, multi-view learning is a more flexible approach that considers the fusion of different data sources to get more stable and reliable estimates. Based on the type of data and the stage of integration, new methodologies have been developed spanning a landscape of techniques comprising graph theory, machine learning and statistics. Depending on the nature of the data and on the statistical problem to address, the integration of heterogeneous data can be performed at different levels: early, intermediate and late. Early integration consists in concatenating data from different views in a single feature space. Intermediate integration consists in transforming all the data sources in a common feature space before combining them. In the late integration methodologies, each view is analysed separately and the results are then combined.

The purpose of this thesis is twofold: the former objective is the definition of a data integration methodology for patient sub-typing (MVDA) and the latter is the development of a tool for phenotypic characterisation of nanomaterials (INSIdEnano). In this PhD thesis, I present the methodologies and the results of my research.

MVDA is a multi-view methodology that aims to discover new statistically relevant patient sub-classes. Identify patient subtypes of a specific diseases is a challenging task especially in the early diagnosis. This is a crucial point for the treatment, because not all the patients affected by the same disease will have the same prognosis or need the same drug treatment. This problem is usually solved by using transcriptomic data to identify groups of patients that share the same gene patterns. The main idea underlying this research work is that to combine more omics data for the same patients to obtain a better characterisation of their disease profile. The proposed methodology is a late integration approach

based on clustering. It works by evaluating the patient clusters in each single view and then combining the clustering results of all the views by factorising the membership matrices in a late integration manner. The effectiveness and the performance of our method was evaluated on six multi-view cancer datasets related to breast cancer, glioblastoma, prostate and ovarian cancer. The omics data used for the experiment are gene and miRNA expression, RNASeq and miRNASeq, Protein Expression and Copy Number Variation.

In all the cases, patient sub-classes with statistical significance were found, identifying novel sub-groups previously not emphasised in literature. The experiments were also conducted by using prior information, as a new view in the integration process, to obtain higher accuracy in patients' classification. The method outperformed the single view clustering on all the datasets; moreover, it performs better when compared with other multi-view clustering algorithms and, unlike other existing methods, it can quantify the contribution of single views in the results. The method has also shown to be stable when perturbation is applied to the datasets by removing one patient at a time and evaluating the normalized mutual information between all the resulting clusterings. These observations suggest that integration of prior information with genomic features in sub-typing analysis is an effective strategy in identifying disease subgroups.

INSIdE nano (Integrated Network of Systems bIology Effects of nanomaterials) is a novel tool for the systematic contextualisation of the effects of engineered nanomaterials (ENMs) in the biomedical context. In the recent years, omics technologies have been increasingly used to thoroughly characterise the ENMs molecular mode of action. It is possible to contextualise the molecular effects of different types of perturbations by comparing their patterns of alterations. While this approach has been successfully used for drug repositioning, it is still missing to date a comprehensive contextualisation of the ENM mode of action. The idea behind the tool is to use analytical strategies to contextualise or position the ENM with the respect to relevant phenotypes that have been studied in literature, (such as diseases, drug treatments, and other chemical exposures) by comparing their patterns of molecular alteration. This could greatly increase the knowledge on the ENM molecular effects and in turn

contribute to the definition of relevant pathways of toxicity as well as help in predicting the potential involvement of ENM in pathogenetic events or in novel therapeutic strategies. The main hypothesis is that suggestive patterns of similarity between sets of phenotypes could be an indication of a biological association to be further tested in toxicological or therapeutic frames. Based on the expression signature, associated to each phenotype, the strength of similarity between each pair of perturbations has been evaluated and used to build a large network of phenotypes. To ensure the usability of INSIdE nano, a robust and scalable computational infrastructure has been developed, to scan this large phenotypic network and a web-based effective graphic user interface has been built. Particularly, INSIdE nano was scanned to search for clique sub-networks, quadruplet structures of heterogeneous nodes (a disease, a drug, a chemical and a nanomaterial) completely interconnected by strong patterns of similarity (or anti-similarity). The predictions have been evaluated for a set of known associations between diseases and drugs, based on drug indications in clinical practice, and between diseases and chemical, based on literature-based causal exposure evidence, and focused on the possible involvement of nanomaterials in the most robust cliques. The evaluation of INSIdE nano confirmed that it highlights known disease-drug and disease-chemical connections. Moreover, disease similarities agree with the information based on their clinical features, as well as drugs and chemicals, mirroring their resemblance based on the chemical structure. Altogether, the results suggest that INSIdE nano can also be successfully used to contextualise the molecular effects of ENMs and infer their connections to other better studied phenotypes, speeding up their safety assessment as well as opening new perspectives concerning their usefulness in biomedicine.

Contents

List of Original Articles

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

- A. Serra, M. Fratello, V. Fortino, G. Raiconi, R. Tagliaferri, and D. Greco, MVDA: a multi-view genomic data integration methodology, BMC bioinformatics, vol. 16, no. 1, p. 261, 2015.
- Fratello, M., Serra, A., Fortino, V., Raiconi, G., Tagliaferri, R., and Greco, D. A multi-view genomic data simulator. BMC bioinformatics, vol 16, no,.1, 2015.
- A. Serra, D. Greco, and R. Tagliaferri (2015, July). Impact of different metrics on multi-view clustering. In 2015 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
- A. Serra, M. Fratello, D. Greco, and R. Tagliaferri (2016, July). Data integration in genomics and systems biology. In 2016 World Congress on Computational Intelligence (WCCI). IEEE.
- A. Serra, I. Letunic, V. Fortino, B. Fadeel, R. Tagliaferri and D. Greco. Integrated network analysis links metal nanoparticles and neurodegeneration. Manuscript Submitted

Chapter 1

Introduction

Cells are the basic structural, functional, and biological units of life and can be considered as the building block of all living beings [1].

They carry precise instructions in DNA concerning how they grow and function. Complexity arose in the study of the cell phenotype, when it comes to study the dynamic aspects of the DNA at the level of genes, RNA transcripts, proteins, metabolites and their interactions.

In fact, the components of a biological system (for example, genes, proteins, metabolites and so on) function in networks and these networks interact with each other [2]. This gave birth to systems biology science whose main idea is that the molecular level of cells must be studied together organically and comprehensively, rather than separately. Since the objective of systems biology is to model the interactions in a system, the experimental techniques need to be system-wide and be as complete as possible. Therefore, there is a need to collect and to integrate different kinds of data that are used to develop new approaches for the contextualisation of the behaviour of biological networks and to design and validate models [3–5].

Thanks to the technological advances in omics technologies and the decrease of storage costs, different types of omics data have become available, among

them, there are gene expression, microRNA expression (miRNA), protein expression, copy number variation (CNV) etc. Each of these experimental data provides potentially complementary information about the whole studied organism.

Integrating these different sources is an important part of current systems biology research, since it is necessary to understand the whole biological system while the information coming from a single experiment is not sufficient. For example, the goal of functional genomics is to define the function of all the genes in the genome under a given condition. This is a difficult task that requires the integration from different experiments in order to be achieved [6]. Lanckriet et al.[7] for example, aimed at classifying proteins as membrane proteins or not. They demonstrated that the integration of genomic data, amino acid sequences and protein-protein interactions increase the classification performance compared to the use of only protein sequencing information.

Data integration methodologies arise in a wide range of clinical applications as well. Recently, new data integration techniques have been proposed to increase the clinical relevance of patients' sub-classifications. Wasito et al. [8] proposed a kernel based integration method for lymphoma cancer sub-typing. They demonstrated that using the integration of DNA microarray and clinical data using Support Vector Machines (SVM) and Kernel Dimensionality Reduction (KDR) improves the accuracy in the identification of cancer subtypes. Sun et al. [9] employed multi-view bi-clustering to subtype cocaine users. They proposed a matrix decomposition approach that integrates already known genetic markers with clinical features to identify significant subtypes of the disease. Data integration also plays an important role in toxicogenomics, when researchers want to understand the interaction between the genome and the environment to investigate the response of the genes to toxins and how they modify the gene expression function. Patel et al.[10] describes the contribution of different data sources in advancing this field. The goals of data integration are to obtain higher precision, better accuracy, and greater statistical power than those provided by single datasets. Moreover, integration can be useful in validating results from different datasets, under the assumption that if information from independent

data sources agrees, it is more likely that the information is more reliable than information from a single source. Unfortunately, there is a big gap between the amount of information and the knowledge in which it is translated. Even though many computational strategies have been developed to pre-process and analyse gene expression data, there is still a huge need of computational methods able to integrate and analyse gene expression data with other omics data.

1.1 Gene expression

The gene expression is the process by which the information contained in a gene is used to obtain a gene product that is often represented by proteins [11]. In a non-protein coding genes, such as the tRNA and small nuclear RNA (snRNA) genes, the final product is a functional RNA. It is worth mentioning that it is a common assumption to take the gene expression level as proportional to the amount of proteins translated. Indeed, several studies show that there exists a strong correlation between expression levels and protein abundance [12–14]. However, it is well known that there exist situations where the opposite is also true [15], [16]. In these cases, it is said that a post-transcriptional activity occurred.

This makes the role of gene expression of paramount importance in the study of the molecular profile of the cells. In fact, considering the existent relationship between gene expression and protein translation, the level of mRNA can provide indirectly knowledge on the state of the cell [17]. For example, by comparing the level of gene expression in healthy and diseased subjects, the molecular basis of the disease can be determined. Furthermore, by measuring the level of gene expression as a function of serial processes, the molecular changes over time (cell cycle), or the response to a specific drug or metabolite, can be identified.

Three are the options available for investigating the molecular dynamics of the cell, that give rise to three different approaches: (1) proteomics, where the set of proteins in the cell are analysed, (2) transcriptomics, where the set of mRNA transcripts that lead to the production of proteins are analysed, (3) metabolomics, where the set of metabolites generated by the proteins are ana-

lysed. Research in proteomics and metabolomics has been ongoing for many years, but there is still a lack of standardised methodologies and poor reproducibility of the experiments. On the other hand, the DNA microarray technology is a well-established tool for transcriptomic studies [18], [19]. DNA microarrays are not the only tool available to study the gene expression, new technologies such as RNA Sequencing, is becoming widely used for transcriptomic data analysis.

1.2 1.2 High-throughput omics technology

Technological advancement allows simultaneous examination of thousands of genes with high-throughput techniques. Two of the major technologies used for transcriptomic analyses are DNA microarray and RNA sequencing. They both allow to measure gene expression but their protocol is substantially different. Zhao et al.[20] demonstrate that RNA-Seq has some advantages with respect to DNA microarray. RNA-Seq technology has a higher capability in detecting low abundance transcripts. It also has a broader dynamic range than microarray, that allows to detect more differentially expressed genes with higher fold-changes. However, despite the benefits of RNA-Seq, microarrays are still widely used by researchers to conduct transcriptional profiling experiments. This is probably because microarrays are better known, data is easier to analyse and they are less expensive that RNA-Seq.

1.2.1 1.2.1 DNA microarrays

DNA microarrays have been proposed by Shena et al. in 1995 as a technology able to simultaneously monitor the level of mRNA transcript for tens of thousands genes [21]. Even if the microarrays are used for a variety of different purposes, such as comparative genomics hybridisation (CGH) [22] or CHIP-on-CHIP [23], their most popular application is still the large scale gene expression analysis. Microarrays have been also used to study several diseases, the cell cycle of various organisms as well as the regulation of many biological mechanisms.

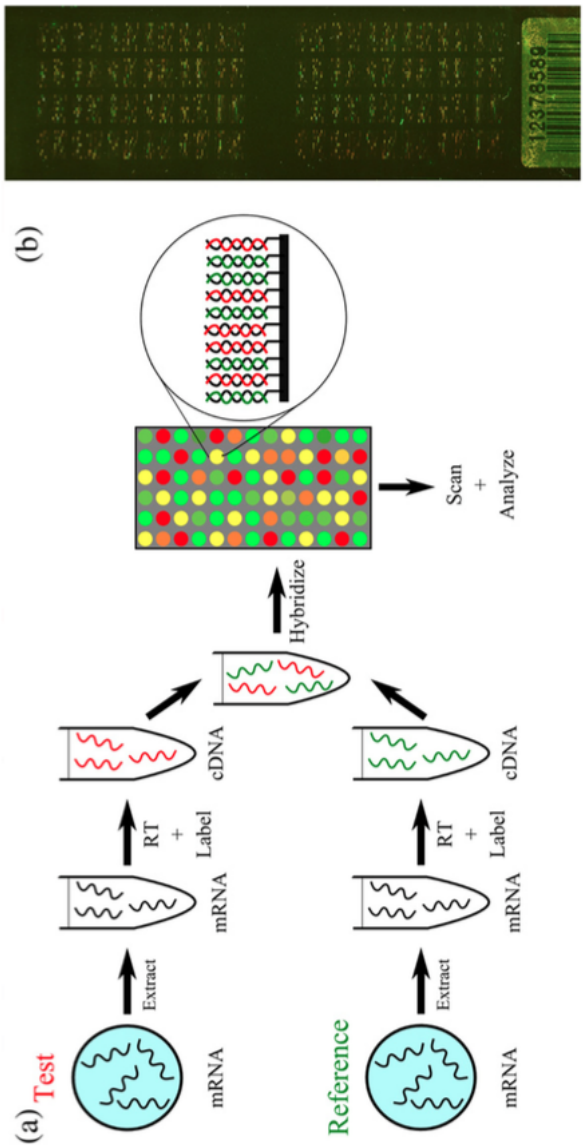


Figure 1.2.1: Microarray Experiment: (a) Spotted microarray experimental set-up. mRNA extracts (targets) from cells under two distinct physiological conditions are reverse transcribed to cDNA and then labelled with different fluorescent dyes e. g. Cy3 and Cy5. Equal amounts of the dye-labelled targets are combined and applied to a glass substrate onto which cDNA amplicons or oligomers (probes) are immobilised. (b) Scanned image of an Atlantic salmon cDNA microarray. Figue from Tobias et al.

Microarrays are small slides to which DNA molecules are bound. From the samples of each experimental condition the mRNA present in the cells is extracted and manipulated (reverse transcription) to obtain complementary DNA (cDNA). At the same time, it is also labelled with fluorescent compounds (Cy3 green fluorescence, Cy5 red fluorescence) to mark each class of sample. The cDNA sequences hybridised with complementary sequences are positioned on the microarray spots. The slide is then scanned with a laser light at different wavelengths. For each spot the fluorescence intensity is recorded. The spots that respond to green light correspond to genes expressed only in first experimental condition, whereas spots responding to red light correspond to genes expressed only in the second experimental condition; finally, the spots responding to both lights (i.e. yielding a yellow fluorescence) indicate genes expressed in both experimental conditions. The intensity of the expression level is given by the intensity average of the corresponding spots in the image. This process is illustrated schematically in Figure 1.2.1.The expression values derived from measurements made on the microarray are noisy and need to be pre-processed, corrected with respect to the background [24] and normalised [25], [26].

1.2.2 1.2.2 RNA Sequencing

Another technology used to quantify gene expression is based on next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment in time [27], [28]. Classical DNA sequencing techniques aims to identify the order of the four bases in a DNA strand. NGS, also called high-throughput sequencing technologies, allows to parallelize the sequencing process, producing thousands or millions of sequences concurrently [29], [30]. The main differences between Microarray technology and RNA-Seq is that with the Microarray only a limited number of genes, those bound on the spots, can be studied. On the other hand, with RNA-Seq methodology a scanning of the whole genome is performed, allowing to investigate both known transcripts and exploring new ones. Therefore, RNA-seq is ideal for discovery-based experiments [28]. For example, 454-based RNA-Seq has been used to sequence the transcriptome of the Glanville fritillary butterfly [31]. Moreover,

RNA-Seq has very low background signal, in fact, DNA sequences can been mapped to unique regions of the genome [28].

The standard work-flow of an RNA-seq analysis is composed of three steps (see Figure 1.2.2): (1) the RNAs in the sample of interest are fragmented and reverse-transcribed to create a library of complementary DNAs fragments (cDNAs). (2) The obtained cDNAs are then amplified (i.e. duplicated millions of times [32]) and subjected to NGS to obtain short sequences. (3) The short reads generated can then be mapped on a reference genome. The number of reads aligned to each gene are called counts and gives a digital measure of gene expression levels in the investigated sample [33].

There are multiple methods for computing counts. The most used method considers the total number of reads overlapping the exons (i.e. the portion of a gene that is transcribed by RNA polymerase during the transcription process) of a gene [34].

However, it can happen that a fraction of reads maps outside the boundaries of known exons [35]. Thus, alternatively, the whole length of a gene can be considered, also counting reads from introns (i.e. non-coding region of the genes that are transcribed by RNA polymerase). The 'Union-Intersection gene' model considers the union of the exonic bases that do not overlap with the exons of other genes [36].

After computing the counting, two kinds of biases have to be removed [37–39]. The former to be taken into account is the sequencing depth of a sample, defined as the total number of sequenced or mapped reads. For two given RNA-seq experiments A and B, if A generates twice the number of reads than the experiment B, it is likely that the counts from experiment A will be doubled. To solve this problem, a common practise is to scale counts in each experiment by the sequencing depth estimated for the sample [40].

The latter is related to gene length [36, 41]; in fact, the expected number of reads mapped into a gene is proportional to both the abundance and length of the isoforms transcribed from that gene. Indeed, longer genes produce more reads than shorter ones. To reduce this bias, Mortazavi et al. [42] proposed to summarise the reads with the "Reads Per Kilobase of exon model per Million

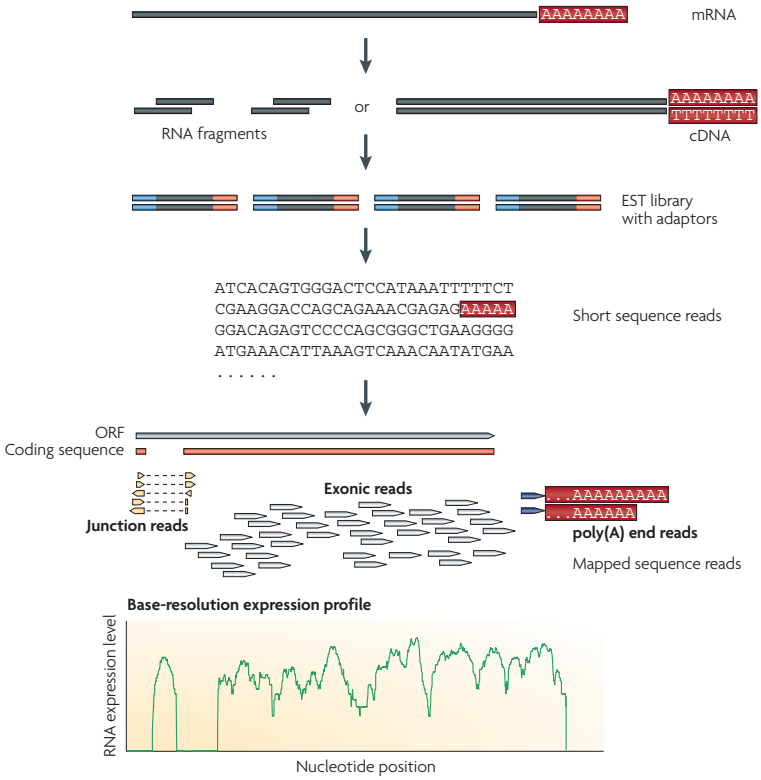


Figure 1.2.2: A typical RNA-seq experiment. Briefly, long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation (see main text). Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. The resulting sequence reads are aligned with the reference genome or transcriptome, and classified as three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a base-resolution expression profile for each gene, as illustrated at the bottom; a yeast ORF with one intron is shown. Figure and Legend from [28].

mapped reads" (RPKM) measure, that is computed by dividing the number of reads aligned to a gene exon, by the total number of reads mapped and by the sum of exonic bases.

microRNA expression

MicroRNAs (miRNAs) are small (approximately 18-24 nucleotides) non-coding RNA molecules of single-stranded DNA, which bind to mRNAs and regulate protein expression, either promoting degradation of the mRNA target and/or by blocking translation [43, 44], or alternatively by increasing translation [45]. miRNA are mainly active in the regulation of gene expression at the transcriptional and post-transcriptional levels. They have been found to be involved in numerous cell functions such as proliferation, differentiation, death [46, 47]. The aberrant expression of miRNAs has been implicated in the onset of many diseases [48, 49] and they can be used for therapeutic purposes [50]. Profiles of miRNAs in various types of tumors have been shown to contain potential diagnostic and prognostic information [51]. The study of mirna expression is slightly more complicated than the study of gene expression. Several technical variables must be taken into account in problems related to microRNA isolation, the stability of stored miRNA samples and microRNAs degradation [52]. As for gene expression, microRNA expression can be quantified by hybridization on microarrays, slides or chips with probes to hundreds or thousands of miRNA targets, so that relative levels of miRNAs can be determined in different samples [53]. microRNAs can be both discovered and profiled by high-throughput sequencing methods (microRNA sequencing) [54].

1.3 Biological Databases

Given the large number of omics data produced, there have been many efforts made to collect the results of the experiments and make them available to the research community. In fact, many are the biological available online databases, which contain the experimental data (either in raw form that pre-processed) and the knowledge produced by such experiment (i.e. the connection between

the gene and disease). Examples of database that provide experimental data are The Cancer Genome Atlas (TCGA - [55]), the Gene Expression Omnibus (GEO [56]) databases, the Connectivity Map (CMAP - [57]) and the NanoMiner ([58]) databases.

The TCGA is a public repository that collects samples related to more that 30 different tumours. It makes available the clinical information of the samples, metadata regarding technicalities of the experiments, histopathology images and molecular different information such as mRNA/miRNA expression, protein expression, copy number, etc. GEO is a public repository that collect array and sequence-based data coming from the scientific community and makes them available to the public. It provides tool to help the user to query and download experiments and curated gene expression profiles. CMAP is a collection of transcriptional expression data from human cells treated with drugs. The main goal of the project was to discover functional connections between drugs, genes and diseases through the transitory feature of common gene-expression changes. NanoMiner database contains in-vitro transcriptomic data on human samples exposed to nanoparticles.

On the other side, some examples of database collecting knowledge extracted from the experiments are the Comparative Toxicogenomics Database (CTD [59]), the Medicaton Indication dataset (MEDI [60]) and the Molecular Signature Database (MSigDB [61]).

CTD is a robust, publicly available database that aims to advance understanding about how environmental exposures affect human health. It provides manually curated information about chemicalâĂŞgene/protein interactions, chemicalâĂŞdisease and geneâĂŞdisease relationships. These data are integrated with functional and pathway data to aid in development of hypotheses about the mechanisms underlying environmentally influenced diseases. MEDI is an ensemble resource of electronic medical record (EMR) data. It contains information related to the drugs that have been prescribed to treat certain diseases. The MSigDB is a collection of annotated gene sets, that are useful for the GSEA analysis or to evaluate the biological meaning of clusters of genes. All these data can be used to perform integrative analysis both with multi-view and

meta-analysis techniques.

1.4 Data Integration

Data integration (or data fusion) methodologies integrate multiple datasets in order to increase the accuracy, to reduce the noise and to extract more accurate information from multi modal datasets by finding correlation across them. Figure 1.4.1 reports a classification of the integration methodologies based on the statistical problem, the type of analysis to be performed, on the type of data to be integrated and on the integration stage as described in [62].

Type of Analysis

The analysis to be performed is somehow limited by the type of data involved in the experiment and by the desired level of integration. Analyses can be divided into two categories: meta-analysis and integrative analysis. Meta-analysis can be thought as an integrative study of previous results, usually performed aggregating the summary statistics from different studies [63, 64]. Due to its nature, meta-analysis can only be performed as a step of late integration involving homogeneous data. On the other hand, integrative analysis is a more flexible approach that considers the fusion of different data sources to get more stable and reliable estimates. Based on the type of data and on the stage of integration, new methodologies have been developed spanning a landscape of techniques comprising graph theory, machine learning and statistics.

Type of Data

Data integration methodologies in systems biology can be divided into two categories based on the type of data: integration of homogeneous or heterogeneous data types. Usually biological data are thought to be homogeneous if they assay the same molecular level, for gene or protein expression, copy number variation, and so on. On the other hand if data is derived from two or more different molecular levels they are considered to be heterogeneous. Integration of this

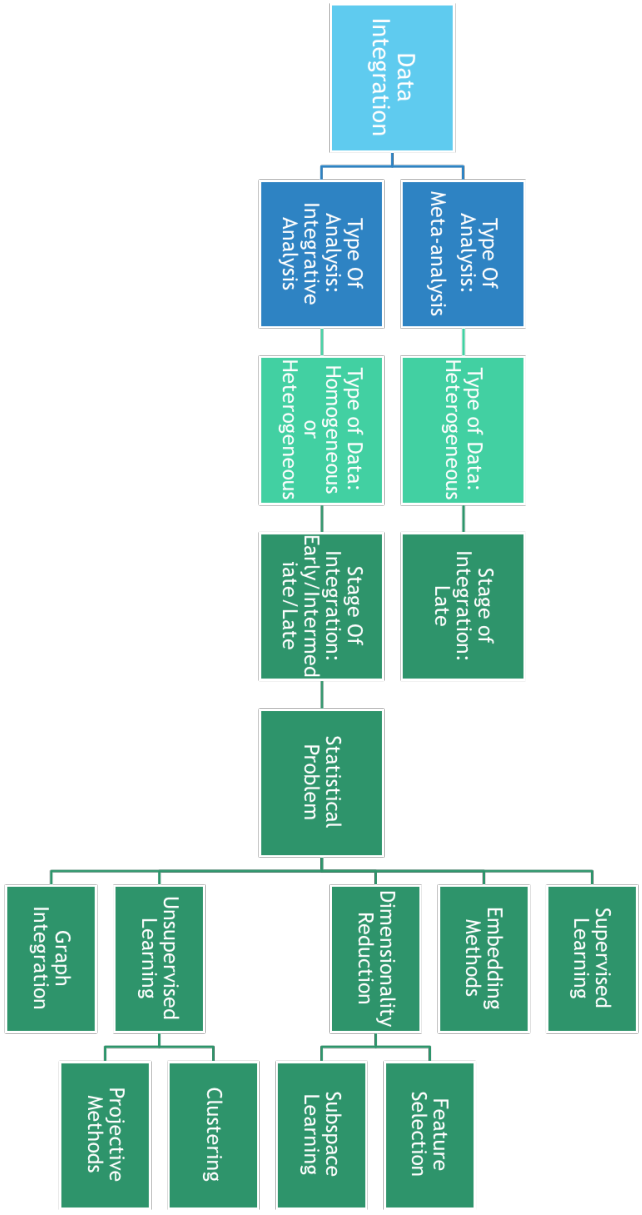


Figure 1.4.1: Data Integration Taxonomy

kind of data poses some issues: first, the data can have different structure, for example they can be sequences, graphs, continuous or discrete numerical values. Moreover, data coming from different sources is subject to different noise levels depending on the platform and on the technologies used to generate data. The integration process must include a step, called batch effect removal where the noise and the random or systematic errors between the different views become comparable [65].

Integration Stage

Depending on the nature of data and on the statistical problem to address, the integration of heterogeneous data can be performed at different levels: early, intermediate and late. Early integration consists in concatenating data from different views in a single feature space, without changing the general format and nature of data. Early integration is usually performed in order to create a bigger pool of features by multiple experiments. The main disadvantage of early integration methodologies is given by the need to search for a suitable distance function. In fact, by concatenating views, the data dimensionality considerably increases, consequently decreasing the performance of the classical similarity measures [66]. Intermediate integration consists in transforming all the data sources in a common feature space before combining them. In classification problems, every view can be transformed in a similarity matrix that will be combined in order to obtain better results. In the late integration methodologies each view is analysed separately and the results are then combined. Late integration methodologies have some advantages over early integration techniques: (1) the user can choose the best algorithm to apply to each view based on the data; (2) the analysis on each view can be executed in parallel.

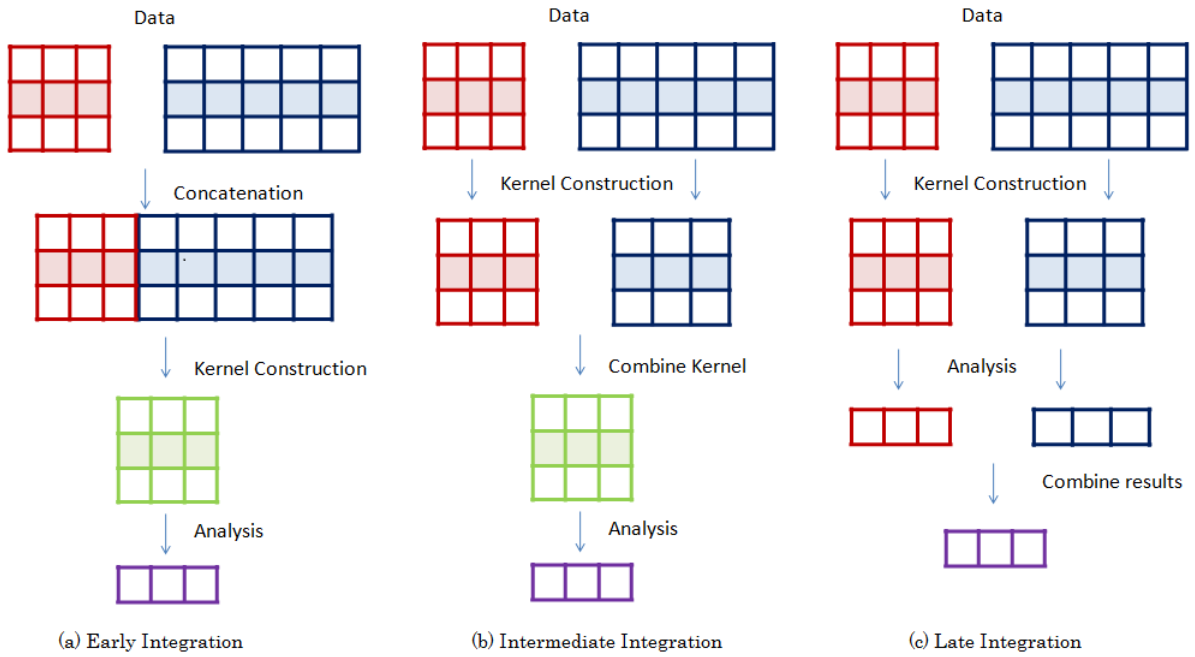


Figure 1.4.2: Data integration stage proposed by Pavlidis et al [67].They proposed an SVM kernel function in order to integrate microarray data. In early integration methodologies SVMs are trained with a kernel obtained from the concatenation of all the views in the dataset (a). In intermediate integration, first a kernel is obtained for each view, and then the combined kernel is used to train the SVM (b). In the late integration methodology a single SVM is trained on a single kernel for each view and then the final results are combined (c).

Chapter 2

Aim of the study

The aim of this work is to provide new data integration tools to help the scientific community. Particularly, I have focused on two main objectives:

- The definition of a methodology that integrates multiple omics feature sets to find statistical relevant patient subtypes.
- The development of a tool for phenotypic characterisation of nanomaterials mode-of-action with respect to human diseases, drugs treatments and chemicals exposures.

Chapter 3

Multi View Learning for Patient Subtyping

In this chapter the multi-view genomic data integration methodology (MVDA) is described. MVDA is a clustering based methodology proposed to identify patient sub-types by combining multiple high-throughput molecular profiling data. It is a late integration approach where the views are integrated at the levels of the results of each single view clustering iterations. By using MVDA, patient sub-classes with statistical significance were retrieved, identifying novel sub-groups previously not emphasised in literature. The content of this chapter is published in [68].

3.1 Introduction

Many diseases - for example, cancer, neuropsychiatric, and autoimmune disorders - are difficult to treat because of the remarkable degree of variation among affected individuals [69]. Trying to solve this problem a new discipline emerged, called precision medicine or personalized medicine [70]. It tries to individualize

the practice of medicine by considering individual variability in genes, lifestyle and environment with the goal of predicting disease progression and transitions between disease stages, and targeting the most appropriate medical treatments [71].

A central role in precision medicine is played by patients subtyping, that is the task of identifying subpopulations of similar patients that can lead to more accurate diagnostic and treatment strategies Identify disease subtypes can help not only the science of medicine, but also the practice. In fact, from a clinical point of view, refining the prognosis for similar individuals can reduce the uncertainty in the expected outcome of a treatment on the individual. Traditionally, disease subtyping research has been conducted as a by-product of clinical experience, wherein a clinician noticed the presence of patterns or groups of outlier patients and performed a more thorough (retrospective or prospective) study to confirm their existence.

In the last decade, the advent of high-throughput biotechnologies has provided the means for measuring differences between individuals at the cellular and molecular levels. One of the main goals driving the analyses of high-throughput molecular data is the unbiased biomedical discovery of disease subtypes via unsupervised techniques. Using statistical and machine learning approaches such as non-negative matrix factorization, hierarchical clustering, and probabilistic latent factor analysis [72], [73], researchers have identified subgroups of individuals based on similar gene expression levels. For example, the analysis of multivariate gene expression signatures was successfully applied to discriminate between disease subtypes, such as recurrent and non-recurrent cancer types or tumour progression stages [74]. To improve the model accuracy for patient stratification, in addition to gene expression, other omics data type can be used, such as miRNA (microRNA) expression, methylation or copy number alterations. For example, somatic copy number alterations provide good biomarkers for cancer subtype classification [75]. Data integration approaches to efficiently identify subtypes among existing samples has recently gained attention. The main idea is to identify groups of samples that share relevant molecular characteristics. Strategies of data integration of multiple omics data types poses several compu-

tational challenges, as they deal with data having generally a small number of samples and different pre-processing strategies for each data source. Moreover, they must cope with redundant data as well as the retrieval of the most relevant information contained in the different data sources. When the integrated data are high dimensional and heterogeneous, defining a coherent metric for clustering becomes increasingly challenging. A number of data integration approaches for patients subgroups discovery were recently proposed, based on supervised classification, unsupervised clustering or bi-clustering [76–79]. Moreover, multi-view clustering methodologies have been intensively used also if in few cases on omics data

3.2 Materials and Methods

3.2.1 Clustering

Clustering is an unsupervised learning technique, able to extract structures from data without any previous knowledge on their distribution. It is one of the main exploratory techniques in data mining and it is used to group a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other clusters.

Clustering have been widely applied in bioinformatics. Two of the main problems addressed by clustering are: (1) identify groups of genes that share the same pattern across different samples [80]; (2) identify groups of samples with similar expression profiles [81]. The number of different clustering techniques proposed in literature is huge. Two of the most common approaches are hierarchical clustering or partitive clustering [82].

The former methods start having each object in different clusters and then they, iteratively, join couple of similar clusters until they reach a stop criterion. This kind of methodology creates a hierarchy of the clusters that is called dendrogram. The crucial point in hierarchical clustering is the evaluation of the measure between two clusters. Different measures have been proposed such as: the single-linkage, where the distance between two clusters is defined as the

minimum distance between each couple of points between the two clusters; the complete linkage computes the distance between two clusters as the maximum distance between each couple of points. In the average linkage the distance between two clusters is computed as the mean of the distances between all the couples of objects between the two clusters. Ward's minimum variance linkage aims to minimise the total within-cluster variance. All these methods produce a hierarchy of the samples, in order to clustyer data, this hierarchy must be cut at a certain height that is arbitrary chosen by the user. To solve this problem the Pvclust [83] algorithm have been proposed. It is a hierarchical clustering algorithm that applies a multi-scale bootstrap re-sampling procedure to the dataset, to compute a p-value that is used to cut the tree to obtain clusters that are statistically supported by data.

The latter methods, conversely, start from a group of initial points, called centroids, that represent the clusters and in an iterative manner, assign each sample in the dataset to a centroid and if necessary update the centroids. The whole process aims to minimise an objective function, and the algorithm runs until convergence or until a stop criterion is reached. Examples of this kind of algorithms are Kmeans [84], Partitional Around Medoids (PAM) [85] and SOM [86].

Given a set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional real vector, k-means clustering [84] aims to partition the n observations into $k(\leq n)$ sets $S = S_1, S_2, \ldots, S_k$ so as to minimise the within-cluster sum of squares (WCSS) In other words, its objective is to find:

$$\arg\min_{x} \sum_{i=1}^{K} \sum_{x \in S_i} ||x - \mu_i||^2 \tag{3.2.1}$$

where μ_i is the mean of the points in S_i. The optimal solution is obtained iterating two different steps: the former in which each point is assigned to the nearest centroid, and the latter in which the centroids are updated to minimise the equation 3.2.1.

Partitional Around Medoids [85] is a clustering algorithm related to the K-means algorithm and the medoids shift algorithm. Both the K-means and

K-medoids algorithms are partitional (breaking the dataset up into groups) and both attempt to minimise the distance between points assigned to a cluster and a point designated as the centre of that cluster. Contrary to the K-means algorithm, K-medoids chooses data points as centres (medoids or exemplars) and works with an arbitrary matrix of distances between data points;

SOM [86], is a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low dimensional (typically two-dimensional), discretised representation of the input space, called a map. Self-organising maps are different from other artificial neural networks in the sense that, during the training phase, they use a neighbourhood function to preserve the topological properties of the input space. A self-organizing map consists of components called nodes or neurons. Associated with each node there are a weight vector of the same dimension of the input data vectors (prototype), and a position in the map space. The usual arrangement of nodes is a two-dimensional regular spacing in a hexagonal or rectangular grid. The self-organising map describes a mapping from a higher-dimensional input space to a lower-dimensional map space. The procedure for placing a vector from input space onto the map is to find the node with the closest (smallest distance metric) weight vector to the data space vector. This makes SOMs also useful for obtaining low-dimensional views of high-dimensional data, akin to multidimensional scaling.

The SOM is trained iteratively. At each training step, a sample vector x is randomly chosen from the input data set. Distances between x and all the prototype vectors are computed. The best-matching unit (BMU), which is denoted here by b, is the map unit with prototype closest to x

$$||x - m_b|| = min_i||x - m_i||. \tag{3.2.2}$$

Next, the prototype vectors are updated. The BMU and its topological neighbours are moved closer to the input vector in the input space. These two iterations are repeated until convergence. At the end, all the points assigned to the same neuron in the SOM will be allocated in the same cluster.

A clustering algorithm different from these two families is the spectral clustering [87]. The general approach to spectral clustering starts from a similarity

matrix, then computes the relevant eigenvector of its Laplacian. In the space of these eigenvectors, a classical clustering algorithm, such as K-means is executed. Traditional state-of-the-art spectral methods [87] aim to minimise RatioCut [88], by solving the following optimisation problem:

$$\min_{Q \in R^{n \times c}} Trace(Q^T L^+ Q) \quad s.t. Q^T Q = I \tag{3.2.3}$$

where $Q = Y(Y^T Y)^{-1/2}$ is a scaled partition matrix, L^+ denotes the normalised Laplacian matrix $L^+ = I - D^{-1/2} W D^{-1/2}$ given the similarity matrix W.

When performing a clustering analysis, the first problem is to choose the algorithm that best fits the problem. There are many different clustering algorithms and the application of each one will usually produce different results. Moreover the results of the algorithms strongly depend on the input parameters (such as the number of clusters). Without additional evaluation, it is difficult to determine which solutions are better. To solve this problem some indexes have been proposed to asses the clustering solutions [89]. Usually algorithms are executed with different parameters and then solutions that reach the best values of the evaluation indexes are selected.

3.2.2 Multi-View Clustering

The main difference between traditional and multi-view clustering is that the former takes multiple views as a flat set of variables without taking into account differences among different views, while the latter exploits the information from multiple views and takes the differences among the views into consideration in order to produce a more accurate and robust data partitioning. Multi-view clustering has been widely applied in machine learning by using different variants of existing single view clustering methodologies [90–95]. Moreover, it has been widely applied to integrate different genome-wide measurements in order to identify cancer-subtypes [76–78, 93, 96].

Chen et al. [91] proposed the early integration method *Two-level variable Weighting k-means* (TW-kmeans) clustering for multi-view data. This method

extends the classical k-means algorithm by incorporating the weights of the views and of the variables into the distance function that identifies clusters of objects. The algorithm is able to identify the set of k clusters, the important views and the relevant variables for each view. The authors evaluated the performance of the TW-k-means algorithm for classification of real life data, by testing it on three data sets from UCI Machine Learning Repository.

The algorithm proposed by Long et al. [94] exploits the intuition that the optimal clustering is the consensus clustering shared across as many views as possible. This can be reformulated as an optimisation problem where the optimal clustering is the closest to all the single view clusterings under a certain distance or dissimilarity measure. Clusterings are again represented as membership matrices. Formally the model can be described as follow: given a set of clustering membership matrices $M = [M_1, \ldots, M_h] \in R_+^{n \times l}$, a positive integer k and a set of non-negative weights $\{w_i R_+\}_{i=1}^m$, the optimal clustering membership matrix $B \in R_+^{n \times k}$ and the optimal mapping matrices $P = [P_1, \ldots, P_h] \in R_+^{k \times l}$ are given by the minimisation:

$$\begin{aligned} \min_{B,P} \quad & \sum_{i=1}^{h} w_i GI(M^{(i)} || BP^{(i)}) \\ \text{s. t.} \quad & P \geq 0 \\ & B \geq 0, B\mathbf{1} = \mathbf{1} \end{aligned} \tag{3.2.4}$$

where $GI(M||BP)$ is the generalised Kullback-Leibler divergence such that

$$GI(X||Y) = \sum_{ij} (logX_{ij} log \frac{X_{ij}}{Y_{ij}} - X_{ij} + Y_{ij})$$

subject to the constraint that both P and B must be non-negative and that each row of B must sum to one. The method has been evaluated on both synthetic and two real data sets: the former is a web page advertisement data-set and the latter is a newspaper dataset. They executed the multi-view clustering algorithm ten times on each dataset by imposing the number of clusters equal to the number of real classes and evaluated the final multi-view clustering with

respect to the real class labels with the Normalised Mutual Information. They demonstrated that the integration methodology gives better results compared to those obtained with the use of single view separately and the experiments showed that the algorithm efficiently learns robust consensus patterns from multiple view data with different levels of noise.

Green et al. [95] proposed a meta-analysis technique for multi-view clustering in a late integration manner. The main idea is to use the matrix factorisation approach to combine clustering results on each single view expressed in form of membership matrices.The method first transposes all the membership matrices and stacks them vertically in order to obtain the cluster matrix X in R l×n where l is the total number of clusters in C and n is the number of samples. Then, it iteratively finds the best approximation of X such that $X \approx PH$ and $P \geq 0$, $H \geq 0$ by measuring the error with the Frobenius norm and the update rules proposed by Lee et al. [97]. The results of the factorisation are two matrices, $P \in R^{\{l \times k\}}$ that projects the clusters into a new set of k meta-clusters and $H \in R^{\{k \times n\}}$ whose columns can be viewed as the membership of the original objects in the new set of meta-clusters. Starting from the values of P, a matrix $T \in R^{\{v \times k\}}$ is computed, with v being the number of views. T_{hf} indicates the contribution of the view V_h to the fth meta-cluster. The method has been evaluated on both synthetic and real-world document datasets. In both settings it produced more informative clusterings with respect to the single view clustering counterparts. It turned out that the method can effectively take advantage of cases when a variety of different clusterings are available for each view and in fact out-performed popular ensemble clustering algorithms.

Yang et al. [98] proposed a biclustering algorithm, based on matrix factorisation, for module detection in multi-view genomic data sets. Their method, called iNMF is a modification of the jSNF algorithm [99]. Both methods are able to factorise multi-view datasets but, while jSNF considers all the views as having the same effect on the resulting factorisation, iNMF allows each view to bring its own contribute to the factorisation process. The method was tested on a real ovarian cancer dataset coming from TCGA. The dataset was composed of three views: DNA methylation, gene expression and miRNA expression. The

method was able to detect multi-modal modules and sample clusters that agree with the ones already in the literature.

iCluster [96] uses a joint latent-variable model to identify the grouping structure in multi *omics* data. This method simultaneously achieves data integration and dimension reduction, by reporting all the views to a common space, that has a number of latent variables significantly smaller than the originals ones, and clustering patients in that space. The optimal space is identified in an optimization process that uses the EM algorithm and a Lasso model that penalises the norm of the coefficient vectors and continuously shrinks the coefficients associated with non-informative genes toward zero, to ensure the data sparseness. The method was tested by simultaneously clustering gene expression, genome-wide DNA copy number and methylation data derived from the TCGA Glioblastoma Multiforme samples. The authors compared their method with a naive integration obtained by concatenating the views and applying PCA. Their results showed that iCluster had better capability to stratify patients by integrating different *omics* views.

On the other hand, SNF [93] is an intermediate integration network fusion methodology able to integrate multiple genomic data (e.g., mRNA expression, DNA methylation and microRNA expression data) to identify relevant patients' subtypes. The method first constructs a patients' similarity network for each view. Then, it iteratively updates the network with the information coming from other networks in order to make them more similar at each step. At the end, this iterative process converged to a final fused network. The authors tested the method combining mRNA expression, microRNA expression and DNA methylation from five cancer data sets. They showed that the similarity networks of each view have different characteristics related to patients similarity while the fused network gives a clearer picture of the patients' clusters. They compared the proposed methodology with iClust and the clustering on concatenated views. Results were evaluated with the silhouette score for clustering coherence, Cox log-rank test p-value for survival analysis for each subtype and the running time of the algorithms. SNF outperformed single view data analysis and they were able to identify cancer subtypes validated by survival analysis.

MEREDITH [77] is an intermediate late integration approach methodology to discover cancer sub-typing by integrating multiple *omics* feature sets. This methodology is composed of several steps comprising data normalisation, data integration, clustering and validation. Since, it is a gene centred method, as a first step, all the features in the different views are mapped to the corresponding genes. Then, a PCA analysis is performed to reduce the dimensionality of each view. The first 50 PCs per dataset with the highest eigenvalues are retained and then scaled and concatenated to construct a single integrated dataset. A mapping into a two-dimensional space is performed by means of the t-distributed stochastic neighbourhood embedding. The clustering analysis and survival analysis is performed. The method was tested on more that 4000 patients coming from the The Cancer Genome Atlas (TCGA)across 19 cancer-types and four views: gene expression, DNA-methylation, copy-number variation and microRNA expression. The performed DBSCAN clustering was able to identify 18 clusters significantly over-represented by a cancer type ($p < 0.001$). MEREDITH was then able to identify known cancer types and subtypes. This results suggest that data integration methodologies could enable novel insight in patients characterization.

The proposed methodology for the analysis of multi-view biological datasets takes in input n matrices $M_i \in R^{F_i \times P}$ for $i = 1, \ldots, n$ where F_i is the number of features (genes, miRNAs, CNV, methylation, clinical information, etc.) and P is the number of patients and a vector cl of classes labels, and yields a multi-view partitioning $G = \bigcup_{i=1}^{k} G_i$ of patients. The multi-view integration methods also return a matrix C where $c_{i,j}$ is the contribution of view i to the final multi-view cluster j.

The approach consists of four main steps as shown in Figures 3.2.1 and 3.2.2: (a) Prototype Extraction; (b) Prototype ranking; (c) Single view clustering; (d) late integration.

In the prototype extraction step, the features with low variance across the samples were eliminated. Therefore, the data were clustered with respect to the patients and the cluster centroids were selected as the prototype patterns. The centroid of each cluster was selected as the most correlated element with

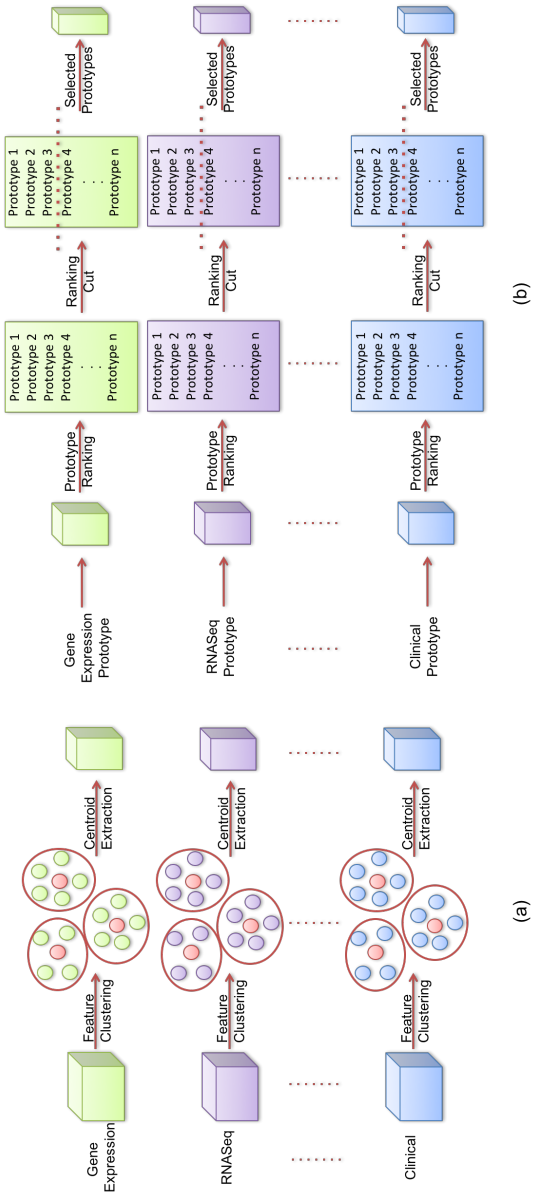


Figure 3.2.1: First two steps of the MVDA methodology. A dimensionality reduction is performed by clustering the features. A prototype is extracted for each cluster to represent it in the following steps (a) The prototypes are ranked by the patient class separability and the most significant ones are selected (b).

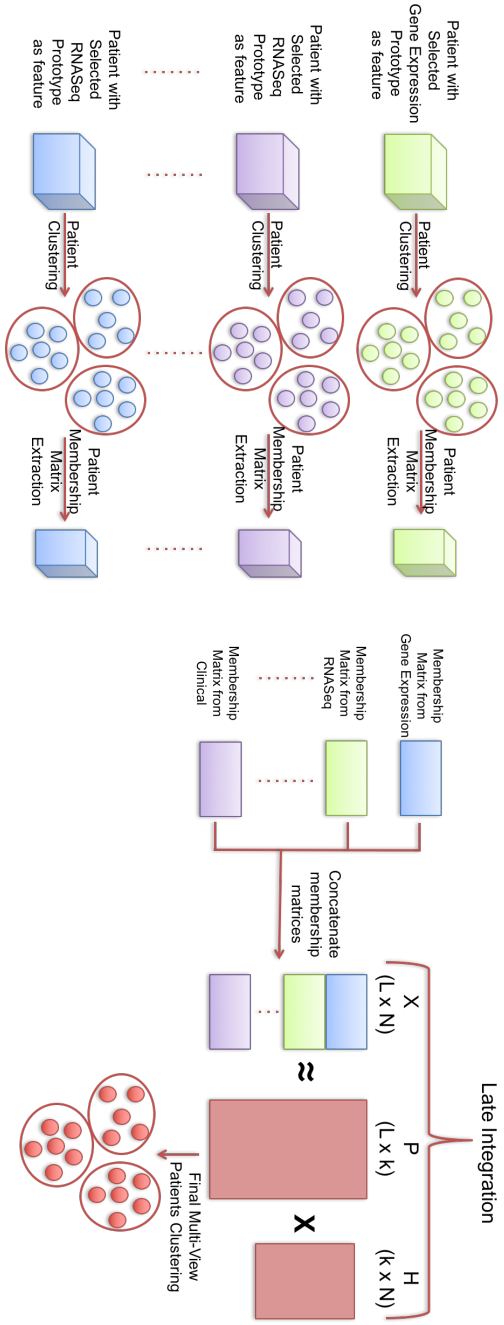


Figure 3.2.2: Last two steps of the MVDA methodology. Single view clustering methods are applied in each view to group patients and obtain membership matrices (c). A late integration approach is utilised to integrate clustering results (d).

respect to the other elements in that cluster. Different clustering algorithms were used such as Hierarchical clustering with Ward's method [100], Pvclust [83], Kmeans [84], PAM [85] and SOM [86]. See section 3.2.1 for more detail on these clustering algorithms.

The idea is to evaluate several popular clustering techniques and compare their behaviour on the different views with respect to the hierarchical method that is the standard algorithm used to cluster genes. Cluster analysis is a complex and interactive process and results change based on its parameters [101]. Therefore, each algorithm was executed for different values of K. For each algorithm and for each K, clustering performance was evaluated according to the following evaluation function:

$$VAL = \frac{1}{4}\left(\frac{IC+1}{2} + 1 - \frac{IE+1}{2} + (1-S) + CG\right)$$

where IC is the complete diameter measure, representing the average sample correlation of the less similar objects in the same cluster; EC is the complete linkage measure, representing the average sample correlation of the less similar objects for each pair of clusters; S is the singleton factor and CG is the compression gain. The evaluation function was defined in order to obtain the output value normalised between 0 and 1. The complete diameter and the complete linkage measures were calculated with the R "clv" package [102]. The number of singleton was normalised in a range (0,1) in order to be comparable with the correlation measure. It was defined as $S = N/(K-1)$. The compression gain was defined as $CG = 1-(K/Nelem)$, where K is the number of clusters and N is the number of elements to be clustered. Each clustering algorithm was executed on n different values of K and the corresponding results were evaluated with the function VAL. Values close to 1 indicate a clustering with similar objects in the clusters, weakly linked clusters, with few singletons and with a good compression rate. A numeric score was then assigned to each K value by considering the average values of the VAL function compiled over the clustering results obtained with the different algorithms. Then, the K showing the highest score was chosen and subsequently used to identify the best clustering algorithms having the first

two highest scores with respect to the selected k value.

The VAL index is a validity index measure such as the Dunn Index or David Bouldin Index [103]. It measures the compactness and separation of the clusters. A "good" clustering should have two characteristics: compactness and separation. A clustering is compact when the members of each cluster are as close to each other as possible. The compactness is measured by the IC value. The separation is therefore measured by the EC value, the less the correlations between two different clusters the more they are separated.

More detail on the computational procedure followed to fine-tuned the k-values for the cluster analysis can be found in the article [68].

In the prototype ranking, further dimensional reduction by feature selection was performed. Feature ranking was computed by means of the CAT-score [104] and the Mean Decreasing Accuracy index calculated by Random Forests [105]. For each rank, the cumulative sum of the ranking score was computed and four different cuts based on the cumulative values were taken. Cuts considered all the features needed to maintain 60 %, 70 %, 80 % and 90 % of the cumulative value. These different groups of features were used to cluster patients in each single view (single view clustering step), with the same single view clustering algorithms used in the first step. The number of clusters K was considered as the number of classes. For each clustering, the error was calculated as the dispersion obtained in the confusion matrix between class labels and clustering assignments. The clustering algorithm that reached the minimum error for each view was then selected.

These clustering results were used as the input to the late integration step to obtain the final muti-view meta clusters. The integration was performed by using both Greene [95] and Long [94] approaches. Once the multi-view clusters were obtained, a subclass was assigned to each one. For each cluster, the number of objects of each class was calculated and the class with more representative patterns was assigned as the cluster label. Then, a p-value was calculated in order to verify the statistical significance of the subclass by the Fisher's exact test [106]. Experiments were performed in two ways: the former uses all the prototypes for classification; the latter uses only the most relevant

ones for class separability. Each one of these approaches were performed both in unsupervised and semi-supervised manners, respectively. The semi-supervised approach consists of giving a priori information as input to the techniques of late integration via a membership matrix of patients with the exact information of their classes. This information is combined with the membership of the patients compared to the single view clustering and integrated in meta-clusters. This can be a useful approach mainly when the data set is composed of unbalanced or under represented classes.

MVDA was compared with classical single view clustering algorithms (Kmeans, Hierarchical and Pam), early (TW-Kmeans) and intermediate (SNF) integration approaches. For each method clustering impurity, normalised mutual information (NMI) and cluster stability were evaluated. Cluster impurity was defined as the number of patients in the cluster whose label differs from that of the cluster. Given two clustering solutions the NMI was computed as the mutual information between the two clustering normalised by the cluster entropies. The NMI was computed between clustering results and real patient classifications.

The stability of the system was tested by giving different inputs to the algorithm and comparing the results. In order to perform the highest number of comparison, and avoid to have unbalanced patient classes, the dataset was altered with leave-one-out technique.

A test was run on the first step to generate a stability index for the prototypes of the obtained clusters. Then, the steps 2, 3 and 4 were evaluated jointly to assess the stability of the selected features and to evaluate the robustness of the multi-view clustering results. Furthermore, a borda-count method [107] was performed to find the final list of features selected over the leave-one-out experiments for the integration step. At the end of this process, N different clustering assignments were obtained, one for each removed patient. An matrix was created, where $M(i, j)$ was the normalised mutual information (NMI) between the clustering obtained removing patient i and the clustering obtained removing patient j. Then the mean of the matrix was calculated, indicating the stability measure of the method.

The method was tested on large genomic data sets including different omics

data types, such as the Cancer Genome Atlas (TCGA) data sets ([55]). The comparison experiments suggest that MVDA outperforms other existing integration methods, such as Tw-Kmeans and SNF.

3.3 Dataset collection and preparation

The experiments were performed on six genomic datasets (see Table 3.1). They were downloaded from The Cancer Genome Atlas (TCGA) ([55]), Memoral Sloan-Kettering Cancer Center ([108]) and from NCBI GEO ([56]) (See Table 3.1). Since all the data downloaded were already pre-process ed, only the batch effect was removed by the comBat method in the R âĂIJsvaâĂİ package [109].

For the dataset TCGA.BRC, the RNASeq and miRNASeq (level 3) data related to breast cancer patients, with invasive tumors, were downloaded from the TCGA repository (`https://tcga-data.nci.nih.gov/tcga/` - Breast invasive carcinoma [$BRCA$]). Patients were subsequently divided into four classes (Her2, Basal, LumA, LumB), using PAM50 classifier [110, 111].

mRNA (GSE22219) and microRNA (GSE22220) expression data related to breast cancer patients, from a study performed at Oxford University [112], were downloaded from Gene Expression Omnibus Dataset [56]. Patients were divided into four classes (Her2, Basal, LumA, LumB), using PAM50 classifier [110, 111]. This dataset was namend OXF.BRC.1. The same patients were then classified into four classes (Level1, Level2, Level3, Level4) using clinical data also retrieved from the same source. This dataset was namend OXF.BRC.2.

For the TCGA.GBM dataset, the gene and miRNA expression (level 3) data related to patients affected by Glioblastoma, were downloaded from the TCGA repository (`https://tcga-data.nci.nih.gov/tcga/` - Glioblastoma multiforme [GBM]). Also, clinical data was retrieved. The patients were divided info four classes: Classical, Mesechymal, Neural and Proneural as described in [113].

For the dataset TCGA.OVG the gene expression, protein expression, and miRNA expression (level 3) data related to patients affected by ovarian cancer, were downloaded from the TCGA repository (`https://tcga-data.nci.nih.gov/tcga/` - Ovarian serous cystadenocarcinoma [OV]). Clinical data were

Table 3.1: Datasets: Description of the datasets used in this study. "N" is the number of subjects for each dataset. $N(i)$ is the number of samples in the i-th class. An x denotes if that view (column) is available for a specific dataset (row).

Dataset	Response	N(0)	N(1)	N(2)	N(3)	Gene Expression	RNASeq	microRNA Expression	miRNASeq	Protein Expression	Copy Number	Clinical Data
Breast Cancer from The Cancer genome Atlas, N = 151												
TCGA.BRC	Pam50 (Her2,Basal,LumA,LumB)	24	13	55	59		x		x			
Breast Cancer from The Gene Expression Omnibus, N = 201												
OXF.BRC.1	Pam50 (Her2,Basal,LumA,LumB)	26	6	117	52	x		x				
OXF.BRC.2	Clinical (Level1, Level2, Level3, Level4)	73	54	42	32	x		x				
Prostate Cancer from Memorial Sloan-Kettering Cancer Center, N=88												
MSKCC.PRCA	Tumor stages T1 vs. T2, T3, T4	53	35			x		x			x	x
Ovarian Cancer from The Cancer Genome Atlas, N=398												
TCGA.OVG	Tumor stage I,II, Tumor stage III, Tumor stage IV	33	315	50		x		x		x		
Glioblastoma Multiforme from The Cancer genome Atlas, N = 167												
TCGA.GBM	(Classical, Mesechymal, Neural, Proneural)	37	54	24	52	x		x				

downloaded to classify patients in three categories. In particular patients were classified by clinical stage: first class: stage IA, IB, IC, IIA, IIB and IIC, second class: IIIA, IIIB and IIIC, third class Stage IV.

For the dataset MSKCC.PRCA, the gene expression, microRNA expression, copy number variation (CNV) and clinical data related to patients affected by prostate cancer, were downloaded from the Memorial Sloan Kettering Cancer Center ([114]). Clinical data were downloaded to classify patients in three categories. Patients were classified in two classes by using clinical data by the tumour stage: class one is Tumour Stage I and class two is Tumour Stage II, III and IV. Classification of patient was done according to a previous study performed on the same dataset [115].

3.4 Results

The MVDA method was compared with other multi-view clustering methods, such as SNF [93] and Tw-Kmeans [91]. Using TCGA datasets from 4 different

tumour types (Table 3.1), the cluster impurity error, the Normalised Mutual Information and the cluster stability of all the considered algorithms was evaluated. The evaluation metrics computed for each dataset are summarised in Table 3.3. The unsupervised version of MVDA, shows a mean error of 22.47%, normalised mutual information (NMI) of 28% and stability of 85%. Moreover, the error significantly decrease when using prior information. Indeed, the MVDA methodology applied with prior information reduces the error to 6.30%.

The other methods used in the comparison study show a higher mean error from the lowest $30,83\%$ of SNF to the highest $30,93\%$ of Kmeans. They also show a lower NMI (the maximum value reached is 26% of Ward's method) and variable stability from the lowest 51% of the Kmeans to the highest 96% of the partitioning around medoids (pamk).

A class label and a p-value was assigned to each cluster obtained after the integrative step. Figure 3.4.1 shows, as example, the results obtained for the dataset OXF.BRC.1. The label indicates the subclass to which patients in the cluster belong, while the p-value measures the statistical significance of a cluster. In the case of the dataset OXF.BRC.1, the patients are divided into four classes: LumA, LumB, Her2 and Basal. Eight relevant clusters were observed, four of which are subclasses of class LumA (cluster 4 - pvalue $2.51E^{-4}$; cluster 5 - pvalue $8.71E^{-8}$; cluster 6 - pvalue $2.92xE^{-3}$; cluster 11 - pvalue $1.97E^{-3}$) and two are subclasses of class LumB (cluster 2 - pvalue $3.93E^{-14}$; cluster 10 - pvalue $5.14E^{-3}$). The influence of each view on the final cluster is also reported. While it is obvious that the clusters are obtained considering all the genomic data views, the information needed to identify a specific subclass can be more relevant in a particular data type instead over the others. For example, the clusters 3, 6 and 11 of the OXF.BRC.1 dataset are both labelled as LumA. miRNA expression contributes for the 100% to define the cluster 11, the gene expression is mainly determining the cluster 3 (57%), while for cluster 6 they are equally important. This could mean, for example, that patients in cluster 11 are particularly characterised by miRNA expression while patients in cluster 3 by gene expression. For all the six datasets, the results showed that the matrix factorisation method gives lower classification error and better accuracy than

Figure 3.4.1: Multi-View Clusters Statistics for the OXF.BRC.1 dataset. For each cluster class label, the p-value and the view contribution are reported.

the approach with general linear integration

As shown in Figure 3.4.2, the integrative clustering performed generally better that the clustering on each single data view. In the TCGA.BRCA dataset, the mean cluster impurity is about 26% when patients are grouped by the gene expression and 43% when they are grouped by their miRNA expression profiles. However, combining the gene and the miRNA expression profiles, $26,50\%$ of error in unsupervised mode and 9% in semi-supervised mode are obtained, respectively. Only in a few cases, the patient grouping based on a single data view performs better than the one obtained with multiple data types.

Figure 3.4.3 depicts the comparison between the two integration methods, either with or without prior information. The matrix factorisation based method reaches the higher stability (about 85%) in all the cases. With respect to the cluster impurity, the difference is almost always negligible. The greatest difference occurs when passing from the unsupervised to the semi-supervised approach. The cluster impurity for the unsupervised clustering is about 30% and about 7% for semi-supervised. Therefore, for more accurate sub-typing of classes semi-supervised integration was used, which maintains high stability and reduces the classification error compared to the classes. However, in case of unbalanced patient classes, the prior information is needed to increase the prediction.

Since we tested different algorithms at each step of our methodology, we aimed at understanding if a common pipeline for all the datasets could be applied. After the execution of all the analyses, we observed that the best algorithms for the first and second steps strongly depend on the data. We found that K-means is the best algorithm for step 3 for the TCGA.BRACA, OXF.BRCA.1 and OXF.BRCA.2 datasets (Table 3.2).

At the last step, the matrix factorisation approach provided lower errors and greater stability as compared with the general linear integration methods on most of the datasets. This result corroborates our hypothesis that a late integration approach is better for it allows using the best algorithms for each data type.

In order to evaluate the performance of the proposed method, we systemat-

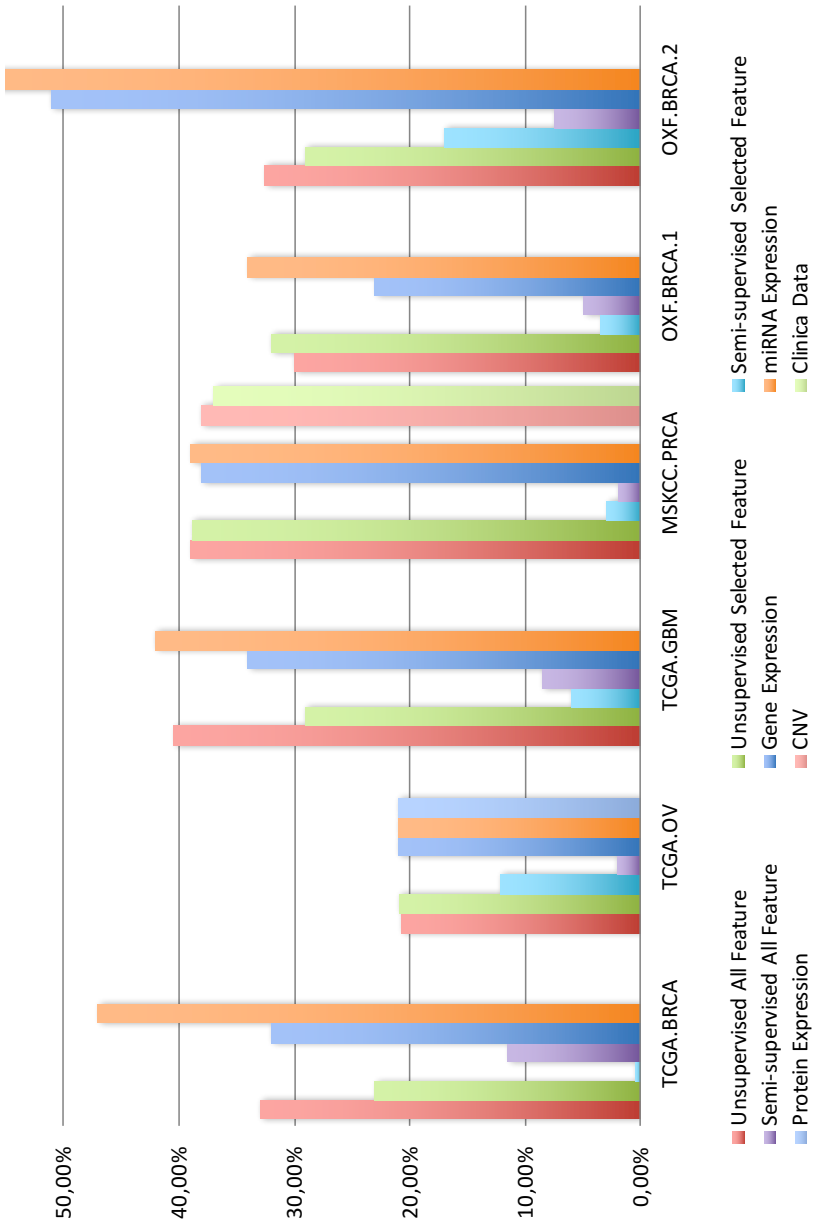


Figure 3.4.2: Cluster Impurity difference between single view and integration analysis. Errors decreased with the integration approach in particular when the semi-supervised methodologies were used.

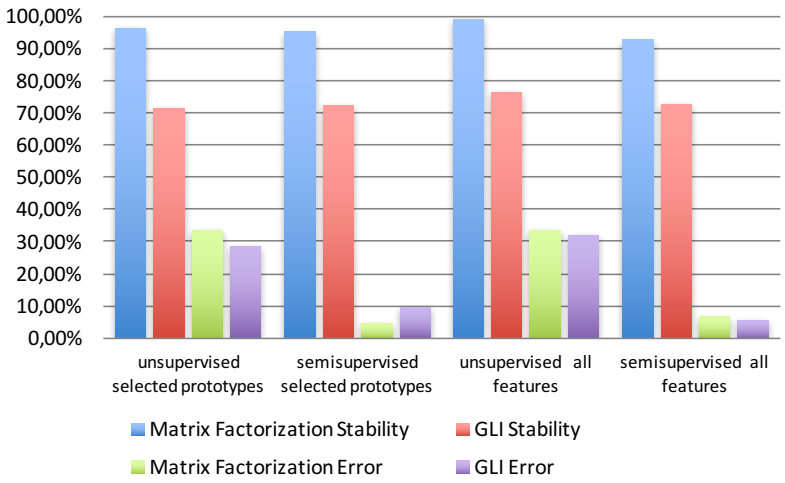


Figure 3.4.3: Difference between alternative integration methods: The mean cluster stability is reported. Clustering stability was calculated by comparing the unsupervised and the semi-supervised mode, both using either all the features or only the selected prototypes

Table 3.2: Summary of combination of algorithms for each view that give the best grouping of patients. The symbol (-) means that feature selection was not executed. Symbol (DM) means that same classification error was obtained with all the algorithms used.

Dataset	Views	(a) Feature Clustering	(b) Feature Selection	(c) Patients Clustering	(d) Late integration
TCGA.BRCA	RNASeq	Pam	CAT-score	Kmeans	MF
	miRNASeq	Pam	CAT-score	Pam	
TCGA.OV	Gene Expression	Pam	Random Forest	DM	MF
	Protein Expression	Pam	-	DM	
	miRNA Expression	Pam	-	DM	
TCGA.GBM	Gene Expressions	Spectral	CAT-score	Kmeans	MF
	miRNA Expression	Ward	-	Kmeans	
OXF.BRCA.1	Gene Expressions	Pam	Random Forest	Ward	GLI
	miRNA Expression	Pam	Random Forest	Kmeans	
OXF.BRCA.2	Gene Expressions	Pvcluster	CAT-score	Kmeans	MF
	miRNA Expressions	Pam	Random Forest	Kmeans	
MSKCC	Gene Expressions	Pam	CAT-score	Kmeans	MF
	miRNA Expressions	Pam	-	Pam	
	CNV	Spectral	CAT-score	Kmeans	
	Clinical	-	-	Pam	

Table 3.3: Validation Results: The mean classification error, normalized mutual information (NMI) and stability, on all datasets, are shown, measuring the agreement between the clusters resulting from an approach and the real patient classification. Bold font in percentage indicates best performance in the experiments.

	Feature	Integration	Algorithm	Error	NMI	Stability
Single View	All Feature	-	Ward	30,08%	26%	86%
		-	Kmeans	30,93%	25%	51%
		-	Pamk	30,75%	24%	94%
	Selected Prototype	-	Ward	30,72%	26%	89%
		-	Kmeans	30,36%	25%	52%
		-	Pamk	30,78%	24%	96%
Multi-View	All Feature	Early	Tw-kmeans	37,10%	24%	69%
	All Feature	Intermediate	SNF	30,83%	22%	83%
	All Feature in Cluster of Selected Prototype	Intermediate	SNF	31,31%	18%	82%
	Selected Prototype	Late / unsupervised	MF/GLI	**27,47%**	**28%**	**85%**
	Selected Prototype	Late / semi-supervised	MF/GLI	**6,30%**	**63%**	**84%**

ically compared it with Tw-Kmeans and SNF algorithms (Table 3.3). Anyhow, we did not compare our method with iClust, as it has been shown to have worse performance than SNF, with which we deal in this study [93]. We confirmed that late integration works more efficiently in integrating different views of genomic data. This is due to the large complexity and difference between the views. When views have different numerical and statistical characterisations, it is more convenient to individually analyse single data types and then combine the results in a multi-view analysis. This becomes more and more important as the number of views involved in the analysis increases.

3.5 Discussion

Biomedical research, gives focus on the identification of patients' subtypes to produce accurate diagnosis and targeted treatments in the field of precision medicine [69]. Initially this problem was addressed by identifying groups of patients who shared similar patterns of gene expression [116–118].

Thanks to the advancement of omics techniques, capable of producing data related to different molecular aspects of the cell, research has moved on supplementary techniques. In fact, efforts have been made in the use of multi-view

clustering techniques that identify subtypes of patients considering different information at the same time (for example, gene expression, mirna expression, protein expression, etc) [62]. Obviously, the joint use of different types of data poses different problems such as the use of suitable metrics to all experiments or the type of integration. To address some of these a new data integration methodology, called MVDA, was proposed. MVDA can integrate different omics experiments in a late integration manner with the aim of identify patients subtyping. The methodology is composed of four steps using state of the art algorithms. It was evaluated on six cancer benchmark datasets and compared with classical single view clustering algorithms and two state of the art multi-view algorithms: TW-kmeans and SNF.

TW-KMeans [91] and SNF [93] were selected for comparison with MVDA, because each of them represents a different data integration methodology: early - (TW-Kmeans), intermediate - (SNF) and late - (MVDA) integration. Moreover, TW-Kmeans is the multi-view version of the classical K-means algorithm, SNF uses a spectral clustering algorithm (based on k-means) applied to the fused kernels, while in MVDA the multi-view clustering is performed by using factorisation approaches. This is not a partitive clustering algorithm, but it evaluates the probability of each point to be part of each meta-clusters, and then assigns each point to the meta-cluster with the highest membership. Between all of theme, TW-Kmeans is the only method that was not proposed for clustering biological data. But, it has the advantage to have a double weighting scheme, meaning that two vectors are associated to the multi-view clustering solutions: the former specifying the contribution of each variable, and the latter specifying the contribution of each view. This is not the case neither for MVDA nor for SNF, in fact MVDA gives information only regarding the contribution of each view, while SNF does not give any information. On the other side, the late integration methodology proposed for MVDA, offers several significant advantages: (1) the optimal algorithm and similarity measure can be chosen with respect to each single view. Since the clustering solutions are strongly affected by these two factors, this point should not be underestimated. In fact, the data in each view can have not comparable distribution, and then applying the same

metric and algorithm in each view could not lead to the best solution. This is the main disadvantage of the TW-Kmeans method where all the views are concatenated together and treated as a unique bigger dataset. SNF solves the problem, because it first creates a kernel that can be specific for each view, and then fuses them together. (2) Moreover, each view can be processed independently from the others, then the process can be naturally parallelized; This is also the case for SNF. On the other hand, the computation with TW-KMeans can be slower because the number of features is higher, being the combination of all the views. The performed analysis shows that the integrative clustering outperforms the single view approaches on all the datasets. Moreover, MVDA prove to be stable on perturbation dataset analyses performed by executing clustering on perturbed datasets removing one patient at a time and evaluating the normalized mutual information between all the resulting clusterings. Furthermore, the analysis suggests that the use of a late integration technique lead to better classification results. This is probably because, with late integration technique, many operations can be performed before the integration, such as the best algorithms for dimensionality reduction, feature selection and patients clustering can be applied.

Chapter 4

Integrated Network of Systems bIology Effects of nanomaterials (INSIdEnano)

This chapter describes INSIdEnano (Integrated Network of Systems bIology Effects of nanomaterials), a novel tool for the systematic contextualisation of the effects of nanomaterials in the biomedical context. The methodology and the data used to construct the database and the tool are described in section 4.2. Also example of tool usage are reported.

4.1 Introduction

Due to their physical, electronic, and biological characteristics, engineered nanomaterials (ENMs) are increasingly used in a wide spectrum of applications such as energy production, vehicles construction, architecture, computers manufacturing, medicine and in various everyday consumer products [119, 120].

Interactions between nanomaterials and the humans seem to be inevitable

[121]. Indeed nanomaterials in the air, like nanoparticles emitted from laser printers, are easily inhaled through breathing and therefore affect the respiratory system [122, 123]. Others, instead, such as sunscreen or body lotions entering the bloodstream through the skin [124]. Others are even administered as drugs [125, 126].

Even if considerable advanced was performed in the last years, the understanding of the biological effects of the huge number of existing and emerging ENMs is partially unknown and there is still a lack of tools for the contextualisation of nanomaterials. Transcriptomic studies, performed with DNA microarray, may help in characterising the mode-of-action (MoA) of ENMs, opening new possibilities for their safety valuation based on systems biology approaches.

Systems biology approaches have already been applied to study complex phenomena. A well-established principle in biomedical research is the concept that the pattern of molecular alteration, of any phenotypic perturbation, can be used as its signature. By exploiting this principle, many computational strategies to explore similarities between drugs MoA and to perform drug repositioning task have been proposed [127–131]. Moreover systems biology approaches have been applied to study the interactions between different diseases [132] or to predict the toxicology of specific compounds [133].

Here the hypothesis is that similar strategies could be used to contextualise nanomaterials with respect to human diseases, drug treatments, and chemical exposures. A considerable amount of data related to these biological entities is already publicly available. The use of systems biology techniques to integrate and analyse them, could lead to a better understanding of the molecular effects of ENM. The idea is to compare the behaviour of gene mode of action (MoA) in cells exposed to nanomaterials with the one of cells treated with drugs or chemicals, and with the gene MoA of diseases. This goal was reached by constructing a network of interactions between the four phenotypic entities (ENMs, drugs, chemicals and diseases), where the nodes are the entities and the edges between them represents how similar or dissimilar is their gene MoA.

Moreover, the INSIdE nano tool was developed to scan the network in search of heterogeneous cliques (a clique is a completely connected sub-network) that

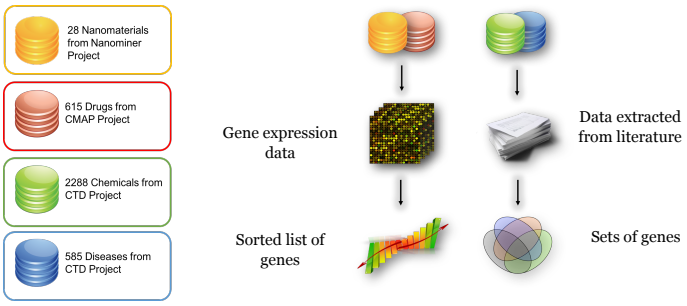


Figure 4.2.1: The data used in INSIdEnano

contains one ENM, a drug, a chemical and a disease. These cliques are the environment in which a nanomaterial can be contextualise. Indeed, by comparing its effects on the genes with respect to the one of a drug that cure the disease, one can hypothesise that the nanomaterial can be used as adjutants of the drugs to treat the disease if their effects on the genes are the same. Moreover, a nanomaterial can be supposed to be toxic if it has the same effect on the genes such as the one of a chemical that cause the disease.

4.2 Materials and Methods

INSIdEnano is a graphical tool able to contextualise the molecular effects of nanomaterials perturbations by systematically comparing their pattern of molecular alterations with different types of perturbations (diseases, drugs, etc).

4.2.1 Input Data

INSIdEnano was design to integrate a number of public available dataset related to nanomaterials, drugs, chemicals and diseases. See figure 4.2.1 part A. Raw gene expression data related to in-vitro experiment about the exposure of 28 nanomaterials (see Table 4.1 for details) to different human cell was downloaded from the NanoMiner database web page [134]. The raw data files (coming from

three technologies: Affimetrix, Illumina and Agilent) were imported into R and pre-processed by using a nested batch effect removing process. Figure 4.2.2 shows the work-flow of the pre-processing framework. First, probes of each chipset were re-annotated according to NCBI Entrez Gene Database [135, 136].

For each chipset background correction and probe summarization was performed. This resulted in a separate expression data matrix for each microarray study that were integrated in a big data matrix that was further quantile normalised. ComBat [147] and SVA [148, 149] were used for removing known batch effects and other unwanted variations in each microarray study. In details, ComBat was exploited to adjust each microarray study for known batch covariates, such as the dye-effect, as well as the slide and the array position. The model matrix used in ComBat included the exposure as the variable of interest and other covariates, such as the cell type, the exposure time and the dose to be preserved during the batch effect removal. Moreover, the microarray studies were also adjusted for the batches/covariates variables related to any pre-treatments. However, covariates/batches confounded with the outcome of interest were not considered. Principal component analysis (PCA) was used to investigate the variance in each expression data matrix before and after applying ComBat, followed by ANOVA analyses to explore the associations between the covariates of interest and the most relevant components. Indeed, after removing batch effects with ComBat, if the ENM exposure variable of interest was not significantly associated to the first principal component, the SVA method was used. The SVA algorithm allows identifying and removing unknown and sources of variation while protecting the variance correlated to the variables of interest. SVA was exploited to discover batch effects, which were subsequently removed using ComBat. Gene expression data for drug treatments was downloaded from the Connectivity Map (CMap) web page [127]. Raw data for 615 drugs was downloaded and pre-processed as in Napolitano et al [129]. Raw data files were quality checked in order to discard suboptimal data points. Then, the probes were re-annotated using the NCBI Entrez Gene database similarly to the process used in the NanoMiner data set, and the final matrix was normalised with the quantile method from the RMA algorithm, as described above. Next, the

Table 4.1: Nanomaterials description

Name	Description	Ref
TiO2T200	Submicron TiO2(T-200)	[137]
TiO2T20	Submicron TiO2(T-20)	[137]
TiO2T7	Ultrainfe TiO2(T-7)	[137]
TiO2NB	Titanium Nanobelts	[138]
TiO2	NanoTiO2 (Titanium)	[139]
RS	Eudragit RS nanoparticles	[140]
AuNP	Gold - EGFP oligonucleotide complex	[141]
ZnO	Zinc Oxide - IBU-tec advanced materials AG	[139, 142]
ZnO-1	Zinc - IBU-tec advanced materials AG;	[142]
ZnO-2	Zinc - mandelic acid coated;	[142]
ZnO-3	Zinc - mercaptopropyl-trimethoxysilane coated	[142]
ZnO-4	Zinc - methoxyl coated	[142]
ZnO-5	Zinc - diethylene glycol modified;	[142]
ZnO-9	Zinc - folic acid modified	[142]
ZnCl2	Zinc chloride	[139]
GSNO	S-nitrosoglutathione	[143]
UP	Ultrafine Particle	[144]
MWCNT	Multi Wallet Carbon Nanotube	[138]
CBNP	Nano-carbon Black	[139]
SiO2	Silicon dioxide	[139]
PSNP	Polysthylene nanoparticles	[145]
AgNP	Silver nanoparticles	[145]
Ag2CO3	Silver nanoparticles	[145]
Al203	Aluminum Oxide	[139]
microZnO	microZnO	[139]
WC	WC nanoparticles	[146]
WCCo	WC-Co nanoparticles	[146]
Fe203	Iron (III) oxide	[139]

Figure 4.2.2: Outline to the microarray data integration steps at interpretative level.

batch effect was estimated and removed by using the ComBat algorithm.

Manually curated information about chemical-gene and disease-gene interaction were retrieved from the Comparative Toxicogenomics Database (CTD) website [150] for 2288 chemicals and 585 diseases. For each disease-gene interaction, a score representing the strength of association is provided. The distribution of these scores was investigated and used to define a threshold to filter out associations. A connection between a disease-gene and chemical-gene was considered reliable when its strength of association was higher than the 95th percentile of the overall distribution of the scores. Disease-genes associations are not based on gene expression data. On the other hand, the connections between chemicals and genes indicate whether the genes are up or down regulated by the chemicals. Moreover, only the connections between chemicals and genes from the human genome were considered.

4.2.2 Integration Process

The fact that the a phenotypic perturbation can be identify by its pattern of molecular alteration is a well-established idea in biomedical research. The comparison of such molecular signature have been used to develop many computational strategies in order to explore similarities between drugs MoA and perform drugs repositioning tasks [127–130].

The main effort in this kind of analysis is the identification of a good strategy to perform these comparisons. For example Lamb et al. [127] created a collection of 563 gene expression profiles, called "Connectivity Map", with the aim of systematically characterise small-molecule perturbations (and their corresponding vehicle controls) in human cell lines and infer functional connections between them, diseases and drugs action. They identify each expression profile in their dataset, with a query signature consisting in a list of genes correlated with the biological state of interest. The list is ranked, from the most up-regulated to the most down-regulated gene, according to their differential expression relative to the control. Then, each list of genes is compared to the other by means of a nonparametric strategy based on the Kolmogorov-Smirnov statistic as they described in [151–153].

The hypothesis is that similar strategies could be used to contextualise nanomaterials with respect to human diseases, drug treatments, and chemical exposures. In fact, to compare nanomaterials exposure with drug treatment, each nanomaterials and drugs in the INSIdE nano was represented by a ranked list of genes resulting from the differentially expressed analysis relative to their controls. The genes are ranked from the most up- to the most down- regulated according to the following score: $\pm logFC \cdot -log(Pval)$ where $logFC$ is the logarithm of the genes fold change and $Pval$ is the adjusted p-value (the adjustment is performed with the FDR method) coming from the analysis. The similarity between each couple of lists were evaluated by using the Kendall Tau Distance [154], that is a ranked based non parametric strategy already used in bioinformatics to compare ranked gene lists [155, 156].

On the other hand, each chemical and disease available in INSIdE nano was represented by a set of genes with no prior information related to their order. See figure 4.2.1 part B. To compare chemicals with diseases the Jaccard Index was used. The Jaccard index has already been used in literature to compare sets of genes alone [157, 158] or integrated with other similarities measures [129].

The Gene Sets Enrichment Analysis (GSEA) [153] was used to compare the similarity between nanomaterials or drugs, represented by ordered lists of genes, and chemicals or diseases represented by sets of genes (See Figure 4.2.3).

Kendall Tau distance

The Kendall Tau distance between two ranked lists T1 and T2 is defined as follow:

$$K(\tau_1, \tau_2) = \sum_{\{i,j\} \in P} \bar{K}_{i,j}(\tau_1, \tau_2) \tag{4.2.1}$$

where P is the set of unordered pairs of distinct elements in τ_1 and τ_2, $\bar{K}_{i,j}(\tau_1, \tau_2) = 0$ if i and j are in the same order in τ_1 and τ_2, $\bar{K}_{i,j}(\tau_1, \tau_2) = 1$ if i and j are in the opposite order in τ_1 and τ_2. Its values range between 0 and $n(n-1)$, where n is the length of the list. A value of 0 means that elements in the list are in the same order; A value of $n(n-1)$ means that elements in the list are in the opposite order. The values were normalised in the range $[-1, 1]$, where

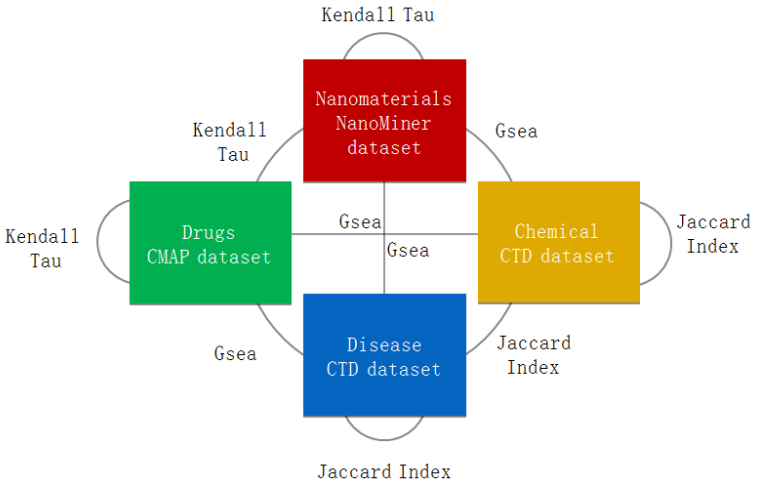


Figure 4.2.3: Integration Process

-1 means that the two list have opposite order, and 1 means that they have the same order.

Jaccard Index

The Jaccard Index between two sets A and B is defined as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

This measure is 0 if the intersection between A and B is empty, while it is 1 if it contains exactly the same elements. For each chemical, two sets of genes were considered: those whose expression is up-regulated and those whose expression is down-regulated by the chemical exposure. For the down-regulated genes, the Jaccard Index was multiplied by -1 in order to take into account the effects on the genes.

Gene Sets Enrichment Analysis

The GSEA was implemented by using the Kolmogorov-Smirnov test [159, 160]. It is used to compare a sample distribution with a reference probability distribution. The empirical distribution function F_n for n *iid* observations X_i is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I[-\infty, x](x_i) \tag{4.2.2}$$

where

$$I[-\infty, x](X_i)$$

is the indicator function defined on a set X that indicates the membership of an element to a subset A of X, having the value 1 for all elements of A and the value 0 for all elements of X not in A. The Kolmogorov-Smirnov statistic for a given cumulative distribution function $F(x)$ is

$$D_n = sup_x|F_n(x) - F(x)|$$

As in [128], the Kolmogorov-Smirnov statistic was used without the absolute value in order to preserve the sign. This helps understanding if the genes in the sets are up or down-regulated. The value of the statistic is used as weight for the edges in the network.

Data Normalisation

These three measures were used to build a pairwise similarity matrix between all the considered elements. The distributions of the values coming from the three similarities were quite different and had different ranges. In order to make them comparable they were transformed in uniform distributions in the range $[0, 1]$ by means of the cumulative function as shown in Figure 4.2.4. In fact if a distribution X has the (cumulative) distribution function $F(x) = P(X < x)$, then $F(X)$ has a uniform distribution on $[0, 1]$. The signs have been left unchanged to keep track of similar (+) or dissimilar (-) behaviour between each pair of elements.

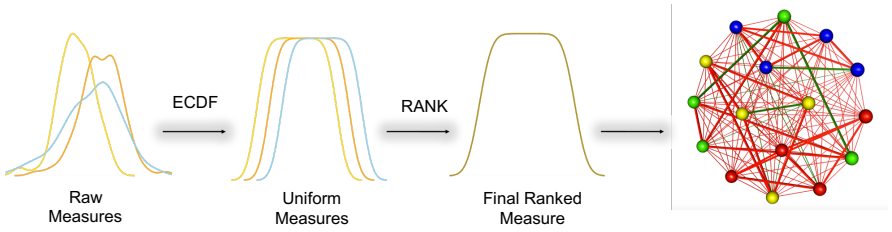


Figure 4.2.4: Normalization Process

Similarity Network Inference

A well-established concept in biomedical literature is to represents complex phenotypes and their interactions as network and to make inferences about them by using network theory results [130, 132]. Then the pairwise similarity matrix, obtained from the previous step, was used as an adjacency matrix to build a final interaction network, that represents the core of INSIdEnano that has 3,516 nodes and 12,362,256 edges. Because of the network is completely connected, there is a need to apply a threshold to retain only the most relevant connections. Instead of apply a predetermined threshold, a ranking system was applied. See figures 4.2.5 and 4.2.4. For each vertex, its neighbours were ranked based on the similarity score, then when the network is queried the user can set a percentage of the top edges to select (e.g. first 10%,20%,30% of the rank). However, rankings are not symmetric, meaning that if a node A is in the top $x\%$ of the nodes connected to node B, is not always true that B is in the $x\%$ of the nodes connected to node A. To solve this problem, a more stringent threshold is applied by computing the mutual neighbourhood of a node. It is defined as:

$$\mathcal{N}(i) = \{j : rank_i(j) \leq th \wedge rank_j(i) \leq th\}$$

where $rank_i(j)$ is the position of node j in the ranked list of nodes connected to i and th is the user defined threshold.

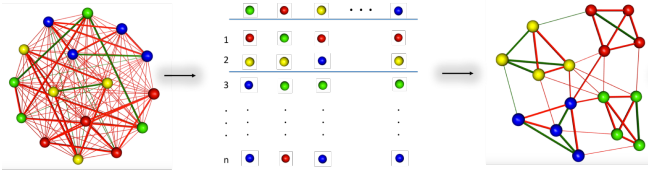


Figure 4.2.5: The complete connected network is pruned by using a method based on ranked list

4.2.3 Validation of the Similarities Measures

The pairwise phenotype similarities, that are based on the MoA, were compared with other independently similarities based on different characteristics such as the molecular structure of the drugs and chemicals, the symptoms of the diseases, the use in clinical practice of a drug to treat certain diseases and the pathogenic exposures to a chemical that can cause a disease.

The idea to validate the similarity measures used, by comparing them with other independent, is not a new technique in systems biology. For example, Zhou et al. [132], compared the similarities between the diseases based on symptoms, with similarities between the diseases based on genes affected by them. They verified that the overlap between the two measurements was significant by performing a permutation test.

Here the comparison was performed by means of the Mantel Test. The Mantel test is used to evaluate the correlation between two similarity matrices by adopting a procedure that is a kind of permutation test [161]. The drugs smiles were downloaded from the Drug Bank Database (`https://www.drugbank.ca/`). The chemical smiles were downloaded from the Chemspider Database (`http://www.chemspider.com/`). The pairwise similarity between drugs and chemicals smiles were evaluated with the Optimal string alignment algorithm implemented in the R stringdist package [162] The associations between drugs and diseases based on their clinical usage were downloaded from the MEDI Prescription Database (`https://medschool.vanderbilt.edu`) [163, 164]. The associations between chemicals and diseases were downloaded from the Com-

	INSIdEnano	Others	Coverage
Disease - Disease	585	426	$72,82\%$
Drug - Drug (Target)	615	410	$66,67\%$
Disease - Disease	615	608	$98,86\%$
Chemical - Chemical	2288	2236	$97,73\%$

Table 4.2: For each category of object the number of elements in INSIdE nano is reported (column INISIdEnano). Moreover, the number of element used to perform the Mantel test (column Others) and its percentage of coverage (column Coverage) is shown.

parative Toxicogenomics Database (`http://ctdbase.org/`). The similarities between diseases comes from a recent study published in Nature by Zhou et al. [132] where they used a biomedical literature database to construct a symptom-based human disease network and investigate the connection between clinical manifestations of diseases and their underlying molecular interactions.

The other similarity are not available for all the elements in INSIdE nano, but they are numerous enough to cover at least 60% of objects in each category. The exact number of element used to perform the Mantel Test is reported in Table 4.2.

4.2.4 Nanomaterials characterisation

Graphs (or networks) can efficiently represent complex phenomena and they can be rapidly analysed with ad hoc algorithms that consider the topological relatedness of their constituents. The main hypothesis is that patterns of similarity between sets of phenotypes could be used as an indication of biological association. In previous studies network based models were used to perform drug repositioning tasks [130] and to construct network of interactions between diseases based on their symptoms [132]. Both the works used substructure of the network to make inferences between the entities.

For example Iorio et al. [130], starting from transcriptomic data related to drugs treatment on human cells, constructed a network of interactions between drugs to characterise their mode of action. Each drug was represented by the

ranked list of genes sorted by their differential expression value with respect of their control. Then the similarities between each couple of drug, that represents the edges of the network, were computed by using the Inverse Total Enrichment Score (TES) that is based on the Kolmogorov-Smirnov test [159, 160]. Then they scanned the network in search of communities, to identify groups of drugs with similar effects. Moreover, to repositioning a new drug, the distances between its molecular alteration pattern and the one of the drugs in the communities were calculated. Then the drugs were predicted to have the same behaviour of those in the closest community.

Zhou et al [132], constructed a Human diseases network based on symptoms similarity. They parsed thousands of research articles in PubMed [165] related to diseases, computed the term frequency of each symptom (Mesh term[166]) associated to each disease, and then used a bipartite network projection method to compute similarities between diseases based on how many symptoms did they share. Then they created a network of diseases interactions based on how many genes or proteins are shared between each couple of disease. Then, in order to identify similarities between diseases they used global measures coming from network theory to characterise couple of diseases or disease communities. For example, they used the Dijkstra's algorithm[167] to compute the shortest path between diseases, then using this information as a dissimilarity measures they clustered diseases with the complete linkage algorithm.

The idea behind INSIdE nano is a bit different from the others. While in both Iorio's and Zhou's works the idea was to scan the network in search of homogeneous groups of nodes or to characterise nodes based on network theory global measures. Here, instead, INSIdE nano was scanned in search of heterogeneous clique sub-networks, that are quadruplet structures of heterogeneous nodes (a disease, a drug, a chemical and a ENM) completely interconnected by strong patterns of similarity (or anti-similarity). Figure 4.2.6 reports the pseudo-code of the cliques search method.

The study of the interactions between the elements in the clique allow to make inferences between the similarities in the behaviour of nanomaterials and the other elements. For example, the interactions with diseases, can suggest

Procedure CliqueSearch:

Input: nanomaterials, drugs, chemicals, diseases
Output: list of all the heterogeneous cliques of size k=4. Each clique contains a nanomaterial, a drug, a chemical and a disease.

```
CliquesList ← list()

for n in nanomaterials:
    drugs_n ← drugs connected to n
    for dr in drugs_n:
        chem_dr ← chemicals connected to dr
        for c in chem_dr:
            disease_c ← diseases connected to c
            for d in disease_c:
                if (n, dr, c, d) is a clique and (n, dr, c, d) is not in CliquesList:
                    add (n, dr, c, d) to CliquesList
                end if
            end for
        end for
    end for
end for

return CliquesList
```

Figure 4.2.6: Clique search pseudo code.

associations between the patterns of gene alteration due to the disease and the one due to the nanomaterials. With respect to drugs it is possible to make inference related to the nanomaterials drug-ability and their use in therapeutics: a nanomaterial can be a co-adjutant for a specific drug to treat a disease, if the nano and the drug have the same effect on the genes affected by the disease. On the other hand, a nanomaterial can be thought to be toxic, if it behaves like a chemical that cause a specific disease.

It is interesting to note that the approach used in nano inside, allows the use of transcriptome data for in-vitro experiments, to infer the effect of nanomaterials on humans. This is a big advantage as it eliminates the cost and the time required to carry out experiments in vivo.

INSIdE nano tool description

INSIdEnano is a web-based tool (available at `http://inano.biobyte.de`) that highlights connections between phenotypic entities based on their effects on genes. All the data prepossessing and the integration strategy was implemented in R. The graphical tool and the routine to scan the network were implemented in Python and Javascript using the d3 library for the Graphical User Interface (GUI). The system was developed in a client-server structure: the client is responsible for managing the user interface, collecting the user input and displaying the outputs. The server, instead, processes the data from the database according to the user inputs, and outputs the results to the client.

The system implements two major types of functions: the former a query analysis of the phenotypic network, where the user can retrieve specific information about selected items, the latter is an exploratory analysis of the phenotypic network. A complete tutorial is available on-line at `http://inano.biobyte.de/help.cgi`. The tool provides two different types of queries. The former, called simple query, allows the user to investigate connections of a specific element in the network. Given a node and a threshold, the tool shows all its neighbours divided into four categories: nanomaterials, diseases, drugs and chemicals. The latter, called conditional query analysis, allows the users to query the network by applying different filters and search for the cliques and it is the core of

the tool. Since the purpose of the analysis is to compare the behaviour of an element, with respect to the others, the user must specify at least two different kind of item. Moreover, the level of similarity necessary to report a connection between selected items, the number of items that must be in the same resulting cliques and the number of query items being connected to the other nodes in the sub-network, are requested as input. First, the tool retrieve the sub-network of all the elements, connected to the query items, that satisfy the user requirement. Then it scans the network in search of cliques. The cliques can contain three heterogeneous elements, that will be any one of the possible combinations of three elements between nanos, drugs, chemicals and diseases in the sub-network (eg. a nano, a drug, a chemical; a nano, a drug, a disease; etc), or they will contain exactly 4 elements (a nano, a chemical, a drug and a disease). Those cliques are then grouped with respect to the nature of the connections between each couple of items that they contain. As a result of the analysis the tool give the opportunity to visualise the sub-network of all the nodes connected to the query inputs that satisfy the user requirements. It displays the list of all the cliques with the opportunity to the user to analyse each one of them and to inspect the genes underlying the connections. Moreover, there are direct link to external source of information that the user can follow to have deep insights into each phenotype.

Thus, INSIdE nano is thought to be an exploratory tool through which the user can make new inferences about phenotypic connections. The inference process can be cyclic, as shown in 4.2.7. The purpose of the analysis is to characterise the behaviour of nanomaterials using the conditional query tool. This tool takes as compulsory input objects of at least two different classes (nano, drug, chemical or disease). If the user wants to investigate only an object, he may identify other entities of interest through the single query tool and then use the conditional query. On the other hand, if the user wants to investigate the connections between two (or more than two) objects, he can go to the conditional query tool directly. The outputs of the tool will enable the user to verify his hypothesis (sub-network and lists of clique) and to study the phenotypic entities in more detail (external link). This process can, however, lead to the

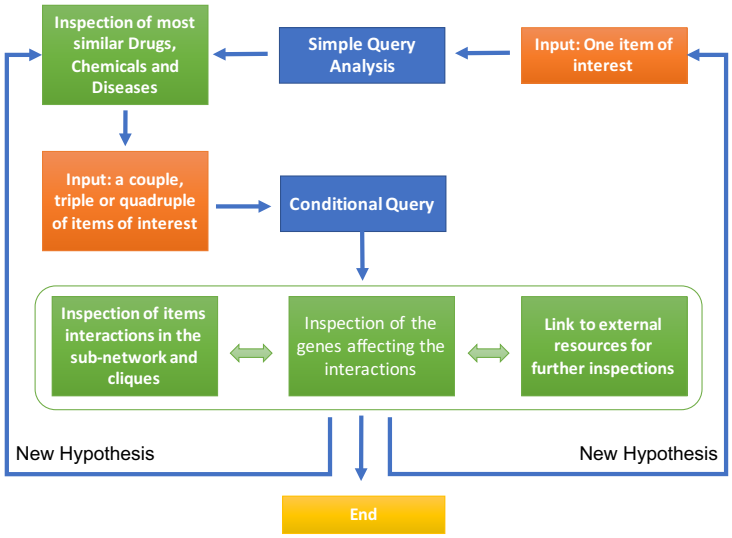


Figure 4.2.7: INSIdE nano workflow

consideration of new hypotheses, and then the process can be repeated more and more refining the initial analysis.

A tool with a similar exploratory analysis is ToxEvaluator. It was developed by the Pfizer company to facilitate the interpretation of toxicity findings for chemicals substances by using prior knowledge [133]. It includes proprietary resources such as in vitro pharmacological activity, in vivo preclinical toxicology study findings and more that 4 million compound structure libraries. It uses a proprietary tool to link intended pharmacological targets with toxicology-related gene information (ToxReporter [168]), computes compounds similarity and includes functions to predict unintended pharmacology (Polypharmacology). The Polypharmacology tool consists of a collection of 771 (one for each target Incorporated in the model) individual classifiers trained with internal screening data and external bioassay data [169, 170]. Moreover, given a compound the tool give in output a list of similar compounds based on the Tanimoto similarity [171] algorithm applied to their structure-encoding molecular fingerprint descriptors

[172].

The main difference with INSIdE nano is that ToxEvaluator tool is target specific, in fact, given an input compound it returns the probability that this compound will bind the targets present in their dataset. On the other hand, INSIdE nano is based on a different principle: the genes are used to evaluate the similarities between object and the identification of toxicity or anti-toxicity effects of the nanomaterials is performed by compare the effect of the nanomaterials on all the gene (present in the studies) with the one of other phenotypic entities.

4.3 Results

4.3.1 Network Description

The final network underling the INSIdE nano tool is composed by 3.516 nodes and 12.362.256 edges. Nodes are divided into four categories: 28 nanomaterials, 615 drugs, 2288 chemical and 585 disease.The number of known connection between diseases and drugs is 3.383 ($0,94\%$ of the total) and the number of known connections between diseases and chemicals is 8.960 ($0,67\%$ of the total).

The edges weight distribution was investigated. It is comprised between -1 and 1, where 1 means that the molecular alteration pattern of two entities is the same, and -1 meaning that it is opposite. The distribution is shown in figure 4.3.1 from which is easy to see that the majority of the edges have positive sign.

Moreover, the properties of the network were investigated. The clustering coefficient [173] and the degree distribution were investigated. Since the network is interrogated with different user-selected thresholds, the properties of the networks were calculated to vary the threshold (from 10% to 90% in steps of 10). The clustering coefficient increases as the threshold (Table 4.3). This means that the more edges in the network, the higher the probability that the nodes are grouped into clusters. Moreover, across all the thresholds, the average path length is quite small (Table 4.3). This suggests that when using a higher threshold the network behind INSIdE nano starts having the properties of a

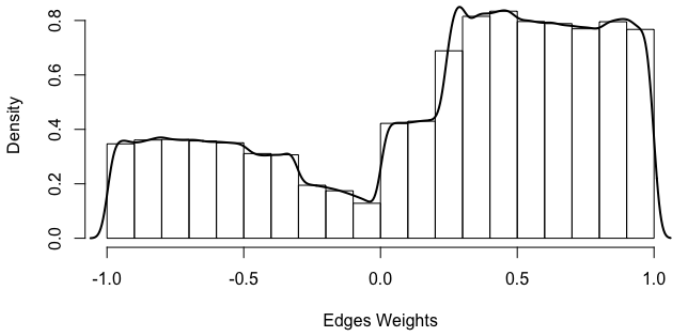


Figure 4.3.1: INSIdE nano weight distribution

small-world network.

To test if the INSIdE nano network has small-world characteristics, its clustering coefficient C and average path length L, were compared with those of random networks, that are known to have small clustering coefficient and short average path length [173]. For each different threshold, 100 random networks were generated by re-sampling the edges in the original network to be used as null-model networks. The average of the mean shortest path length L_r and clustering coefficient C_r over the null-model networks were evaluated. Then the normalised shortest path $\lambda = L/L_r$ and normalised clustering coefficient $\gamma = C/C_r$ was evaluated.

The resulting λ are approximately equal to one for all the thresholds (Figure 4.3.2, part A) and γ are greater than one for all the threshold (Figure 4.3.2, part B), meaning that the INSIdE nano network is small-world for each threshold. In fact, having $\lambda = 1$ means that the mean shortest path of INSIdE nano network is equal to the one of the random networks (that is known to be short). Moreover having $\gamma > 1$ means that the clustering coefficient of the INSIdE nano network is higher that the one of random networks (that is usually small).

Furthermore, for each threshold (from 10% to 90%) the hubs in the network

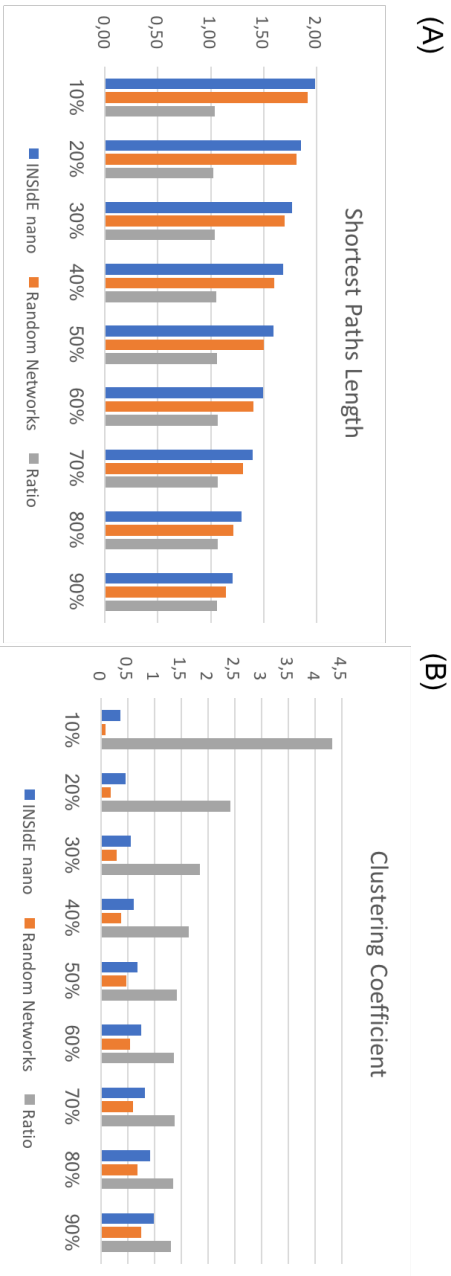


Figure 4.3.2: The first barplot (A) shows the average shortest path length in INSIdE nano network and in the 100 random networks at the varying of the threshold. Their ratio is always almost equal to 1. The second barplot (B) shows the clustering coefficient of the INSIdE nano network and in the 100 random networks at the varying of the threshold. Their ration is always greater that 1.

Th	10%	20%	30%	40%	50%	60%	70%	80%	90%
ClCoef	0.37	0.42	0.56	0.62	0.68	0.75	0.83	0.92	0.99
PathLen	2.00	1.86	1.77	1.68	1.59	1.50	1.40	1.30	1.21

Table 4.3: INSIdE nano network properties. The clustering coefficient (ClCoef) and the average path length (PathLen) are reported in the table. The network properties have been evaluated for different thresholds.

Chemicals	NTimes	Drugs	NTimes
C547593	9	clindamycin	8
1-3MeOPh-3-2MeOOHPhC3	9	bufexamac	7
Torin1 cpd	9	bupropion	7
1-AI-1,5-DCA	9	dihydrostreptomycin	7
MPTP	9	trifluridine	7
Diseases	NTimes	Nanos	NTimes
Akathisia, Drug-Induced	6	AuNP	4
Aphasia	6	PSNP	4
Asthenia	6	TiO2T20	3
Atrial Flutter	6	TiO2T7	3
Back Pain	6	AgNP	2

Table 4.4: The first five hubs for each category are reported

were identified. Their consistency between the different thresholds was evaluated 4.3.3. More than 55% of items were identified as hubs in 5/9 thresholds. The most stable hubs were the chemicals, followed by drugs, the diseases and then nanos. In table 4.4 the first five hubs for each class of phenotypic entities are reported.

4.3.2 Nanomaterials Sub-network

Even though the aim of this work is to compare the molecular alteration patterns of nanomaterials with other patterns, an analysis of the interactions between the nanomaterials in the network was conducted. As we can see from Figure 4.3.4 part (A) nanomaterials made of the same elements are closer in the network. For example, there is a group of zinc based nanomaterials that are completely

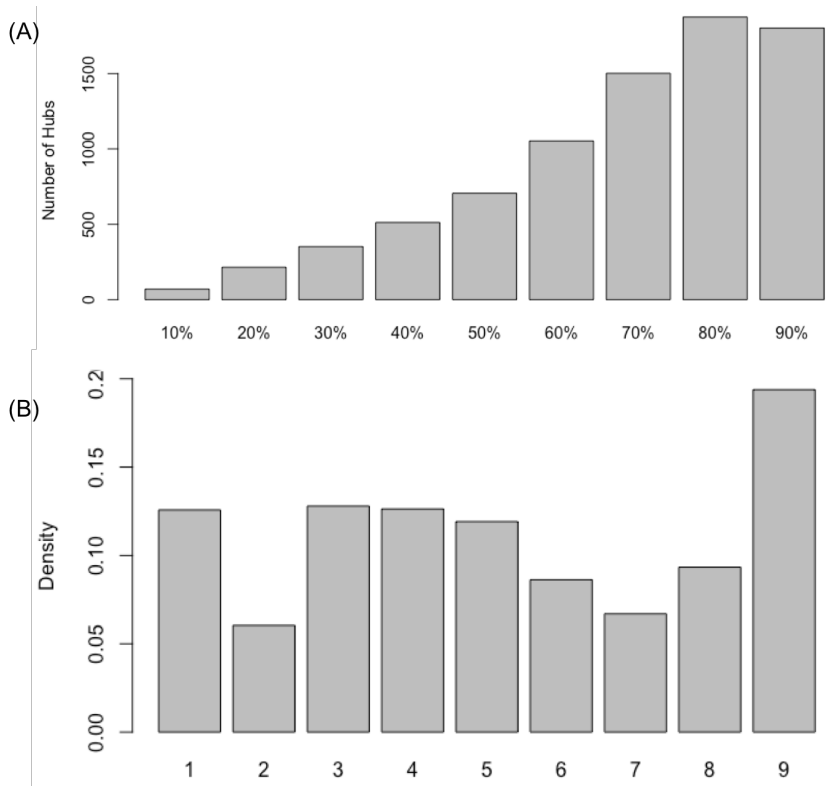


Figure 4.3.3: The barplot (A) shows the number of hubs for each threshold. The barplot (B) shows the percentage of items that were consistently considered hubs in 1 threshold, 2 thresholds, 3 thresholds and so on.

connected between them as well as a group of titanium based nanomaterials. The same behaviour is shown by the hierarchical clustering of the nanomaterials reported in Figure 4.3.4 part (B). This reinforces the belief that the similarity measure used (Kendall Tau distance) to compare the molecular changes pattern of nanomaterials is suitable for the analysis.

4.3.3 Nanomaterials-Drugs connections

The similarities between ENMs and Drugs were analysed by counting how many times each ENMs was connected with a drug of a certain ATC code. The Anatomical Therapeutic Chemical (ATC) Classification System is used for the classification of active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. It is controlled by the World Health Organisation Collaborating Centre for Drug Statistics Methodology (WHOCC) [174]. For each nanomaterial, the number of connections between each class of drugs was computed. Positive and negative connection were treated separately. Each count is normalised by the number of drugs in each category. The results is shown in the stack plot in Figure 4.3.5.

4.3.4 Nanomaterials-Disease connections

It was also possible to analyse the similarities between nanomaterials and diseases. For example, for each nanomaterial it can be interesting to investigate which is the most strongly correlated disease of a certain category. Figures 4.3.6 and reports the most strong associated respiratory disease to each one of the 28 nanomaterials included in INSIdE nano.

4.3.5 Connections Validation

To validate the connections between the different phenotypic entities, their similarity matrices were compared with other similarities matrices based on different metrics independently computed by means of the Mantel Test. Given n objects and two similarity matrices S_1 and S_2 the Mantell Test is used to evaluate if the two matrices are correlated or no. The null hypothesis test is that there

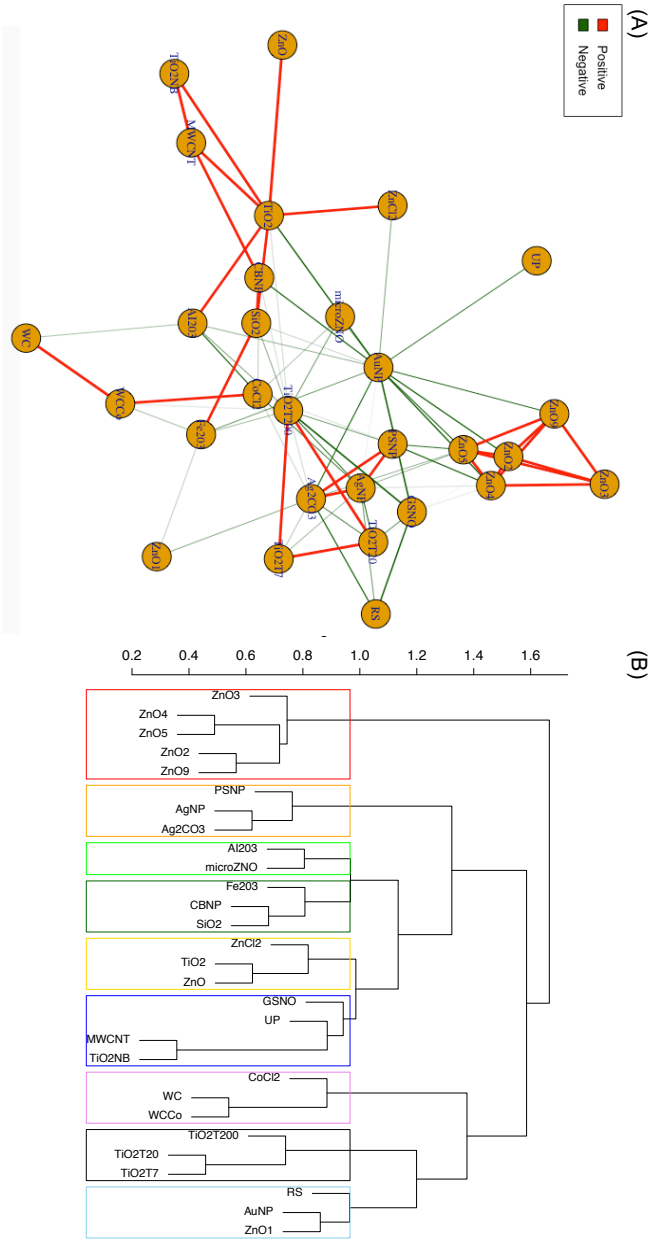


Figure 4.3.4: Nanomaterials sub-network (A). Red edges mean positive connections, while green edges mean negative connections. (B)Hierarchical clustering of Nanomaterials based on the Kendal Tau Similarity

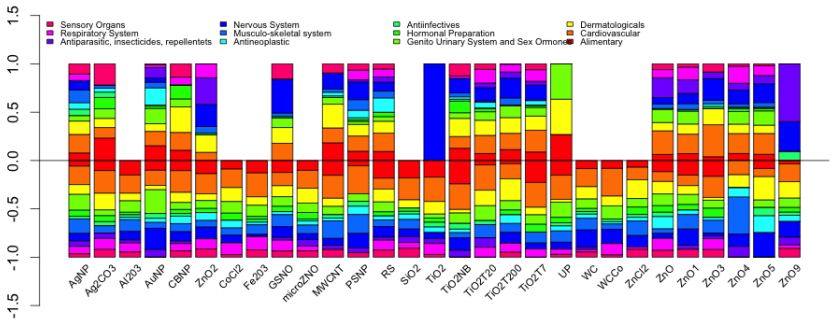


Figure 4.3.5: Number of connections between nanomaterials and drugs

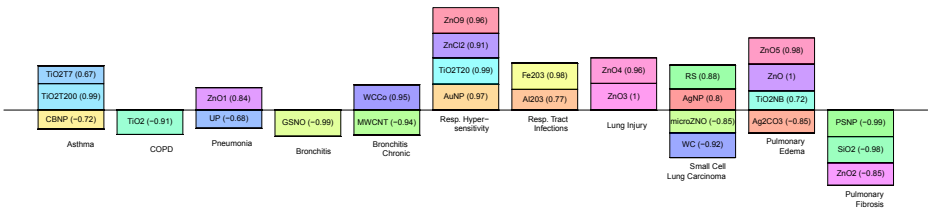


Figure 4.3.6: The bar plots show the association of ENM MoA and the molecular alterations for a set of respiratory diseases. For each nanomaterial, the most strongly connected disease with respect to the network is depicted. The connection strength is represented in each bar by height.

is no relation between the two matrices. First of all the correlation between the $n(n-1)/2$ distances is evaluated, then the rows and column of the matrix are permuted and the correlation is evaluated for each permutation. The significance of the observed correlation is the proportion of such permutations that lead to a correlation coefficient higher than the first one. The Mantel Test confirmed the biological relevance of the phenotype similarities calculated based on the MoA compared with other similarities independently computed considering alternative characteristics, such as the symptoms for the diseases (Mantel's test $P < 1E-05$), the molecular structure of the drugs (Mantel's test $P < 0.01$) and chemicals (Mantel's test $P < 1E-05$), respectively (See Table 4.5). This is an important result, in fact, it shows that the measures used to evaluate the similarities between objects are reliable, unlike what happened in other works, such as in Iorio et al. [130], where they found no significant correlation between measures built starting from the lists of genes and the information on the chemical structure of the medications. Moreover, a Kolmogorov-Smirnof test was performed by comparing the ranked lists of the drugs-diseases connections and the chemicals-diseases connections (based on the connection strength) with the sets of known drugs-diseases and chemical-diseases connections. The tests had significantly p-values (0.001 and 0.002 respectively), indicating that the know connections are distributed in the top of the ranked lists and that most of the strongest connection in the network are known in literature.

4.5).

4.3.6 Relevant cliques

INSIdE nano was scanned in search of cliques of heterogeneous nodes (one nano, one drug, one chemical and one disease). All the cliques identified with a threshold lower that 40% were considered relevant. Moreover, the number of known connections within the cliques increases their relevance. A clique with both disease-drug and disease-chemical known connections is considered more relevant than those in which only one of this information is known. This is because contextualise nanomaterials in these cliques is simpler, as it requires less investigation. Figure 4.3.7 shows how many of the relevant cliques associated to

INSIdE nano (MoA)	Similarity by	Mantel Test
Drugs - Drugs	molecular targets	$1E^{-05}$
Drugs - Drugs	chemical structures	$1E^{-02}$
Drugs - Chemicals	chemical structure	$1E^{-04}$
Drugs - Diseases	use in clinical practice	$1E^{-05}$
Chemicals - Chemicals	chemical structure	$1E^{-05}$
Chemical - Diseases	pathogenic exposures	$1E^{-04}$
Diseases - Diseases	symptoms	$1E^{-05}$

Table 4.5: Comparison of the INSIdE NANO associations based on MoA similarity against independent sets of associations representative of other biochemical aspects. Mantel's test P is reported, under the null hypothesis that the two matrices compared are different (the lower the P, the more similar are the two correlation matrices to each other).

each nanomaterials has both this connection known, or only one of them. It is hi-lighted that the number of cliques with two known connection is lower than the others.

An important aspect in nanomaterials characterisation is the study of the drugs involved in his clique. For each nanomaterial, his relevant clique identified with a lower threshold of 30% is achieved, two known connections were studied. As can be seen in the Figure 4.3.8, nanomaterials based on metal, such as AuNP, TiO2T20 TiO2T7 have a high prevalence of connections with drugs related to the nervous system. Others nanomaterials, such as TiO2NB, PSNP, RS, WC have a high number of connection with drugs affecting the cardiovascular system.

It is also interesting to identify which are the disease that are involved in the relevant cliques related to nanomaterials. As an example, all the relevant cliques involving AuNP were analysed (see figure 4.3.9. It was found to be major associated with circulatory systems disease, most symptoms and signs not well defined and mental and neuro-developmental disorders.

4.3.7 Use-case study

Here and example of analysis in included. Assuming that the user wants to investigate the relationships between Asthma and multi-wallet carbon nanotube

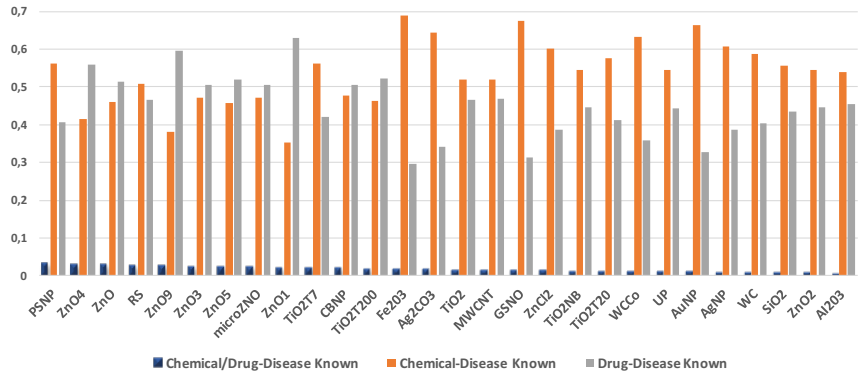


Figure 4.3.7: The bar-plot shows the percentage of relevant cliques associated to each nanomaterials that have both the disease-drug and disease chemical connections already known (blue), that have only the disease-chemical connection known (orange) or that have only the disease-drug connection known (grey). All the cliques were retrieved with a threshold lower that 40%.

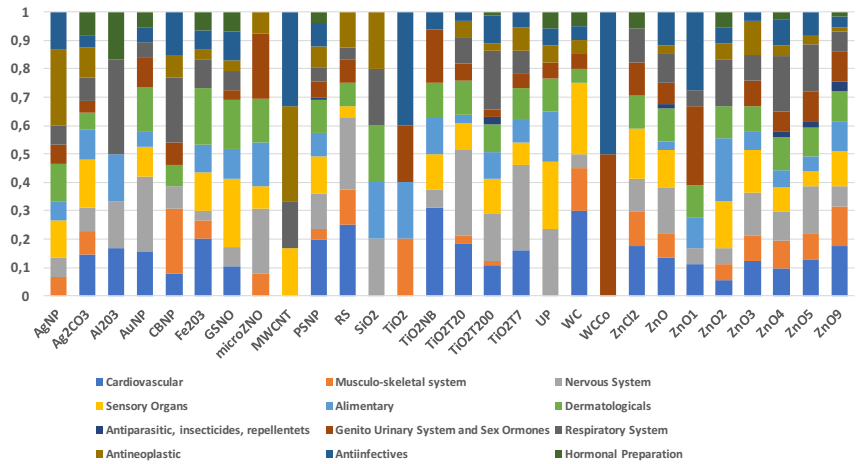


Figure 4.3.8: Drugs involved in relevant cliques. All the cliques were retrieved with a threshold lower that 30% and have two known connections.

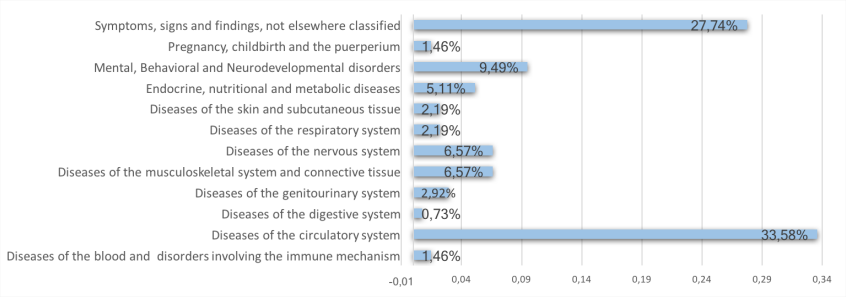


Figure 4.3.9: Diseases involved in relevant AuNP cliques. The cliques considered was retrieved with a threshold lower than 40% and have two known connections.

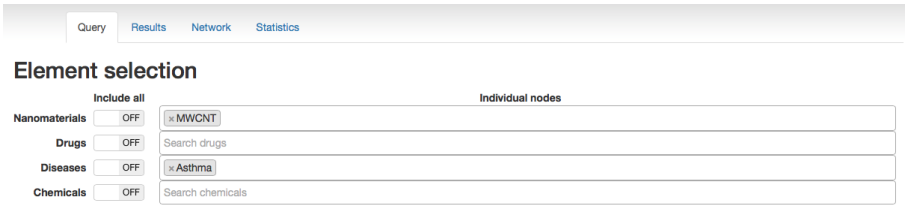


Figure 4.3.10: Conditional Query between MWCNT and Asthma

(MWCNT). Then the use can query the conditional tool by giving MWCNT and Asthma as input. An example of query can be the one depicted in Figure 4.3.10. With these parameter the user set a threshold of 70%, he asked that all the element in the sub-network must be connected both to Asthma and MWCNT, and that both of them must be in the resulting cliques as shown in Figure 4.3.11.

As a results the tool displays a list of cliques, that contains both Asthma and MWCNT (see Figure 4.3.12). The user can filter the list by switch on the 'Filter by known interactions' button. In this way, only the cliques, where at least on between the disease-drug and disease-chemical connections is known. Moreover, the user can start to contextualise the nanomaterial with respect to the other elements by applying filters on the edges. For example in this case, the user asked to have all the cliques where the nanomaterial has different behaviour

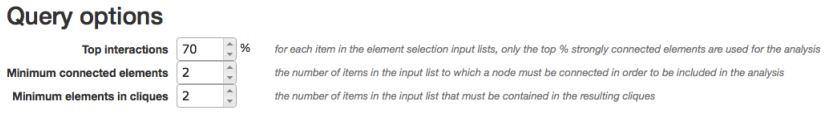
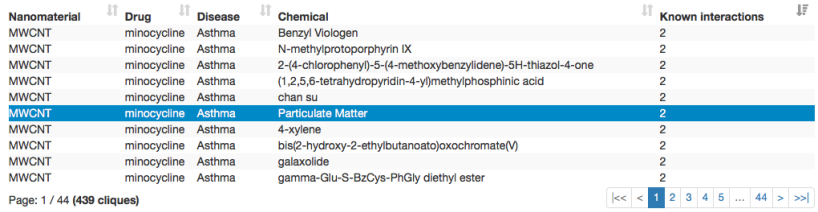


Figure 4.3.11: Conditional Query between MWCNT and Asthma, with a threshold of 70% and 2 minimum connected items, and both Asthma and MWCNT needed to be in the same cliques.



Nanomaterial	Drug	Disease	Chemical	Known interactions
MWCNT	minocycline	Asthma	Benzyl Viologen	2
MWCNT	minocycline	Asthma	N-methylprotoporphyrin IX	2
MWCNT	minocycline	Asthma	2-(4-chlorophenyl)-5-(4-methoxybenzylidene)-5H-thiazol-4-one	2
MWCNT	minocycline	Asthma	(1,2,5,6-tetrahydropyridin-4-yl)methylphosphinic acid	2
MWCNT	minocycline	Asthma	chan su	2
MWCNT	minocycline	Asthma	Particulate Matter	2
MWCNT	minocycline	Asthma	4-xylene	2
MWCNT	minocycline	Asthma	bis(2-hydroxy-2-ethylbutanoato)oxochromate(V)	2
MWCNT	minocycline	Asthma	galaxolide	2
MWCNT	minocycline	Asthma	gamma-Glu-S-BzCys-PhGly diethyl ester	2

Figure 4.3.12: List of cliques resulting from the conditional query between MWCNT and Asthma.

from the drug, and positive behaviour with the disease and the chemical (see Figure 4.3.13). This information can be used to infer knowledge about the fact that the effect on the genes of the nanomaterial and chemical and nanomaterial and disease is the same, meaning that the nano can cause the disease, and that the nano and the drug have different effect on the genes. Moreover, the user asked to select only the cliques where the drug and the disease has negative connection, meaning that the drug can cure the disease. The final output is

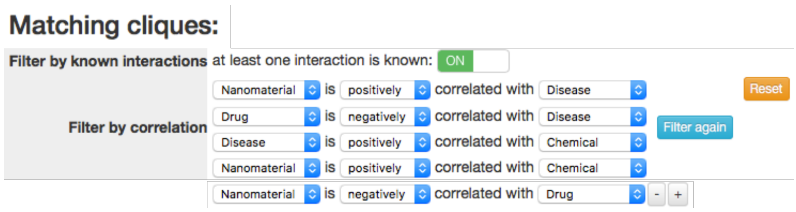


Figure 4.3.13: Filters applied to the results

a list of 87 cliques, 12 of them with 2 known connection. The user then can investigate each clique, by clicking on its row in the table. In this case the user selected the clique composed by MWCNT, minociclyne, Asthma and Particulate Matter. Then he investigated the list of genes underlying the connection. The list of genes confirmed the positivity/negativity of the connection: for example if we consider MWCNT and Particulate matter, their connection is positive, indeed their effect on the gene in the list is the same in 80% of the genes (see Figure 4.3.14).

The user can then, investigate each single element in the clique, by clicking on its name in the 'Clique information' panel. For example by clicking on the name of the drug (minociclyne) the pop-up window reported in Figure 4.3.15 is shown. The available information for drugs are the ATC code, and external link to DrugBank, Wikipedia etc... For example, by clicking on the wikipedia link the user will discover that minociclyne is effectively used to treat Asthma thanks its immune suppressing effects [175].

4.4 Discussion

Many are the possible scenarios of exposure of humans to nanomaterials. For example, when humans breathe millions of natural nanoparticles or byproducts of engine combustion, deposit into the lung. Once in the lung, they are able of overcoming the thin air-blood barrier to transmigrate into the blood [176]. It has also been showed that nanoparticles can reach the brain directly by passing the olfactory epithelium and the nervus olfactorius located in the roof of the nose [177]. Moreover, very small particles ($< 10nm$) are capable of penetrating through to the epidermis or dermis [178]. Another issue is the exposition of workers to nanomaterials during industrial production processes [179]. One example, are the carbon nanotubes, that thanks to their incredible tensile strength are widely used in industries. However, recent studies demonstrate that long thin carbon nanotubes showed the same effects of long thin asbestos fibres [180]. This is because exposition to this nanomaterial can lead to pleural abnormalities such as mesothelioma.

Figure 4.3.14: Results of the conditional query between MWCNT and Asthma.

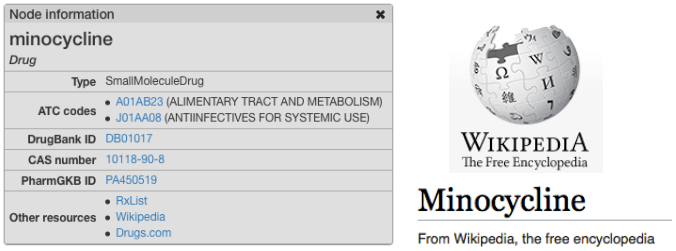


Figure 4.3.15: Minociline investigation

To accurately predict the hazards of these new materials for humans, different biological models are used to determine their potential exposure and toxicity. Figure 4.4.1 elucidates the in vitro-in vivo relationship and its extrapolation to humans. In vitro studies are understood as being very simplified biological models that enable a rapid, low-cost estimation of the effects of xenobiotic substances or nanomaterials. A comparison of different cell types isolated from different tissues or organisms enables the evaluation of more than just the tissue-specific effects. Only animal experiments (in vivo) can provide sufficient answers to the complex issues of absorption, distribution, metabolism, and excretion (ADME). However, the constant improvement of in vitro models to simulate complex multicellular systems [181–183] or entire organs [184] allows an ever more differentiated investigation of possible mechanisms of action and will reduce the need for animal experiments in the long run.

In this scenario, systems biology approaches assume extremely importance due to the computational tool ability to interpret and integrate different information to construct predictive models able to describe the response of the biological system to the nanomaterial perturbation.

Transcriptomics experiments, both performed with microarray or next generation sequencing, are the most likely approaches to survey effects and mechanisms within the toxicological sciences, because they can quantify changes in gene expression through the detection of the number of mRNA copies. Microarrays are very popular in toxicological sciences, in fact, they can be applied to a broad range of in vivo and in vitro models and they are provided with a

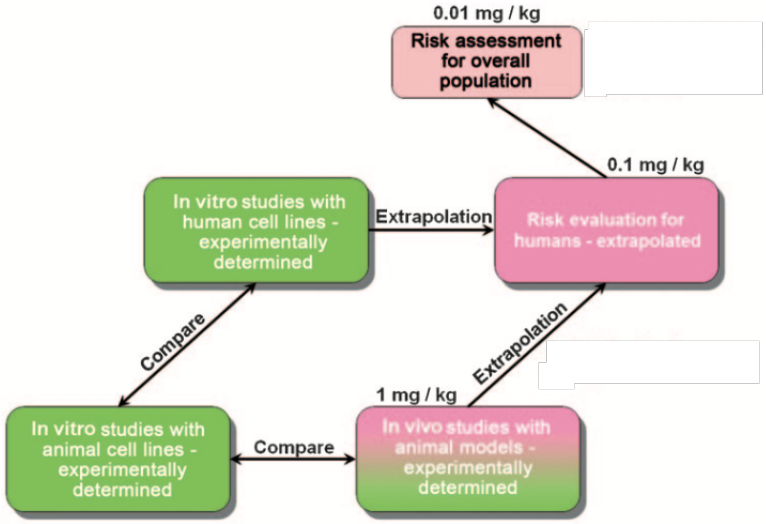


Figure 4.4.1: The evaluation process of toxicity of nanoobjects for humans.

high level of available genomic annotations. As an example, in the frame of the EU-fundend project, FP7-NANOIMMUNE, the NanoMiner transcriptomic database was developed. It comprises a comprehensive set of data based on in vitro studies of nanomaterials [185].

Therefore, INSIdE nano was proposed as an exploratory tool able to compare the nanomaterials exposure mode-of-action with those of drugs treatment, chemical exposure and human disease. The main idea is to infer information regarding the ENMs behaviour by integrating the in-vitro studies coming from the NanoMiner (ENM) and CMAP (drug) databases, and disease-genes and chemicals-genes connections download from the CTD database. The integration process lead to the construction of a big network with phenotypic entities as nodes and their strength of similarity/anti-similarity on the edges. Then the network was scanned in search of cliques of four strongly interconnected nodes (a nano, a drug, a disease and a chemical) that are used to compare the effect of the nano on the genes with the one of the other entities. The analysis of

the network showed that known connections between the phenotypic entitles are retrieved and that this tool can open a new paradigm for future studies on the characterization of nanomaterials effects and mode-of-action. In fact, it is interesting to note that the approach used in nano inside, allows the use of transcriptome data coming for in-vitro experiments, to infer the effect of nanomaterials on human diseases. This is a big advantage as it eliminates the cost and the time required to carry out experiments in vivo.

Chapter 5

Discussion

In the last years, the advancement of high-throughput technologies has led to the production of large amounts of data. Most of them concerns different experiments that characterise the same entity of interest, others that use the same measures to characterize different phenotypic entities.

This aspect significantly increases the importance of data integration in the bioinformatics fields. In fact, many approaches based on systems biology have been designed to exploit such complex and rich data and to integrate them to better characterise complex phenotypes. Nevertheless, nowadays, there is a big gap between the amount of data produced and the knowledge obtained from them. This thesis is an answer to this request, proposing computational methods able to integrate and analyse the data to fill in this gap. In this PhD thesis two new methodology for the integrative analysis of high throughput genomic have been presented, the former called MVDA and the latter called INSIdEnano. MVDA is an integrative tool, based on multi-view clustering techniques, that can identify statistically relevant patients subtype. This problem is usually solved by using transcriptomic data to identify groups of patients that share the same gene patterns. The main idea underlying this research work is that to combine more OMIC data for the same patients to obtain a better characterisation of

their disease profile. This problem is usually solved by using transcriptomic data to identify groups of patients that share the same gene patterns. The main idea underlying this research work is that to combine more OMIC data for the same patients to obtain a better characterisation of their disease profile. The proposed methodology is a late integration approach based on clustering. It works by evaluating the patient clusters in each single view and then combining the clustering results of all the views by factorising the membership matrices. The effectiveness and the performance of my method has been evaluated on six multi-view cancer datasets related to breast cancer, glioblastoma, prostate and ovarian cancer. The omics data used for the experiment are gene and miRNA expression, RNASeq and miRNASeq, Protein Expression and Copy Number Variation. In all the cases patient sub-classes, with statistical significance were found, identifying novel sub-groups previously not emphasised in literature. To obtain higher accuracy, the experiments have been also conducted by using prior information, with respect to patients' classification, as a new view in the integration process. The method outperformed the single view clustering on all the datasets; moreover, it performed better as compared to other multi-view clustering algorithms and, unlike other existing methods, it can quantify the contribution of single views on the results. The method has been also shown to be stable after applying perturbation to the datasets by removing one patient at a time and evaluating the normalised mutual information between all the resulting clusterings. These observations suggest that integration of prior information with genomic features in the sub-typing analysis is an effective strategy in identifying disease subgroups.

Despite that, further experiments should be performed to evaluate the biological differences, at the molecular level, between the sub-classes.

The main drawback of MVDA methodology is its computational complexity. On the other side it allows to have high flexibility in the patient sub-type analysis. In fact, the user can choose the most appropriate clustering algorithm or whether to perform or not the feature selection phase to identify the most significant prototypes. Therefore, it is an effective tool for integrating multi-view homogeneous data that can provide support for the scientific community

INSIdE nano is a tool for the contextualization of nanomaterials mode-of-action. It integrates gene expression data related to: (1) nanomaterials exposure on human cells (2) drugs treatment of human cells (3) known connections between diseases and genes (4) known connections between chemicals and genes. The tool is based on the idea that it is possible to contextualise the molecular effects of nanomaterials perturbations by comparing their patterns of alterations with respect to those of other phenotypic entities (drugs, diseases and chemicals). This tool could greatly increase the knowledge on the ENM molecular effects and in turn contribute to the definition of relevant pathways of toxicity as well as help in predicting the potential involvement of ENM in pathogenetic events or in novel therapeutic strategies. Based on the expression signature associated to each phenotype, the strength of similarity between each pair of perturbations was evaluated and used to build a large network of phenotypes. To ensure the usability of INSIdE nano, a robust and scalable computational infrastructure was developed to scan this large phenotypic network, and a web-based effective graphic user interface was built. Particularly, INSIdE nano was scanned to search clique sub-networks, i.e. quadruplet structures of heterogeneous nodes (a disease, a drug, a chemical and a nanomaterial) completely interconnected by strong patterns of similarity (or anti-similarity). The predictions have been evaluated for a set of known associations between diseases and drugs, based on drug indications in clinical practice, and between diseases and chemicals, based on literature-based causal exposure evidence, and focused on the possible involvement of nanomaterials in the most robust cliques. The evaluation of INSIdE nano confirmed that it highlights known disease-drug and disease-chemical connections. Moreover, disease similarities agree with the information based on their clinical features, as well as drugs and chemicals, mirroring their resemblance based on the chemical structure. Altogether, the results suggest that INSIdE nano can also be successfully used to contextualise the molecular effects of ENMs and infer their connections to other better studied phenotypes, speeding up their safety assessment as well as opening new perspectives concerning their usefulness in biomedicine. From a technical point of view the integrative analysis performed by the two tools are different. The integration performed in

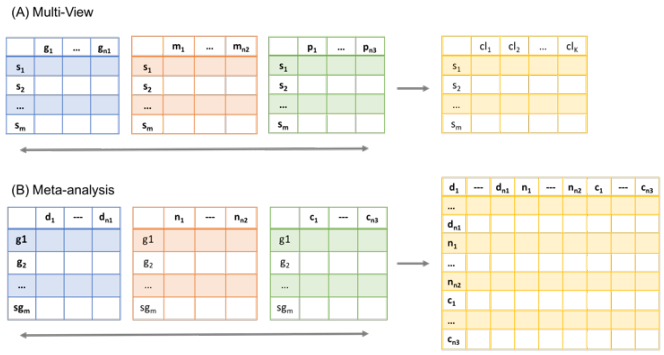


Figure 5.0.1: Difference between the multi-view integration methodology used in MVDA and the meta-analysis methodology used in INSIdE nano. The multi-view methodology (A) integrates different experiments performed on the same samples. The Meta-analysis methodology (B) integrate the results of the same experiment on different samples. In the first case the goal is to find clustering of samples by considering their similarities in all the views. In the second case the goal is to find similarities between differed samples by comparing how they affect the same features.

MVDA can be thought to be horizontal across the views (See Figure 5.0.1 Part A). In fact, the tool requires to have different omics experiments regarding the same patients. The aim of the integration is to evaluate the different measurements performed on the same samples, to identify and enhance the information common across the views, but also highlight the differences that exist between them, which probably will characterize the specific subclasses of patients. On the other hand, the integration strategy proposed in INSIdE nano is a vertical integration across different entities on the same features (See Figure 5.0.1 Part B). In fact, the integration is performed on the same set of features (the genes) evaluated in different experiments related to four specific entities.

Giving the nature of the data, these two integrative strategies can be combined to better characterize phenotypic entities. For example, in the precision medicine field, a useful tool would be one that can identify the patients' subclasses and suggests what the possible factors that cause the disease and the right

drugs treatment and prescription for each of them. Such tool could be implemented by integrating several genomic data to identify the molecular alteration pattern of each disease subtype. These alterations could then be integrated and compared with the molecular changes caused by drugs, chemical substances or environmental factors, to identify which of them can be the underlying causes of the disease and which could, however, cure it.

Chapter 6

Conclusions and future work

Due to the advancement of omics technologies and the development of new systems biology approaches, there is a huge need of methodologies able to integrate and analyse biological data to better cope with clinical, environmental and open research problems. In this PhD thesis, I have proposed two integrative methodologies that solve two relevant biological problems. MVDA is a multi-view methodology that aims to discover new statistically relevant patient sub-classes. Identifying patients' subtypes of a specific diseases is a challenging task especially in the early diagnosis. This is a crucial point for the treatment, because not all the patients affected by the same disease have the same prognosis or need the same drug treatment. INSIdEnano (Integrated Network of Systems bIology Effects of nanomaterials), is a novel tool for the systematic contextualisation of the effects of engineered nanomaterials (ENMs) in the biomedical context. The idea behind the tool is to use analytical strategies to contextualise or position the ENM with respect to relevant phenotypes that are better known, (such as diseases, drug treatments, and other chemical exposures) by comparing their pattern of molecular alteration. The main hypothesis is that suggestive patterns of similarity between sets of phenotypes could be an indication of a biological association to be further tested in toxicological or therapeutic frames.

As the final part of this research a critical analysis of the proposed techniques was carried out by identifying some weaknesses and possible solutions to be developed in the future. Since the integration performed in MVDA is a late integration one, it is not easy to identify the actual contribution of each feature to the results and it lacks of information related to the relationships between the multi-features that characterise a subclass of patients. This is an aspect that could be studied and expanded, for example, by using the canonical correlation analysis. Another solution would be to add a weight to the features to give a confidence level of the feature contribution to results. Moreover, a better biological characterisation of the biological phenotype of each subclass can be identified. This can be done by creating a network of interaction between genes, miRNAs, proteins etc. that characterise the subclasses. This network can be then scanned for searching structural motifs. INSIdE nano has been designed as a building block system of four modules integrated together by using the right similarity measures. This scheme makes INSIdE nano easily expandable with other information blocks such as tissue specific information. This would allow a more precise contextualisation of the nanomaterials and increase the actual level of knowledge that can be inferred. The main limitation of the current state of the system is that to add new phenotypic entities to the system, there is a need of experiments performed on the same genes underlying the actual database. One attractive scenario would be to make INSIdE nano an online learning system, where users can add other phenotypic entities (nodes) or information on the connections (e.g. prior knowledge on the existence/non-existence of a connection), and the system will automatically modify its structure to incorporate these data. For example, the existence of information known a priori on two elements connection is a key shortcoming in the current system, and could be increased after confirming the information entered by users. Moreover, the network in INSIdE nano could be integrated as a multi-level network with epigenetic and metabolic layers. This would allow to perform more complex queries that can contextualise nanomaterials in an even wider scenario. INSIdE-nano can be also used to contextualise new experimental drugs, and infer their behaviour on the cell without performing too many experiments in the wet lab.

To achieve this, the new drug must be integrated to the network by calculating the appropriate measures. After that the network can be scanned in searching cliques involving that drug with a chemical and a disease. Moreover, in the field of data integration in bioinformatics, there is a lack of a complete framework of multi-view learning that includes the various methods proposed so far. Such tools would be of great support for researchers that would like to apply these tools to new kinds of data. Moreover, there is no a general criterion to choose a priori a method among the others. The choice of the methodology mainly depends on the statistical problem, on the type of analysis to be performed, on the type of data to be integrated and on the integration stage. In conclusion, despite the limitations outlined, results showed in the thesis are almost always very encouraging and this suggests that this research area is very promising when working with outmost problems with biomedical, complex and big data.

Appendices

Appendix A

Differentially expressed genes analysis

Let us assume, for simplicity, that for each gene G a set of measurements $a_1^c, \ldots, a_{n_c}^c$ and $a_1^t, \ldots, a_{n_t}^t$ (where n_c and n_t are the number of control and treated samples) representing the expression level are available in both control and treatment situations. Microarray experiments typically aim to identify whether the expression level is significantly different between the biological conditions under examination [186].

One approach commonly used in the literature is based on the analysis of the log fold change of the genes [21, 187–189]. The log fold change (logFC) is defined as follow:

$$logFC = log_2(\overline{a}_c/\overline{a}_t) \tag{A.0.1}$$

where $\overline{a}_c$ and $\overline{a}_t$ are the mean expression values of the gene in the control and in the treatment conditions, respectively. Following this approach, a gene is declared to have a differentially expressed level, if its log fold change is higher than a constant factor, typically 2. Inspection of gene expression data suggests, however, that such a simple "2-fold rule" is unlikely to yield optimal results,

since a factor of 2 can have quite different significance in different regions of the spectrum of expression levels [190]. Moreover, an analysis solely based on fold change however does not allow the assessment of significance of expression differences in the presence of biological and experimental variations, which may differ from gene to gene.

More strict statistical evaluation has been established and the number of methodological papers introducing novel statistical approaches has been increased with the number of biological papers presenting microarray results [191–193]. Usually, in gene-wise analyses, p-values are computed for each gene present on the microarray by using the t-test or some other analytical strategies such as the ANOVA, which helps to estimate the contribution of experimental factors with respect to the distribution of the measured gene expression [194].

Suppose that Y_{jk} is the expression level of gene j in the array $k(j = 1, \ldots, n; k = 1, \ldots, K_1, K_1 + 1, \ldots, K_1 + K_2)$ and that the first K_1 and last K_2 arrays are obtained under the two conditions (i.e., controls and treatments) respectively.

A general statistical model is

$$Y_{jk} = a_j + b_j x_k + \epsilon_{jk} \tag{A.0.2}$$

where $x_k = 1$ for $1 \leq k \leq K_1$ and $x_k = 0$ for $K_1 + 1 \leq k \leq K_1 + K_2$ and ϵ_{jk} ia a random error with mean 0. Hence $a_j + b_j$ and a_j are the mean expression levels of gene j under the two conditions respectively. Determine if a gene has differential expression is equivalent to testing for the null hypothesis $H_0 : b_j = 0$ against $H_1 : b_j \neq 0$.

A statistical test consists of two parts, the former is the construction of a summary statistics, while the latter is to determine the significance level (or p-value) associated to the statistics. Usually the p-value is evaluated based on the null distribution of the test statistics (i.e. under the H_0 hypothesis) which may be specified or estimated via model assumption (e.g. permutation test).

In a t-test, the empirical means $\overline{Y}_{j(1)}$ and $\overline{Y}_{j(2)}$ and variances $s^2_{j(1)}$ and $s^2_{j(2)}$ are used to compute a normalised distance between the two populations (control and treatment) in the form

$$z_j = \frac{(\overline{Y}_{j(1)} - \overline{Y}_{j(2)})}{\sqrt{\frac{s^2_{j(1)}}{K_1} + \frac{s^2_{j(2)}}{K_2}}} \tag{A.0.3}$$

From statistical literature it is known that, under the normality assumption for Y_{jk}, z_j follows a Student distribution with

$$d_j = \frac{(s^2_{j(1)}/K_1 + s^2_{j(2)}/K_2)^2}{(s^k_{j(1)}/K_1)^2/(K_1 - 1) + (s^k_{j(2)}/K_2)^2/(K_2 - 1)} \tag{A.0.4}$$

degree of freedom. When z_j exceeds a certain threshold, depending on the selected confidence level, the two populations are considered to be different. This cut-off is usually based on a multiple testing criterion such as the Bonferroni correction [195] or the false discovery rate [195–197]. Post-hoc corrections are also recommended because the number of tested genes is much greater than the amount of samples replicated across two or more biological conditions.

The fundamental problem with t-test for microarray data, however, is that the repetition numbers K_1 and/or K_2 are often small and can lead to significant underestimates of the variance. Moreover the t-test, such as any other statistical model, makes assumptions on the nature of the data.

The model for the 2-sample t-test with pooled variance states that the samples have different means but the same variance. If both samples are sufficiently large, the Welch's t-test [198] can be used, which allows the samples to have different means and different variances. Another assumption of the t-test is that each sample comes from a population that is close to Normal.

Sometimes, in order to follow these assumptions, the model can be manipulated or the data can be normalised. For example, the t-test is quite insensitive (robust) to non-normality as long as data are not too skewed, but is quite sensitive to skewness. Gene expression data is usually skewed [199], and so, taking logarithms of data tend to make the noise more symmetric and hence closer to normal.

Since the most differentially expressed genes are those with higher log fold change and significant p-value, these two types of information are combined.

The following score can be assigned to each gene:

$$abs(FC) \times -log(Pval) \tag{A.0.5}$$

where FC is the fold change and $Pval$ is the p-value obtained from the statistical test. Since significant p-values are small numbers (lower that 0.05), their negative logarithm will be a high number. Genes with higher score values will be the ones with a higher difference in their expression values and with lower p-values. This score is then used to obtain a ranked list of genes where the most differentially expressed genes are at the top.

Since the sign of the log fold change give information of the fact that the gene are up-regulated (+) or down regulated (-), a list with two relevant tails can be created, having the most up-regulated genes at the top and the most down-regulated genes at the bottom. This is obtained by removing the absolute value of the logFC from the score formula.

Appendix B

Nanomaterials

Between the various substances to which the living organisms are exposed there are nanomaterials. Nanomaterials are defined as those materials with at least one dimension ranging between $1nm$ and $1000nm$ ($10^{-9}m$) [200, 201], see Figure B.0.1.

There are two different kinds of nanomaterials following their origin. The former are the particles that are naturally occurring (such as volcanic ash, soot from forest fires) or byproducts of combustion processes (such as welding, diesel engines) [202, 203]. They normally have heterogeneous physical and chemical properties and are often referred to as ultra-fine particles [204]. The latter

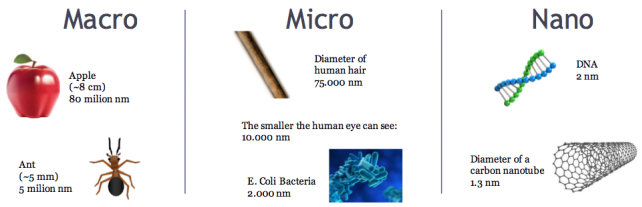


Figure B.0.1: Nanomaterial's Size

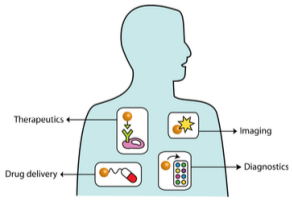


Figure B.0.2: The promise of nanotechnology to improve human health includes diagnostics, drug delivery, imaging, and therapy.Figure from [216]

are the engineered nanomaterials (ENM) that are intentionally produced and designed for various purposes and usually have very specific chemical properties related to shape, size, surface properties [205].

Structural properties affect nanomaterials behaviour more than particle composition itself [206]. Indeed, the ratio between the surface area and the volume is much greater in ENMs than in their conventional bulk forms, enhancing ENMs strength and chemical reactivity [207]. Moreover, quantum effects, at the nano scale level, affect more the nanomaterials properties giving rise to novel optical, electrical and magnetic behaviours. Nanomaterials have already been available for commercial use for several years. Nowadays, they can be found in a wide range of commercial products and everyday items [208, 209].

For example, nanocoatings and nanocomposites are used to produce windows, sports goods, tires, bicycles and automobiles [210]. UV-blocking coatings on glass bottles are used to protect beverages from damage by sunlight [211]. ENMs are also used to make objects last longer such as the butyl-rubber/nanoclay composites used to cover tennis balls [212]. Other examples are the nanoscale titanium dioxide and zinc oxide that find applications in cosmetics and sunscreen [213], and the nanoscale silica that are used as fillers in a range of products, including cosmetics and dental fillings [214]. Moreover, ENMs are even more attractive because they are used in medicine [215, 216] for purposes of diagnosis, imaging and drug delivery. See figure B.0.2 .

Even if nanomaterials are all characterised by extremely small size, they can exist in single, fused, aggregated or agglomerated forms with spherical, tubu-

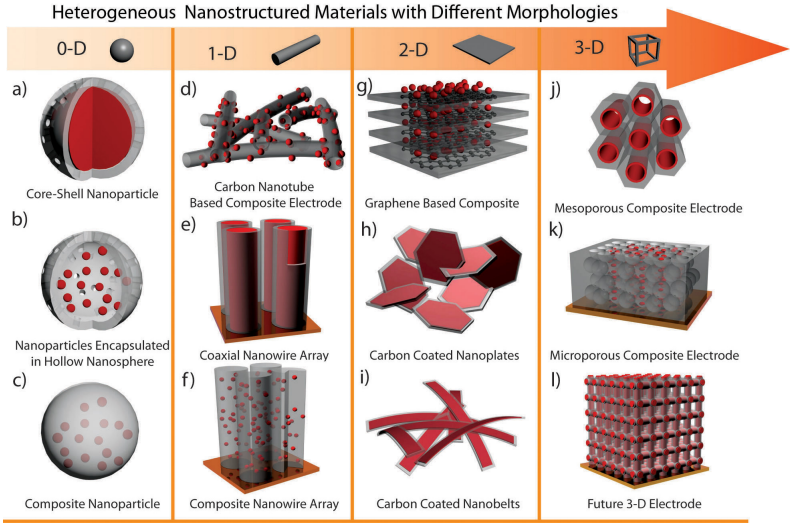


Figure B.0.3: Classification of Nanomaterials (a) 0D spheres and clusters, (b) 1D nanofibers, wires, and rods, (c) 2D films, plates, and networks, (d) 3D nanomaterials.

lar, and irregular shapes. Common types of nanomaterials include nanotubes, dendrimers, quantum dots and fullerenes. Following Siegel [217], they can be grouped based on the number of dimensions in which they are in the nano-scale level: they can be divided into zero dimensional, one dimensional, two dimensional and three dimensional nano materials as shown in figure B.0.3.

The zero dimensional (0-D) nanomaterials have Nano-dimensions in all the three directions [218]. Some examples are metallic nanoparticles such as gold and silver nanoparticles. Most of these nanoparticles have a spherical shape with a diameter in the $1 - 50nm$ range. Moreover, cubes and polygons shapes are also found for this kind of nanomaterials. One dimensional(1-D) nanostructures, have one dimension outside the nanometer range. These include nanowires, nanotubes and nanorods [219]. These materials are several micrometer long, but with a diameter of only a few nanometer. Two dimensional(2-D) nanomaterials have two dimensions outside the nanometer range [220]. These include different kinds of Nano films such as coatings and thin-film-multilayers, nano sheets or nano-walls. The area of the nano films can be large (several square micrometer), but the thickness is always in nano scale range. Three Dimensional(3-D) structures have all dimensions outside the nano meter range. These include bulk materials composed of the individual blocks which are in the nanometer scale $(1 - 100nm$ [221].

On the basis of their structural configuration, nanomaterials can be classified into four types:

Carbon Based Nano materials: The nature of this kind of nanomaterials is hollow spheres, ellipsoids, or tubes. Spherical and ellipsoidal configured carbon nanomaterials are defined as fullerenes, while cylindrical ones are described as nanotubes. Graphite is widely use to engineer various types of carbon-based nanomaterials (CBNs), including single or multi-walled nanotubes, fullerenes, nanodiamonds, and graphene. Carbon Nanotubes (CTNs) are used in a wide range of biomedical applications such as Cell and tissue labeling and imaging, drugs delivery, Reinforcing tissue engineering scaffolds, etc. Despite many successful applications in biomedical engineering, there is a growing concern for safety with CNTs. Some recent in vitro studies have reported increased cytotox-

icity of CNTs due to their cellular uptake, agglomeration, and induced oxidative stress [222].

Metal Based Nano Material: These nanomaterials include nanogold, nanosilver and metal oxides, such as titanium dioxide and closely packed semiconductors like quantum dots [223]. These materials find applications to solve many engineering problems such as to reduce the impact of car exhaust gases on the environment or to create self cleaning windows by the decomposition of dirt [224].

Dendrimers: Highly branched, star-shaped macromolecules with nanometer-scale dimensions. Dendrimers are defined by three components: a central core, an interior dendritic structure (the branches), and an exterior surface with functional surface groups. Applications highlighted in recent literature include drug delivery, gene transfection, catalysis, energy harvesting, photo activity, molecular weight and size determination, rheology modification, and nanoscale science and technology [225, 226].

Composites: Multiphase solid materials where at least one of the phases has one, two or three dimensions in nanoscale. The most common examples of these materials are colloids, gels and copolymers [227]. Nanotechnology has gained a great deal of public interest due to the needs and applications of nanomaterials in many areas of human endeavours, including industry, agriculture, business, medicine and public health. Environmental exposure to nanomaterials is inevitable as nanomaterials become part of our daily life, and as a result, nanotoxicity research is gaining attention [228].

Appendix C

Complex Network Theory

Complex network theory has an important role in a wide range of disciplines, ranging from engineering, social sciences, communications to systems biology [229–233].

For example, in the last decade, the Internet and the World Wide Web (WWW) networks, had a huge increase in size and importance. Network theory was widely applied to the development systematic methods for the analysis and understanding of social networks properties, such as Facebook [234, 235]. Moreover, in ecology and sociology, network theory have been used to perform studies on food-webs [236] and human social networks [237]. These methodologies have been also applied to solve public health problems and to perform epidemiological studies on the spread of diseases[238].

The networks are a suitable tool to model complex entities and their interactions. There are many problems that can be solved using these structures. For example, they allow to infer information about the global structure of the connections (network topology), to identify groups of entities which have homogeneous characteristics (communities), to calculate similarity between entities based on the number of paths that join them together. Most of these problems have been applied to study interactions between biological phenotypes.

Complex biological systems can be represented and analysed as computable networks. There are different kinds of biological networks under study in the field of systems biology, the most common ones are: protein-protein interactions (PPIs) networks, gene regulatory networks, gene co-expression networks.

In PPI networks, proteins are nodes and their interactions are edges [239]. PPIs mainly represent information of how different proteins are coordinate to operate with other proteins to perform the biological processes within the cell [240]. For many of the proteins their complete sequence is already known, but their molecular function need to be fully determined. This prediction can be performed by comparing their interactions with other bio-molecules.

Gene regulatory networks are directed graphs that represent a collection of molecular regulators (DNA, RNA, protein) that interact with each other and with other substances in the cell, such as transcription factors [241], to govern the gene expression levels of mRNA and proteins [242]. They are often studies to identify gene motifs, that are small sets of recurring regulation patterns, that are the basic building blocks of transcription networks [243].

Gene co-expression networks are undirected graphs where the nodes are the genes and there is an edge between a couple of genes if they are significantly co-expressed in the samples [244].

Network models are also used to study interactions between different kinds of phenotypic entities. Many studies related to the interactions between genes and diseases have been performed, analysing complex networks where the diseases and the genes represent the nodes and the edges between them represent their interaction strength [245, 246]. For example, DisGeNET [246] is an network based exploratory platform developed to understand the underlying mechanisms of complex diseases. In fact, many efforts have been made to identify connections between genes and diseases [247, 248], but there are increasing evidences that most diseases arise due to complex interactions among environmental risk factors and multiple genetic variants [249].

Due to the importance that networks has in the biological field, their basic definitions are reported in the following section.

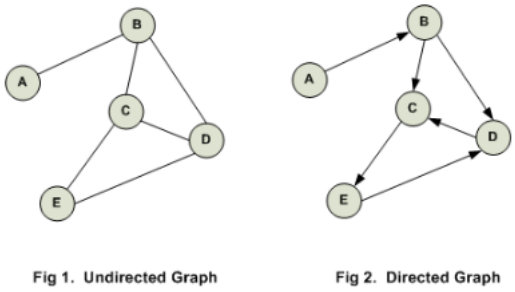


Fig 1. Undirected Graph Fig 2. Directed Graph

Figure C.0.1: Undirected and Directed Graphs representation

Basic Definitions

A graph (or equivalently, a network) is a mathematical abstraction that represents a set of objects, called nodes, and their relationships, called edges. The concept of network is cross disciplinary and it is independent from the kind of objects and relations that represents. Formally, a graph G is defined as the pair $G = (V, E)$, where $V = v_1, \ldots, v_n$ is the finite set of objects representing the nodes of the graph, and $E = e_1, \ldots, e_m$ is the finite set of objects representing the set of edges. Each edge in E is a connection between a pair of nodes (x, y) in V. If there is a relevant sorting order in the pair (x, y) then the graph G will be said to be oriented (or directed) and x will be said to be the source of the edge and y the destination. On the other side, if there is no relevant order, the graph G will be said to be unoriented (or undirected). In terms of information flow into the network, in an oriented graph, the information can transit only from x to y. On the contrary, in an undirected graph, the information can flow in both ways. Moreover nodes and edges can have attributes that identify specific properties of the objects and their interactions represented in the graph. For example, nodes can have labels representing their name, size, colour, etc., while usually edges can have a numeric weight that represents the connection strength between the two end nodes (in this case the graph is said to be weighted). In a visual representation the nodes of a graph are usually denoted as circle and the

edges are denoted as arrows going from the source to the destination. Sometimes undirected edges are represented as lines without arrows. In figure C.0.1 an example of an undirected (a) and a directed graph (b) are shown.

C.0.1 Network properties

Graph density

The graph density shows how sparse or dense a graph is according to the number of connections per node set and is defined as $density = \frac{2|E|}{|V|(|V|-1)}$. A sparse graph is a graph where $|E| = O(|V|^k)$ and $2 > k > 1$. A complete graph is a graph in which every pair of distinct vertices is connected by a unique edge.

Degree Centrality

Given a node i, the degree centrality Cd is defined as the number of edges incident on the node i. Degree Centrality shows that an important node is involved in a large number of interactions. For undirected graph the degree is unique. On the other hand, for directed graphs, each node is characterised by two degree centrality: the "in-degree" counting the number of edges that enter the node, and the "out-degree" counting the number of edges that exit the node. Nodes with very high degree centrality are called hubs since they are connected to many neighbours.

Clustering Coefficient

The clustering coefficient is the measurement that shows the tendency of a graph to be divided into clusters. A cluster is a subset of vertices that contains lots of edges connecting these vertices to each other. Assuming that i is a vertex with degree $deg(i) = k$ in an undirected graph G and that there are e edges between the k neighbours of i in G, then the Local Clustering Coefficient of i in G is given by $C_i = \frac{2e}{k(k-1)}$. Thus, C_i measures the ratio of the number of edges between the neighbours of i to the total possible number of such edges, which is $k(k-1)/2$. It takes values as $0 \leq C_i \leq 1$. The average Clustering Coefficient

of the whole network Coverage is given by $Coverage = \frac{1}{N}\sum_{i=n}^{N}\frac{E_i}{k_i(k_i-1)}$ where $N = |V|$ is the number of vertices. The closer the local clustering coefficient is to 1, the more likely it is for the network to form clusters.

Scale-free Network

Assuming that k is the number of links originating from a given node and $P(k)$ the probability that the degree of a randomly chosen vertex equals k, a scale-free network exhibits a power law distribution $P(k) \sim k^{-\gamma}$ where γ denotes the degree exponent. A scale-free network can be constructed by progressively adding nodes to an existing network and introducing links to existing nodes with preferential attachment so that the probability of linking to a given node i is proportional to the number of existing links k_i that the node has.Thus the connectivity of one node i to any other node j should approximately follow the rule: $P(links\ to\ node\ i) \sim \frac{k_i}{\sum_j k_j}$. The degree distribution $P(k)$ has become one of the most prominent characteristics in network topology. In terms of numerical estimation, a more reliable property, very similar to the previous, is the cumulative degree distribution $P_c(k)$. For a power law distribution $P(k) \sim k^{-\gamma}$ the cumulative degree distribution is of the form $P(k) \sim k^{-\gamma-1}$ and describes the probability of a random chosen node in the network to have a degree greater than k.

Cliques

Given an undirected graph $G = (V, E)$, a clique C is a subset of the vertices, $C \subseteq V$, such that every two distinct vertices are adjacent. The size of a clique comes from the number of vertices it contains. The smallest one contains two vertices with one edge. A maximal clique is a clique that cannot be extended by including one more adjacent vertex, i.e. a maximum clique is a clique of the largest possible size in a given graph. The clique problem refers to the problem of finding the largest clique in any graph G. This problem is NP-complete, and as such, many consider that it is unlikely that an efficient algorithm for finding the largest clique of a graph exists. Figure C.0.2 shows a graph with two

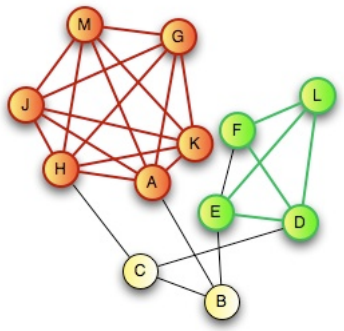


Figure C.0.2: A clique is a sub-network completely connected. In this figure two of the existing cliques are highlighted in the graph. The former is composed of the red nodes while the latter of the green nodes in the completely connected sub-networks.

cliques. Detection and analysis of these structures has found many biological applications: identifying groups of consistently co-expressed genes in microarray datasets, finding cis-regulatory motifs or matching three-dimensional structures of molecules [250, 251].

Adjacency Matrix

Given a graph $G = (V, E)$ with n nodes the adjacency matrix representation consists of a nxn matrix $A = (aij)$ such that $a_{ij} = 1$ if there is an edge that connect the node i and the node j or otherwise $a_{ij} = 0$. If G is weighted graph, the edges can assumes positive or negative values, in that case if there is an edge between the node i and the node j, the a_{ij} would be equal to the weight of the edge w_{ij}. For undirected graphs the matrix is symmetric because $a_{ij} = a_{ji}$. Adjacency matrices require space of $\theta(|V|^2)$ and are best suited for dense and not for sparse graphs.

Chapter 7

Bibliography

1. Marth, J. D. A unified vision of the building blocks of life. *Nature cell biology* **10,** 1015–1015 (2008).
2. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402,** C47–C52 (1999).
3. Gomez-Cabrero, D. *et al.* Data integration in the era of omics: current and future challenges. *BMC systems biology* **8,** 1 (2014).
4. MaâĂŹayan, A. *et al.* Lean Big Data integration in systems biology and systems pharmacology. *Trends in pharmacological sciences* **35,** 450–460 (2014).
5. Mason, O. & Verwoerd, M. Graph theory and networks in biology. *IET systems biology* **1,** 89–119 (2007).
6. Jansen, R., Lan, N., Qian, J. & Gerstein, M. Integration of genomic datasets to predict protein complexes in yeast. *Journal of structural and functional genomics* **2,** 71–81 (2002).
7. Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I. & Noble, W. S. A statistical framework for genomic data fusion. *Bioinformatics* **20,** 2626–2635 (2004).

8. Wasito, I., Istiqlal, A. & Budi, I. *Data integration model for cancer subtype identification using Kernel Dimensionality Reduction-Support Vector Machine (KDR-SVM)* in *Computing and Convergence Technology (ICCCT), 2012 7th International Conference on* (2012), 876–880.

9. Sun, J., Bi, J. & Kranzler, H. R. Multi-view singular value decomposition for disease subtyping and genetic associations. *BMC genetics* **15,** 73 (2014).

10. Patel, S, Parmar, D, Gupta, Y., Singh, M. *et al.* Contribution of genomics, proteomics, and single-nucleotide polymorphism in toxicology research and Indian scenario. *Indian Journal of Human Genetics* **11,** 61 (2005).

11. Brown, D. D. Gene expression in eukaryotes. *Science* **211,** 667–674 (1981).

12. Futcher, B, Latter, G., Monardo, P, McLaughlin, C. & Garrels, J. A sampling of the yeast proteome. *Molecular and cellular biology* **19,** 7357–7368 (1999).

13. Greenbaum, D., Jansen, R. & Gerstein, M. Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics* **18,** 585–596 (2002).

14. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E. M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature biotechnology* **25,** 117–124 (2007).

15. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Molecular and cellular biology* **19,** 1720–1730 (1999).

16. Rogers, S. *et al.* Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics* **24,** 2894–2900 (2008).

17. De Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA expression levels. *Molecular BioSystems* **5,** 1512–1526 (2009).

18. Lockhart, D. J. & Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **405,** 827–836 (2000).

19. Heller, M. J. DNA microarray technology: devices, systems, and applications. *Annual review of biomedical engineering* **4,** 129–153 (2002).

20. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS one* **9,** e78644 (2014).

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

22. Theisen, A. Microarray-based comparative genomic hybridization (aCGH). *Nature Education* **1,** 45 (2008).

23. Buck, M. J. & Lieb, J. D. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83,** 349–360 (2004).

24. Ritchie, M. E. *et al.* A comparison of background correction methods for two-colour microarrays. *Bioinformatics* **23,** 2700–2707 (2007).

25. Smyth, G. K. & Speed, T. Normalization of cDNA microarray data. *Methods* **31,** 265–273 (2003).

26. Park, T. *et al.* Evaluation of normalization methods for microarray data. *BMC bioinformatics* **4,** 1 (2003).

27. Chu, Y. & Corey, D. R. RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic acid therapeutics* **22,** 271–274 (2012).

28. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics* **10,** 57–63 (2009).

29. Hall, N. Advanced sequencing technologies and their wider impact in microbiology. *Journal of Experimental Biology* **210,** 1518–1525 (2007).

30. Church, G. M. Genomes for All. *Scientific American* **294,** 46–54 (2006).

31. Vera, J. C. *et al.* Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular ecology* **17,** 1636–1647 (2008).

32. Scotto-Lavino, E., Du, G. & Frohman, M. A. 5âĂš end cDNA amplification using classic RACE. *Nature protocols* **1,** 2555–2562 (2006).

33. Finotello, F. & Di Camillo, B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Briefings in functional genomics* **14,** 130–142 (2015).

34. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome biology* **17,** 1 (2016).

35. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).

36. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics* **11,** 1 (2010).

37. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-content normalization for RNA-Seq data. *BMC bioinformatics* **12,** 480 (2011).

38. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic acids research,* gks001 (2012).

39. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research* **36,** e105–e105 (2008).

40. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research* **18,** 1509–1517 (2008).

41. Oshlack, A., Robinson, M. D. & Young, M. D. From RNA-seq reads to differential expression results. *Genome biology* **11,** 1 (2010).

42. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5,** 621–628 (2008).

43. Pillai, R. S., Bhattacharyya, S. N. & Filipowicz, W. Repression of protein synthesis by miRNAs: how many mechanisms? *Trends in cell biology* **17,** 118–126 (2007).

44. Nilsen, T. W. Mechanisms of microRNA-mediated gene regulation in animal cells. *TRENDS in Genetics* **23,** 243–249 (2007).

45. Vasudevan, S., Tong, Y. & Steitz, J. A. Switching from repression to activation: microRNAs can up-regulate translation. *Science* **318,** 1931–1934 (2007).

46. Kloosterman, W. P. & Plasterk, R. H. The diverse functions of microRNAs in animal development and disease. *Developmental cell* **11,** 441–450 (2006).

47. Ambros, V. The functions of animal microRNAs. *Nature* **431,** 350–355 (2004).

48. Trang, P, Weidhaas, J. & Slack, F. MicroRNAs as potential cancer therapeutics. *Oncogene* **27,** S52–S57 (2008).

49. Li, C., Feng, Y., Coukos, G. & Zhang, L. Therapeutic microRNA strategies in human cancer. *The AAPS journal* **11,** 747–757 (2009).

50. Fasanaro, P., Greco, S., Ivan, M., Capogrossi, M. C. & Martelli, F. microRNA: emerging therapeutic targets in acute ischemic diseases. *Pharmacology & therapeutics* **125,** 92–104 (2010).

51. Calin, G. A. & Croce, C. M. MicroRNA signatures in human cancers. *Nature Reviews Cancer* **6,** 857–866 (2006).

52. Mráz, M., Malinova, K., Mayer, J. & Pospisilova, S. MicroRNA isolation and stability in stored RNA samples. *Biochemical and biophysical research communications* **390,** 1–4 (2009).

53. Shingara, J. *et al.* An optimized isolation and labeling platform for accurate microRNA expression profiling. *Rna* **11,** 1461–1470 (2005).

54. Buermans, H. P., Ariyurek, Y., van Ommen, G., den Dunnen, J. T. & AC't Hoen, P. New methods for next generation sequencing based microRNA expression profiling. *BMC genomics* **11,** 1 (2010).

55. *The Cancer Genome Atlas* <`https://tcga-data.nci.nih.gov`>.

56. *Gene Expression Omnibus* <`https://www.ncbi.nlm.nih.gov/geo/`>.

57. *Connectivity Map* <`https://portals.broadinstitute.org/cmap/`>.

58. *Nanominer project* <`http://compbio.uta.fi/estools/nanommune/index.php`>.

59. *CTD website* <`http://ctdbase.org`>.

60. *MEDI website* <`https://medschool.vanderbilt.edu/cpm/center-precision-medicine-blog/medi-ensemble-medication-indication-resource`> (2013).

61. *MSIGDB Website* <`http://software.broadinstitute.org/gsea/msigdb`>.

62. Serra, A., Fratello, M., Greco, D. & Tagliaferri, R. *Data integration in genomics and systems biology* in *Evolutionary Computation (CEC), 2016 IEEE Congress on* (2016), 1272–1279.

63. Rhodes, D. R. *et al.* Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 9309–9314 (2004).

64. Choi, J. K., Yu, U., Kim, S. & Yoo, O. J. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **19,** i84–i90 (2003).

65. Luo, J *et al.* A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *The pharmacogenomics journal* **10,** 278–291 (2010).

66. Aggarwal, C. C., Hinneburg, A. & Keim, D. A. *On the surprising behavior of distance metrics in high dimensional space* (Springer, 2001).

67. Pavlidis, P., Weston, J., Cai, J. & Grundy, W. N. *Gene functional classification from heterogeneous data* in *Proceedings of the fifth annual international conference on Computational biology* (2001), 249–255.

68. Serra, A. *et al.* MVDA: a multi-view genomic data integration methodology. *BMC bioinformatics* **16,** 261 (2015).

69. Saria, S. & Goldenberg, A. Subtyping: What it is and its role in precision medicine. *IEEE Intelligent Systems* **30,** 70–75 (2015).

70. Hood, L. & Friend, S. H. Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nature Reviews Clinical Oncology* **8,** 184–187 (2011).

71. Mirnezami, R., Nicholson, J. & Darzi, A. Preparing for precision medicine. *New England Journal of Medicine* **366,** 489–491 (2012).

72. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences* **101,** 4164–4169 (2004).

73. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).

74. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* **98,** 10869–10874 (2001).

75. Vang Nielsen, K. *et al.* The value of TOP2A gene copy number variation as a biomarker in breast cancer: Update of DBCG trial 89D. *Acta Oncologica* **47,** 725–734 (2008).

76. Planey, C. R. & Gevaert, O. CoINcIDE: A framework for discovery of patient subtypes across multiple datasets. *Genome medicine* **8,** 1 (2016).

77. Taskesen, E. *et al.* Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics. *Scientific reports* **6** (2016).

78. Higdon, R. *et al.* The promise of multi-omics and clinical data integration to identify and target personalized healthcare approaches in autism spectrum disorders. *Omics: a journal of integrative biology* **19,** 197–208 (2015).

79. Liu, G., Dong, C. & Liu, L. Integrated Multiple âĂIJ-omicsâĂİ Data Reveal Subtypes of Hepatocellular Carcinoma. *PloS one* **11,** e0165457 (2016).

80. Jiang, D., Tang, C. & Zhang, A. Cluster analysis for gene expression data: a survey. *IEEE Transactions on knowledge and data engineering* **16,** 1370–1386 (2004).

81. Sotiriou, C. & Piccart, M. J. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nature Reviews Cancer* **7,** 545–553 (2007).

82. Theodoridis, S. & Koutroumbas, K. *Pattern Recognition* ISBN: 9781597492720 (Academic Press, 2008).

83. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics Oxford England* **22,** 1540–2. ISSN: 1367-4803 (June 2006).

84. Hartigan, J. a. & Wong, M. a. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society* **28,** 100–108. ISSN: 0035-9254 (1979).

85. Kaufman, L & Rousseeuw, P. J. Clustering by means of medoids. *Statistical Data Analysis Based on the L 1-Norm and Related Methods. First International Conference,* 405–416416 (1987).

86. Vesanto, J. & Alhoniemi, E. Clustering of the self-organizing map. *Neural Networks, IEEE Transactions on* **11,** 586–600 (2000).

87. Ng, A. Y., Jordan, M. I., Weiss, Y. *et al.* On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* **2,** 849–856 (2002).

88. Wei, Y.-C. & Cheng, C.-K. *Towards efficient hierarchical designs by ratio cut partitioning* in *Computer-Aided Design, 1989. ICCAD-89. Digest of Technical Papers., 1989 IEEE International Conference on* (1989), 298–301.

89. Yona, G., Dirks, W. & Rahman, S. in *Methods in Molecular Biology* 479–509 (Springer Nature, 2009). doi:10.1007/978-1-59745-243-4_21. <https://doi.org/10.1007%2F978-1-59745-243-4_21>.

90. De Sa, V. R. *Spectral clustering with two views* in *ICML workshop on learning with multiple views* (2005).

91. Chen, X., Xu, X., Huang, J. Z. & Ye, Y. TW-(k)-Means: Automated Two-Level Variable Weighting Clustering Algorithm for Multiview Data. *Knowledge and Data Engineering, IEEE Transactions on* **25,** 932–944 (2013).

92. S. Bickel and T. Scheffer. Multi-View Clustering. **Proc. IEEE,** 19–26 (204).

93. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nature methods* **11,** 333–337 (2014).

94. Long, B., Philip, S. Y. & Zhang, Z. M. *A General Model for Multiple View Unsupervised Learning.* in *SDM* (2008), 822–833.

95. Greene, D. A Matrix Factorization Approach for Integrating Multiple Data Views. *Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science* **5781** (eds Buntine, W., Grobelnik, M., Mladenić, D. & Shawe-Taylor, J.) 423–438 (2009).

96. Shen, R. *et al.* Integrative subtype discovery in glioblastoma using iCluster. *PLoS One* **7,** e35236 (2012).

97. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401,** 788–791 (1999).

98. Yang, Z. & Michailidis, G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* **32,** 1–8 (2016).

99. Zhang, S. *et al.* Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic acids research,* gks725 (2012).

100. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **58,** 236–244. ISSN: 0162-1459 (Mar. 1963).

101. Handl, J., Knowles, J. & Kell, D. B. Computational cluster validation in post-genomic data analysis. *Bioinformatics (Oxford, England)* **21,** 3201–12. ISSN: 1367-4803 (Aug. 2005).

102. Nieweglowski, L. *clv: Cluster Validation Techniques* R package version 0.3-2.1 (2013). <`http://CRAN.R-project.org/package=clv`>.

103. Kovács, F., Legány, C. & Babos, A. *Cluster validity measurement techniques* in *6th International symposium of hungarian researchers on computational intelligence* (2005).

104. Ahdesmäki, M. & Strimmer, K. Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. *The Annals of Applied Statistics* **4,** 503–519. ISSN: 1932-6157 (Mar. 2010).

105. Breiman, L. Random forests. *Machine learning* **45,** 5–32 (2001).

106. Fisher, R. A. *JSTOR: Journal of the Royal Statistical Society, Vol. 85, No. 1 (Jan., 1922), pp. 87-94* 1922.

107. Lin, S. Space oriented rank-based data integration. *Statistical Applications in Genetics and Molecular Biology* **9** (2010).

108. *Memoral Sloan-Kettering Cancer Center* <`http://cbio.mskcc.org/`>.

109. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. *sva: Surrogate Variable Analysis* R package version 3.10.0 ().

110. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 6567–72. ISSN: 0027-8424 (May 2002).

111. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **27,** 1160–7. ISSN: 1527-7755 (Mar. 2009).

112. Buffa, F. M. *et al.* microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. *Cancer research* **71,** 5635–45. ISSN: 1538-7445 (Sept. 2011).

113. Verhaak, R. G. W. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer cell* **17,** 98–110. ISSN: 1878-3686 (Jan. 2010).

114. <`http://cbio.mskcc.org/cancergenomics/prostate/data/`>.

115. Ray, B. *et al.* Information content and analysis methods for multi-modal high-throughput biomedical data. *Scientific reports* **4,** 4411. ISSN: 2045-2322 (Jan. 2014).

116. Chang, H. Y. *et al.* Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 3738–3743 (2005).

117. West, M. *et al.* Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences* **98,** 11462–11467 (2001).

118. Huang, E. *et al.* Gene expression predictors of breast cancer outcomes. *The Lancet* **361,** 1590–1596 (2003).

119. Aitken, R., Chaudhry, M., Boxall, A. & Hull, M. Manufacture and use of nanomaterials: current status in the UK and global trends. *Occupational medicine* **56,** 300–306 (2006).

120. De Volder, M. F., Tawfick, S. H., Baughman, R. H. & Hart, A. J. Carbon nanotubes: present and future commercial applications. *science* **339,** 535–539 (2013).

121. Oberdörster, G. *et al.* Principles for characterizing the potential human health effects from exposure to nanomaterials: elements of a screening strategy. *Particle and fibre toxicology* **2,** 1 (2005).

122. Pirela, S. V. *et al.* Consumer exposures to laser printer-emitted engineered nanoparticles: a case study of life-cycle implications from nano-enabled products. *Nanotoxicology* **9,** 760–768 (2015).

123. Pirela, S. V. *et al.* Effects of Laser Printer–Emitted Engineered Nanoparticles on Cytotoxicity, Chemokine Expression, Reactive Oxygen Species, DNA Methylation, and DNA Damage: A Comprehensive in Vitro Analysis in Human Small Airway Epithelial Cells, Macrophages, and Lymphoblasts. *Environmental health perspectives* **124,** 210 (2016).

124. Raj, S., Jose, S., Sumod, U., Sabitha, M *et al.* Nanotechnology in cosmetics: Opportunities and challenges. *Journal of Pharmacy and Bioallied Sciences* **4,** 186 (2012).

125. Hubbell, J. A. & Chilkoti, A. Nanomaterials for drug delivery. *Science* **337,** 303–305 (2012).

126. Zamani, M., Prabhakaran, M. P. & Ramakrishna, S. Advances in drug delivery via electrospun and electrosprayed nanomaterials. *Int J Nanomedicine* **8,** 2997–3017 (2013).

127. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *science* **313,** 1929–1935 (2006).

128. Napolitano, F., Sirci, F., Carrella, D. & di Bernardo, D. Drug-set enrichment analysis: a novel tool to investigate drug mode of action. *Bioinformatics* **32,** 235–241 (2016).

129. Napolitano, F. *et al.* Drug repositioning: a machine-learning approach through data integration. *Journal of cheminformatics* **5,** 1 (2013).

130. Iorio, F. *et al.* Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences* **107,** 14621–14626 (2010).

131. Dudley, J. T. *et al.* Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Science translational medicine* **3,** 96ra76–96ra76 (2011).

132. Zhou, X., Menche, J., Barabási, A.-L. & Sharma, A. Human symptoms–disease network. *Nature communications* **5** (2014).

133. Pelletier, D *et al.* ToxEvaluator: an integrated computational platform to aid the interpretation of toxicology study-related findings. *Database* **2016,** baw062 (2016).

134. Kong, L. *et al.* NanoMiner – integrative human transcriptomics data resource for nanoparticle research. *PloS one* **8,** e68414 (2013).

135. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research* **33,** D54–D58 (2005).

136. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic acids research* **39,** D38–D51 (2011).

137. Fujita, K. *et al.* Effects of ultrafine TiO 2 particles on gene expression profile in human keratinocytes without illumination: involvement of extracellular matrix and cell adhesion. *Toxicology letters* **191,** 109–117 (2009).

138. Tilton, S. C. *et al.* Three human cell types respond to multi-walled carbon nanotubes and titanium dioxide nanobelts with cell-specific transcriptomic and proteomic expression patterns. *Nanotoxicology* **8,** 533–548 (2014).

139. Moos, P. J. *et al.* Responses of human cells to ZnO nanoparticles: a gene transcription study. *Metallomics* **3,** 1199–1211 (2011).

140. Hussien, R. *et al.* Unique growth pattern of human mammary epithelial cells induced by polymeric nanoparticles. *Physiological reports* **1** (2013).

141. Kim, E. *et al.* Gold nanoparticle-mediated gene delivery induces widespread changes in the expression of innate immunity genes. *Gene therapy* **19,** 347–353 (2012).

142. Tuomela, S. *et al.* Gene expression profiling of immune-competent human cells exposed to engineered zinc oxide or titanium dioxide nanoparticles. *PLoS One* **8,** e68415 (2013).

143. Ronzani, C., Safar, R., Le Faou, A., Rihn, B. H. & Joubert, O. Comment on: S-nitrosoglutathione (GSNO) is cytotoxic to intracellular amastigotes and promotes healing of topically treated Leishmania major or Leishmania braziliensis skin lesions. *Journal of Antimicrobial Chemotherapy,* dku122 (2014).

144. Karoly, E. D., Li, Z., Dailey, L. A., Hyseni, X. & Huang, Y.-C. T. Up-regulation of tissue factor in human pulmonary artery endothelial cells after ultrafine particle exposure. *Environmental health perspectives,* 535–540 (2007).

145. Kawata, K., Osawa, M. & Okabe, S. In vitro toxicity of silver nanoparticles at noncytotoxic doses to HepG2 human hepatoma cells. *Environmental science & technology* **43,** 6046–6051 (2009).

146. Busch, W., Kühnel, D., Schirmer, K. & Scholz, S. Tungsten carbide cobalt nanoparticles exert hypoxia-like effects on the gene expression level in human keratinocytes. *BMC genomics* **11,** 1 (2010).

147. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)* **8,** 118–27. ISSN: 1465-4644 (Jan. 2007).

148. Leek, J. T. & Storey, J. D. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences* **105,** 18718–18723 (2008).

149. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28,** 882–883 (2012).

150. Davis, A. P. *et al.* The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic acids research* **43,** D914–D920 (2015).

151. Lamb, J. *et al.* A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell* **114,** 323–334 (2003).

152. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics* **34,** 267–273 (2003).

153. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 15545–15550 (2005).

154. Fagin, R., Kumar, R. & Sivakumar, D. Comparing top k lists. *SIAM Journal on Discrete Mathematics* **17,** 134–160 (2003).

155. DeConde, R. P. *et al.* Combining results of microarray experiments: a rank aggregation approach. *Statistical Applications in Genetics and Molecular Biology* **5** (2006).

156. Tembe, W. D. *et al.* Statistical comparison framework and visualization scheme for ranking-based algorithms in high-throughput genome-wide studies. *Journal of Computational Biology* **16,** 565–577 (2009).

157. Vorontsov, I. E., Kulakovskiy, I. V. & Makeev, V. J. Jaccard index based similarity measure to compare transcription factor binding site models. *Algorithms for Molecular Biology* **8,** 1 (2013).

158. Bass, J. I. F. *et al.* Using networks to measure similarity between genes: association index selection. *nature methods* **10,** 1169–1176 (2013).

159. Kolmogorov, A. N. *Sulla determinazione empirica di una leggi di distribuzione; Translated in English in Breakthroughs in Statistics, by Kotz and Johnson* 1933.

160. Smirnov, N. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics,* 279–281 (1948).

161. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer research* **27,** 209–220 (1967).

162. van der Loo, M. The stringdist package for approximate string matching. *The R Journal* **6,** 111–122 (1 2014).

163. Wei, W.-Q. *et al.* Development and evaluation of an ensemble resource linking medications to their indications. *Journal of the American Medical Informatics Association* **20,** 954–961 (2013).

164. Wei, W.-Q., Mosley, J. D., Bastarache, L. & Denny, J. C. *Validation and enhancement of a computable medication indication resource (MEDI) using a large practice-based dataset* in *AMIA Annual Symposium Proceedings* **2013** (2013), 1448.

165. Wheeler, D. L. *et al.* Database resources of the national center for biotechnology information. *Nucleic acids research* **35,** D5–D12 (2007).

166. Lowe, H. J. & Barnett, G. O. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Jama* **271,** 1103–1108 (1994).

167. Cormen, T. H. *Introduction to algorithms* (MIT press, 2009).

168. Gosink, M. ToxReporter: viewing the genome through the eyes of a toxicologist. *Database* **2016,** baw141 (2016).

169. Goldstein, D. M., Gray, N. S. & Zarrinkar, P. P. High-throughput kinase profiling as a platform for drug discovery. *Nature reviews Drug discovery* **7,** 391–397 (2008).

170. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **40,** D1100–D1107 (2012).

171. Willett, P. Searching techniques for databases of two-and three-dimensional chemical structures. *Journal of medicinal chemistry* **48,** 4183–4199 (2005).

172. Xue, L. & Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Combinatorial Chemistry & High Throughput Screening* **3,** 363–372 (2000).

173. Pavlopoulos, G. A. *et al.* Using graph theory to analyze biological networks. *BioData mining* **4,** 1 (2011).

174. Ronning, M. A historical overview of the ATC/DDD methodology. *WHO drug information* **16,** 233 (2002).

175. Joks, R. & Durkin, H. G. Non-antibiotic properties of tetracyclines as anti-allergy and asthma drugs. *Pharmacological Research* **64,** 602–609 (2011).

176. Geiser, M. *et al.* Ultrafine particles cross cellular membranes by nonphagocytic mechanisms in lungs and in cultured cells. *Environmental health perspectives,* 1555–1560 (2005).

177. Oberdörster, G. *et al.* Translocation of inhaled ultrafine particles to the brain. *Inhalation toxicology* **16,** 437–445 (2004).

178. Ryman-Rasmussen, J. P., Riviere, J. E. & Monteiro-Riviere, N. A. Penetration of intact skin by quantum dots with diverse physicochemical properties. *Toxicological Sciences* **91,** 159–165 (2006).

179. Ding, Y. *et al.* Airborne engineered nanomaterials in the workplace - a review of release and worker exposure during nanomaterial production and handling processes. *Journal of hazardous materials* (2016).

180. Poland, C. A. *et al.* Carbon nanotubes introduced into the abdominal cavity of mice show asbestos-like pathogenicity in a pilot study. *Nature nanotechnology* **3,** 423–428 (2008).

181. Wottrich, R., Diabaté, S. & Krug, H. F. Biological effects of ultrafine model particles in human macrophages and epithelial cells in mono-and co-culture. *International journal of hygiene and environmental health* **207,** 353–361 (2004).

182. Rothen-Rutishauser, B., Blank, F., Mühlfeld, C. & Gehr, P. In vitro models of the human epithelial airway barrier to study the toxic potential of particulate matter. *Expert opinion on drug metabolism & toxicology* **4,** 1075–1089 (2008).

183. Mallampati, R. *et al.* Evaluation of EpiDerm full thickness-300 (EFT-300) as an in vitro model for skin irritation: studies on aliphatic hydrocarbons. *Toxicology in Vitro* **24,** 669–676 (2010).

184. Wick, P. *et al.* Barrier capacity of human placenta for nanosized materials. *Environmental health perspectives (Online)* **118,** 432 (2010).

185. Kong, J. *et al.* Integrative, multimodal analysis of glioblastoma using TCGA molecular data, pathology images, and clinical outcomes. *IEEE transactions on bio-medical engineering* **58,** 3469–74. ISSN: 1558-2531 (Dec. 2011).

186. Lee, M.-L. T., Kuo, F. C., Whitmore, G. & Sklar, J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences* **97,** 9834–9839 (2000).

187. Baldi, P. & Brunak, S. *Bioinformatics: The Machine Learning Approach (Adaptive Computation and Machine Learning series)* ISBN: 978-0262025065 (A Bradford Book, 2001).

188. Lockhart, D. J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology* **14,** 1675–1680 (1996).

189. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102,** 109–126 (2000).

190. Mutch, D. M., Berger, A., Mansourian, R., Rytz, A. & Roberts, M.-A. The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data. *BMC bioinformatics* **3,** 1 (2002).

191. Jain, N. *et al.* Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* **19,** 1945–1951 (2003).

192. McCarthy, D. J. & Smyth, G. K. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* **25,** 765–771 (2009).

193. Dembélé, D. & Kastner, P. Fold change rank ordering statistics: a new method for detecting differentially expressed genes. *BMC bioinformatics* **15,** 1 (2014).

194. Cui, X., Hwang, J. G., Qiu, J., Blades, N. J. & Churchill, G. A. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6,** 59–75 (2005).

195. Shaffer, J. P. Multiple hypothesis testing. *Annual review of psychology* **46,** 561 (1995).

196. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological),* 289–300 (1995).

197. Reiner, A., Yekutieli, D. & Benjamini, Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19,** 368–375 (2003).

198. Welch, B. L. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika* **34,** 28–35 (1947).

199. Hardin, J. & Wilson, J. A note on oligonucleotide expression values not being normally distributed. *Biostatistics,* kxp003 (2009).

200. Hochella, M. F. ThereâĂŹs plenty of room at the bottom: Nanoscience in geochemistry. *Geochimica et Cosmochimica Acta* **66,** 735–743 (2002).

201. Buzea, C., Pacheco, I. I. & Robbie, K. Nanomaterials and nanoparticles: Sources and toxicity. *Biointerphases* **2,** MR17 (2007).

202. Sharma, V. K., Filip, J., Zboril, R. & Varma, R. S. Natural inorganic nanoparticles–formation, fate, and toxicity in the environment. *Chemical Society Reviews* **44,** 8410–8423 (2015).

203. Rogers, M. A. Naturally occurring nanoparticles in food. *Current Opinion in Food Science* **7,** 14–19 (2016).

204. Hartland, A., Lead, J. R., Slaveykova, V., O'Carroll, D. & Valsami-Jones, E. The environmental significance of natural nanoparticles. *Nature Education Knowledge* **4,** 7 (2013).

205. Xing, B., Vecitis, C. D. & Senesi, N. *Engineered Nanoparticles and the Environment: Biophysicochemical Processes and Toxicity* doi:10.1002/9781119275855. <http://dx.doi.org/10.1002/9781119275855> (Wiley-Blackwell, Sept. 2016).

206. Lead, J. R. & Wilkinson, K. J. Aquatic colloids and nanoparticles: current knowledge and future trends. *Environmental Chemistry* **3,** 159–171 (2006).

207. Madden, A. S., Hochella, M. F. & Luxton, T. P. Insights for size-dependent reactivity of hematite nanomineral surfaces through Cu 2+ sorption. *Geochimica et Cosmochimica Acta* **70,** 4095–4104 (2006).

208. Kessler, R. Engineered Nanoparticles in Consumer Products: Understanding a New Ingredient. *Environmental Health Perspectives* **119,** a120–a125 (Mar. 2011).

209. Zhang, Y., Leu, Y.-R., Aitken, R. & Riediker, M. Inventory of Engineered Nanoparticle-Containing Consumer Products Available in the Singapore Retail Market and Likelihood of Release into the Aquatic Environment. *International Journal of Environmental Research and Public Health* **12,** 8717–8743 (July 2015).

210. *Nanocoatings and Ultra-Thin Films: Technologies and Applications (Woodhead Publishing Series in Metals and Surface Engineering)* (Woodhead Publishing, 2011).

211. Narasimha, M & Lall, G. Influence of nano materials and their coating in manufacturing industries. *International Journal Science Technology and Management* **5** (Jan. 2016).

212. Vijaykumar B. Sutariya, Y. P. *Biointeractions of Nanomaterials* ISBN: 9781466582385 (CRC Press, 2014).

213. Lu, P.-J., Huang, S.-C., Chen, Y.-P., Chiueh, L.-C. & Shih, D. Y.-C. Analysis of titanium dioxide and zinc oxide nanoparticles in cosmetics. *journal of food and drug analysis* **23,** 587–594 (2015).

214. Yang, J., Han, C.-R., Duan, J.-F., Xu, F. & Sun, R.-C. In situ grafting silica nanoparticles reinforced nanocomposite hydrogels. *Nanoscale* **5,** 10858–10863 (2013).

215. Liang, X.-J., Chen, C., Zhao, Y., Jia, L. & Wang, P. C. Biopharmaceutics and therapeutic potential of engineered nanomaterials. *Current drug metabolism* **9,** 697–709 (2008).

216. Murphy, C. J. *et al.* Biological responses to engineered nanomaterials: Needs for the next decade. *ACS central science* **1,** 117–123 (2015).

217. Siegel, R. W. *Nanophase Materials: Synthesis - Properties - Applications* ISBN: 0792327543 ((NATO Science Series E: Applied Sciences, 1994).

218. Cao, G. & Wang, Y. *Nanostructures and Nanomaterials: Synthesis, Properties, and Applications: Synthesis, Properties, and Applications (2nd Edition) (World Scientific Series in Nanoscience and Nanotechnology)* ISBN: 9814322504 (World Scientific Publishing Company, 2011).

219. Hu, J., Odom, T. W. & Lieber, C. M. Chemistry and physics in one dimension: synthesis and properties of nanowires and nanotubes. *Accounts of chemical research* **32,** 435–445 (1999).

220. Huang, X., Tan, C., Yin, Z. & Zhang, H. 25th Anniversary Article: Hybrid Nanostructures Based on Two-Dimensional Nanomaterials. *Advanced Materials* **26,** 2185–2204 (2014).

221. Wen, D. *et al.* Gold Aerogels: Three-Dimensional Assembly of Nanoparticles and Their Use as Electrocatalytic Interfaces. *ACS nano* **10,** 2559–2567 (2016).

222. Cha, C., Shin, S. R., Annabi, N., Dokmeci, M. R. & Khademhosseini, A. Carbon-based nanomaterials: multifunctional materials for biomedical engineering. *ACS nano* **7,** 2891–2897 (2013).

223. Mody, V., Siwale, R., Singh, A. & Mody, H. Introduction to metallic nanoparticles. *Journal of Pharmacy and Bioallied Sciences* **2,** 282 (2010).

224. Parkin, I. P. & Palgrave, R. G. Self-cleaning coatings. *Journal of Materials Chemistry* **15,** 1689–1695 (2005).

225. Abbasi, E. *et al.* Dendrimers: synthesis, applications, and properties. *Nanoscale research letters* **9,** 1–10 (2014).

226. Fréchet, J. M. Dendrimers and other dendritic macromolecules: From building blocks to functional assemblies in nanoscience and nanotechnology. *Journal of Polymer Science Part A: Polymer Chemistry* **41,** 3713–3725 (2003).

227. Ajayan, P. M., Schadler, L. S. & Braun, P. V. *Nanocomposite science and technology* (John Wiley & Sons, 2006).

228. Ray, P. C., Yu, H. & Fu, P. P. Toxicity and environmental risks of nanomaterials: challenges and future needs. *Journal of Environmental Science and Health Part C* **27,** 1–35 (2009).

229. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Reviews of modern physics* **74,** 47 (2002).

230. Barabasi, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature reviews genetics* **5,** 101–113 (2004).

231. Reka, A., Jeong, H. & Barabasi, A.-L. Diameter of the world-wide web. *Nature* **401,** 130–131 (1999).

232. Mitra, K., Carvunis, A.-R., Ramesh, S. K. & Ideker, T. Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics* **14,** 719–732 (2013).

233. Dorogovtsev, S. N. & Mendes, J. F. *Evolution of networks: From biological nets to the Internet and WWW* (OUP Oxford, 2013).

234. Ugander, J., Karrer, B., Backstrom, L. & Marlow, C. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503* (2011).

235. Backstrom, L. & Kleinberg, J. *Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on facebook* in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (2014), 831–841.

236. Polis, G. A. & Winemiller, K. O. *Food webs: integration of patterns & dynamics* (Springer Science & Business Media, 2013).

237. Burt, R. S., Kilduff, M. & Tasselli, S. Social network analysis: Foundations and frontiers on advantage. *Annual review of psychology* **64,** 527–547 (2013).

238. Keeling, M. J. & Eames, K. T. Networks and epidemic models. *Journal of the Royal Society Interface* **2,** 295–307 (2005).

239. De Las Rivas, J. & Fontanillo, C. Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol* **6,** e1000807 (2010).

240. Pellegrini, M., Haynor, D. & Johnson, J. M. Protein interaction networks. *Expert review of proteomics* **1,** 239–249 (2004).

241. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309,** 1559–1563 (2005).

242. Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E. & Guthke, R. Gene regulatory network inference: data integration in dynamic modelsâĂŤa review. *Biosystems* **96,** 86–103 (2009).

243. Alon, U. Network motifs: theory and experimental approaches. *Nature Reviews Genetics* **8,** 450–461 (2007).

244. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *science* **302,** 249–255 (2003).

245. Zickenrott, S, Angarica, V., Upadhyaya, B. & Del Sol, A. Prediction of disease–gene–drug relationships following a differential network analysis. *Cell death & disease* **7,** e2040 (2016).

246. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research,* gkw943 (2016).

247. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature genetics* **33,** 228–237 (2003).

248. Kann, M. G. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. *Briefings in bioinformatics* **11,** 96–110 (2010).

249. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6,** 95–108 (2005).

250. Zhang, H., Song, X., Wang, H. & Zhang, X. MIClique: an algorithm to identify differentially coexpressed disease gene subset from microarray data. *BioMed Research International* **2009** (2010).

251. Voy, B. H. *et al.* Extracting gene networks for low-dose radiation using graph theoretical algorithms. *PLoS Comput Biol* **2,** e89 (2006).

List of Figures

List of Tables