Εθνικό Μετσόβιο Πολυτεχνείο

Εργαστήριο Βιοϊατρικών Συστημάτων

ΤΟΜΕΑΣ ΜΗΧΑΝΟΛΟΓΙΚΩΝ ΚΑΤΑΣΚΕΥΩΝ ΚΑΙ ΑΥΤΟΜΑΤΟΥ ΕΛΕΓΧΟΥ

Διπλωματική Εργασία

A computational method for comparing pharmacological compounds based on gene expression

Φοιτητής: Σάρδης Αντώνιος

Επιβλέπων καθηγητής: Αλεξόπουλος Λεωνίδας

Αναπληρωτής Καθηγητής ΕΜΠ

ΑΘΗΝΑ, ΟΚΤΩΒΡΙΟΣ 2019

Εθνικό Μετσόβιο Πολυτεχνείο

Εργαστήριο Βιοϊατρικών Συστημάτων

ΤΟΜΕΑΣ ΜΗΧΑΝΟΛΟΓΙΚΩΝ ΚΑΤΑΣΚΕΥΩΝ ΚΑΙ ΑΥΤΟΜΑΤΟΥ ΕΛΕΓΧΟΥ

Διπλωματική Εργασία

A computational method for comparing pharmacological compounds based on gene expression

Φοιτητής: Σάρδης Αντώνιος

Επιβλέπων καθηγητής: Αλεξόπουλος Λεωνίδας
Αναπληρωτής Καθηγητής ΕΜΠ

Εγκρίθηκε την 4η Οκτωβρίου 2019 από την τριμελή επιτροπή:

Αλεξόπουλος Λεωνίδας
Αναπληρωτής Καθηγητής ΕΜΠ

Αντωνιάδης Ιωάννης
Καθηγητής ΕΜΠ

Κυριακόπουλος Κωνσταντίνος
Καθηγητής ΕΜΠ

ΑΘΗΝΑ, ΟΚΤΩΒΡΙΟΣ 2019

Abstract

Compound identification is the most crucial step in the drug discovery process. Today several computational methods focusing on chemical structure that aim to discover the optimal structure for a specific target or disease exist. Since the advent of the Connectivity Map and the release of large scale gene expression data following compound treatment, new methods that aim to discover compounds based on their biological function similarity have been proposed. These methods aim to discover compounds that cause a specific effect on cellular models, rather than relying purely on their chemical structure. The consensus of the two avenues (structure-based and function-based) is that similar compounds in structure will cause similar effects, but the opposite rarely holds. The aim of this project is twofold. First, to create new methods of calculating biological effect similarity based on gene expression data from compound perturbations. Second, to create a model that can predict the biological effect similarity of compounds from structural data, thus augment traditional structure-based screening approaches.

On this front, first of all, the only attempt made on the above mentioned purposes is recreated and examined. It is about investigating the correlation between the chemical structure and the gene expression similarity of compound pairs.

Afterwards, novel ways of processing the data at the transcriptional level are assessed and compared, aiming to the discovery of a characteristic able to decide whether a pair of pharmaceutical compounds are similar or not.

Moreover, a new method of comparing compounds is defined and investigated. The foundation of this method consists of prior biological knowledge about protein networks and is then compared to the already existing ones.

Finally, a machine learning model, which utilizes structural data in order to determine the pairwise transcriptional level similarity of compounds, is tested.

Περίληψη

Η ταυτοποίηση των χημικών ουσιών αποτελεί το πιο κρίσιμο βήμα στην διαδικασία της ανακάλυψης φαρμάκων. Σήμερα υπάρχουν αρκετές υπολογιστικές μέθοδοι που επικεντρώνονται στη χημική δομή προκειμένου να προσδιορίσουν την βέλτιστη χημική δομή για μια συγκεκριμένη ασθένεια. Με την δημιουργία του Connectivity Map και την διάθεση μεγάλου όγκου δεδομένων γονιδιακής έκφρασης, προερχόμενων από την δοκιμή χημικών ουσιών, νέες μέθοδοι που στοχεύουν στην ανακάλυψη νέων ουσιών με βάση την βιολογική τους επίδραση έχουν προταθεί. Αυτές οι μέθοδοι αφορούν την εύρεση ουσιών που εμφανίζουν συγκεκριμένες επιδράσεις σε κυτταρικά μοντέλα, χωρίς να επικεντρώνονται μόνον στη χημική τους δομή. Το κοινό έδαφος και των δύο προσεγγίσεων (χημικής δομής και βιολογικής επίδρασης) είναι πως ουσίες που παρουσιάζουν όμοιες χημικές δομές προκαλούν αντιστοίχως και όμοιες βιολογικές επιδράσεις, με το αντίθετο να ισχύει σπανίως. Ο στόχος της παρούσας μελέτης είναι διπλός. Αρχικά, στοχεύει στην δημιουργία νέων μεθόδων υπολογισμού ομοιότητας της βιολογικής επίδρασης ουσιών, οι οποίες βασίζονται σε πειραματικά δεδομένα γονιδιακής έκφρασης που προκύπτει από τη δράση ορισμένων ουσιών. Κατά δεύτερον, επιχειρείται η δημιουργία ενός μοντέλου που είναι σε θέση να προβλέψει την ομοιότητα της βιολογικής επίδρασης ουσιών βασιζόμενο στην χημική τους δομή, παρέχοντας έτσι μια διαφορετική προσέγγιση πέρα από τις παραδοσιακές που βασίζονται αποκλειστική στη χημική δομή.

Βάσει αυτών, πρώτα από όλα γίνεται αναπαραγωγή και μελέτη της μόνης διαθέσιμης έρευνας που αφορά τα παραπάνω. Η έρευνα αυτή αφορά την μελέτη της συσχέτισης μεταξύ της χημικής δομής και της γονιδιακής έκφρασης ζευγών χημικών ουσιών.

Εν συνεχεία, νέες μέθοδοι επεξεργασίας των δεδομένων μεταγραφικού επιπέδου αξιολογούνται και συγκρίνονται, στοχεύοντας στην εύρεση χαρακτηριστικών που να είναι σε θέση να κρίνουν αν ένα ζεύγος φαρμακολογικών ουσιών είναι όμοιο ή όχι.

Επιπλέον, μια νέα μέθοδος σύγκρισης χημικών ουσιών ορίζεται και μελετάται. Στη βάση της μεθόδου αυτής βρίσκεται η γνώση σαφώς ορισμένων δικτύων πρωτεϊνών. Η μέθοδος αυτή συγκρίνεται με τις ήδη υπάρχουσες.

Τέλος, αξιολογείται ένα μοντέλο μηχανικής μάθησης που αξιοποιεί τα δομικά δεδομένα ουσιών για να προσδιορίσει την μεταξύ τους ομοιότητα σε μεταγραφικό επίπεδο.

Acknowledgements

This project and the work that came with it wouldn’t have been realized without the support and guidance of many great people.

More than an academic project

First of all I would like to thank professor Leonidas Alexopoulos for giving me the opportunity and the privilege to be part of his team and participate in one of the most up to date labs in NTUA. His advisory and guidance meant a lot for me and I highly appreciate it.

Moreover, I am more than grateful to meet and work with Chris Fotis. He is not only a mentor for me but also a friend. Without his leadership and help this project would have been a much harder task.

Finally, I am thankful to every single member of the lab that made this whole time a joyful period for me. They became more friends than lab mates to me.

The ones that were always there

Last but not least, I would like to thank my family and friends for being always there and providing me with love and support all these years. I wouldn’t have been the person I am without them. I am grateful to them and I wish them all the best.

Table of Contents

1 Introduction

1.1 Systems Biology

Systems biology refers to an interdisciplinary field, whose boundaries are not strictly defined. It is characterized by an increasing diversity that stems from its young nature. This makes the establishment of a definition for this field a hard task that usually contains a dose of subjectivity from the person giving it. In the case of Systems Biology field, the qualities that characterize it are diversity, simplicity and complexity [1]. As contradicting as they might be, these values compose the foundation that supports this rapidly developing field. Diversity is directly linked to every aspect of biology, from the living species to the proteins, their structure and functionality. Simplicity is related to the goals of the research in biological sciences. Researchers constantly try to construct or identify well-defined and strict principles that rule various biological phenomena. With these structures, they aim to simplify the complex biological processes through interpretable models that are more likely to provide insight into the biological principles of these processes. Last but not least, complexity is a feature of every living system. Without their complexity, the biological systems would not be as intriguing as they are and the Systems Biology research would not have demonstrated the development at the extent that it has.

A simple definition could address Systems Biology as a holistic and multisided approach to explore and model the subsystems that lie under the complexity of bigger biological systems. The revelation of the possible links between the subsystems and their interactions is attempted through the integration of many sciences, such as biology, engineering, informatics, physics, chemistry and computer science. The modeling of these systems comes with the potential of new discoveries regarding drugs, treatments and diseases' biological profiles.

One of the main applications of Systems Biology is drug discovery. This field exploits large collections of protein microarrays, datasets and algorithms for data mining and analysis. The combination of all the latter mentioned enables the description and understanding of the seemingly complex biological systems that are affected by the drugs [2]. The data which enable this process to take place are originating mainly from OMICS. Their characteristics are mentioned in the following paragraph.

1.2 OMIC data

OMIC data, also known as OMICS refer to a group of technologies that are used to measure some characteristic of a large family of cellular molecules, such as proteins or genes. The results of these technologies are further processed to provide an understanding on the roles, relationships and actions of the different types of molecules that make up the cells of an organism. The suffix "–omics" is appended to this group of technologies. Some of the technologies that are part of this group are the following.

- *Genomics*
  These technologies are focused on the study of an organism's genome. This includes genes, exons, introns, promoters, transcription factors and other genome-related functions (e.g. Illumina HiSeq, Sanger sequencing, de novo assembling).
- *Transcriptomics*
  The technologies included in transcriptomics are aiming to study the transcriptome of an organism (e.g. RNA sequencing, microarrays, real-time PCR).
- *Proteomics*
  Proteomics are technologies that have as main focus the study of the proteins or the proteome of an organism (e.g. ELISA, MSIA).
- *Metabolomics*
  The technologies of this group are targeting on studying the metabolome or the secondary metabolites of an organism, such as amino acids, sugars etc. (e.g. GC-MS, LC-MS).
- *Lipidomics*
  These technologies are focused on studying the cellular lipids of an organism or a biological system (e.g. ESI, MALDI).
- *Catalomics*
  These technologies are used to study enzyme (catalysts) or other biocatalysts of an organism or a system (e.g. chemical reactions).

The present work's initial data are collected using Transcriptomics. More specifically, microarray experiments are used and their results, which are in format of gene expression profiles, are processed with various ways and techniques to acquire the insights displayed in this report.

1.3 Gene Expression

In every living organism, genes encode proteins, which in their turn determine the cell function. As a result, the thousand genes that are available and expressed in a cell define its capabilities. In order for the proteins to be encoded by the genes, two key steps have to happen. These steps are transcription and translation. Through these steps, the information stored in an organism's DNA is converted to instructions for building proteins and other molecules. In order for the process to be properly described, the DNA and RNA definitions have to be provided.

DNA

DNA, or deoxyribonucleic acid, is the hereditary material in almost all of the organisms. The information included in it is stored as a code consisting of four chemical bases. These bases are adenine (A), guanine (G), cytosine (C) and thymine (T). Human DNA is made up of approximately three billion bases. The way that these bases are ordered and their sequence determines the information which are at the cell's disposal for the survival of the organism. DNA's bases pair in a specified way. Adenine pairs with thymine and cytosine pairs with guanine forming some units that are known as the base pairs. All the bases are also attached to a phosphate molecule and a sugar molecule. These three compounds together form the nucleotide. The nucleotides form two long strands and these in their turn form the widely known double helix of the DNA. The base pairs are the linking parts of the helix's vertical sides. These sides are composed by the sugar and the phosphate molecules that were previously mentioned. The DNA structure is presented below.

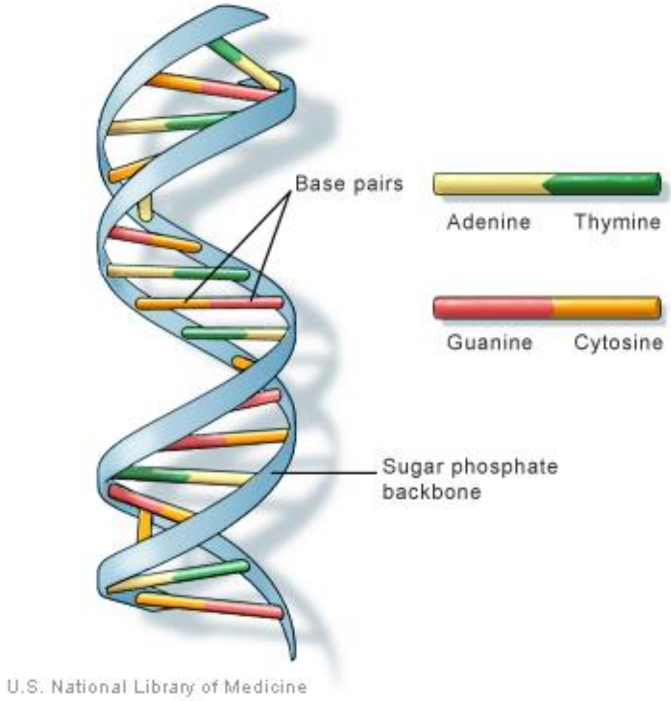


Figure 1 DNA structure (Source: U.S. National Library of Medicine)

RNA

RNA, or ribonucleic acid, is one of the three major biological macromolecules that are essential for all known forms of life, with the remaining two being DNA and proteins. RNA is a nucleic acid similar in terms of structure and properties to DNA, but it consists of only one strand of bases. Moreover, its bases are the same with DNA's, except from one. Instead of the base thymine (T), RNA has another base called uracil (U). So, RNA's bases are adenine (A), guanine (G), cytosine (C) and uracil (U). There are many types of RNA, but the three most significant and well-known are the messenger RNA (mRNA), the transfer RNA (tRNA) and the ribosomal RNA (rRNA). One of the molecules that are useful for the explanation of gene expression process is the messenger RNA. The mRNA is a molecule in cells that carries codes from the DNA that is present in the cell's nucleus to the sites where protein synthesis takes place, in the cytoplasm (mainly ribosomes). The other molecule that is present in the processes is the transfer RNA (tRNA). This is a small molecule in cells which carries amino acids to ribosomes, where they are linked to form proteins.

With DNA and RNA defined, one can proceed to the explanation of the processes which compose the gene expression. These are the transcription and the translation. At first, transcription happens and the following process if translation.

1.3.1 Transcription

Transcription is called the process through which the DNA of a gene is copied to produce the RNA's transcript, messenger RNA (mRNA). The whole process is performed by an enzyme called RNA polymerase. It makes use of the available bases from the cell's nucleus to construct the mRNA. An illustration of the process is presented below.

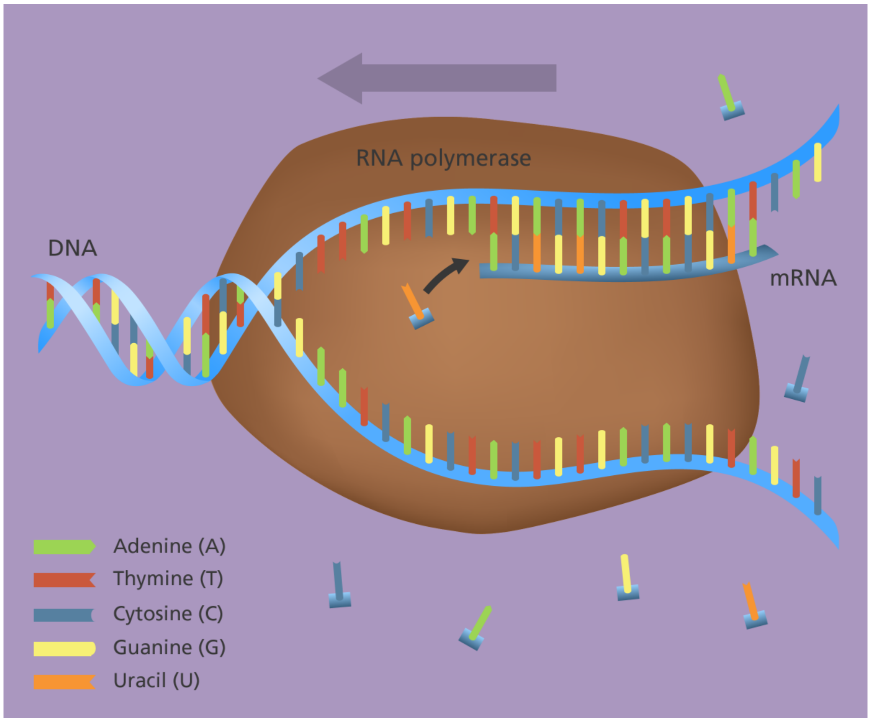


Figure 2 Transcription process (Source: yourgenome.org)

1.3.2 Translation

Translation is the second and final process in the gene expression sequence. In order for it to emerge, the mRNA has to deliver the transcribed information from the DNA to the ribosomes. The ribosomes are the sites where the cell's proteins are produced. Once the mRNA reaches the ribosomes, another RNA molecule is involved in the process. This is the transfer RNA (tRNA), which is a carrier molecule and its task is to read the genetic information provided by the mRNA. The information that mRNA carries is read three letters (one codon) at a time by the tRNA. Each time, the codon read specifies an amino acid (for example, 'GGU' encodes an amino acid called glycine). It is worth mentioning that the number of unique existing amino acids is twenty, while the number of possible codon combinations is sixty-four. This fact points out the case that it is possible that more than one codon correspond to the same amino acid. Each amino acid is bound to its own, specific, tRNA molecule. After all the sequence provided by the mRNA is read, every tRNA molecule delivers its amino acid to the ribosome, where the mRNA is located, and binds temporarily to the corresponding codon on the mRNA molecule. Then, the amino acids are released and all of them are joined together in a long sequence, called a polypeptide. Polypeptides are amino acid chains that make up proteins. One protein can be made up of one or mode polypeptides. So, the described process keeps on happening until the protein is formed. A visual description of the translation's process is shown in the following figure.

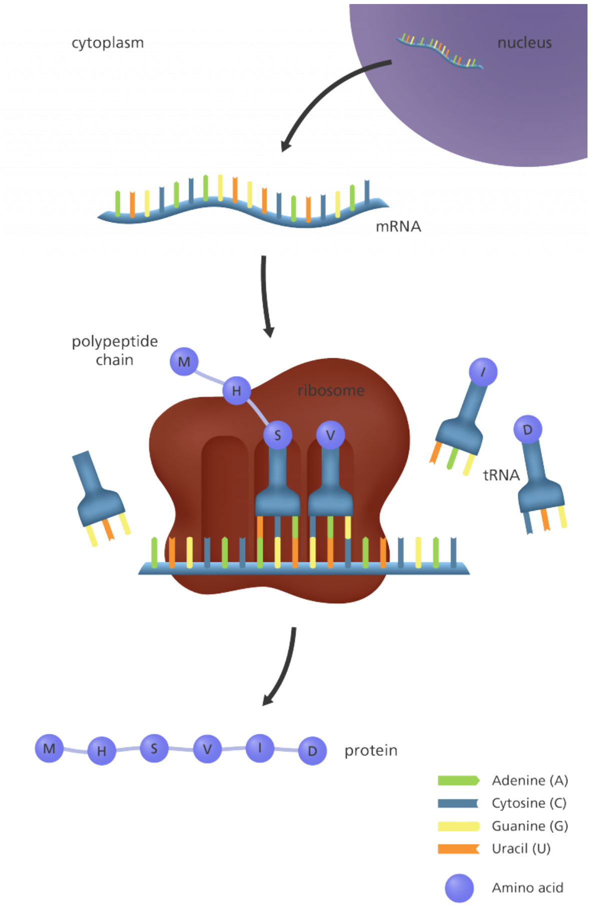


Figure 3 Translation process (Source: yourgenome.org)

The processes described above make up the gene expression process. This process is vital for any living organism, since it is the one that defines the organism's well-being and is critical to its regulatory systems. By monitoring the gene expression activity, an insight can be gained into the biological processes that take place in an organism. These processes might be related to how the organism deals with a disease, or even environmental changes. On this ground, gene expression analysis can provide significant information about the underlying biological systems that operate in a living organism and aid its adaptation to external or internal perturbations.

1.4 New metrics based on prior knowledge

Apart from the Gene Expression analysis, plenty of other ways to approach the biological processes functionality exist. They all exhibit gene expression as a starting point and by utilizing prior biological knowledge, higher level biological characteristics can be defined and constructed. There are four such categories, namely Transcription Factors, Signaling Pathways, Gene Ontology (GO) Terms and Protein Networks.

1.4.1 Transcription Factors

Transcription factors or sequence-specific DNA-binding factors are some proteins that control the rate of transcription of the genetic information included in the DNA to messenger RNA (mRNA)[3]. In fact, transcription factors are proteins that help turn specific genes "on" or "off" by binding to nearby DNA. The transcription factors can be either activators, in case they boost a gene's transcription, or repressors, in case that they decrease a gene's transcription. Moreover, groups of transcription factor binding sites that are called enhancers and silencers can turn a gene on or off in specific parts of the body. The most significant property of the transcription factors is that they allow cells to perform logic operations and in fact combine different sources of information to determine whether a gene will be expressed or not. As mentioned before, at the Transcription process analysis, RNA polymerase has to attach to the DNA of a gene in order to make an RNA molecule. While in some organisms this process doesn't require anything else, in humans and other eukaryotes, an extra step exists. In these cases, the transcription can happen only with the help of some proteins called basal (general) transcription factors. These proteins are part of the cell's core transcription tools and are crucial for the transcription of any gene. The interesting fact about transcription factors is that many of them is not of a general kind, instead they are specifically linked to the activation of well-defined sets of genes. This property of the transcription factors can be utilized as prior knowledge to build new data that stem from the gene expression ones. From the gene expression analysis, the expression levels of each gene that is monitored through these experiments are found. Setting these levels as a benchmark and combining them with the prior knowledge about the gene sets that the transcription factors affect, new data about the activity of the transcription factors can be constructed. On the one hand, these new data might contain some level of noise implied from the prior knowledge data, but on the other hand, these new data refer to a higher biological level and this fact leads to less errors. The transcription factors' level is considered as higher than the gene level, as it groups the information that the genes provide in a smaller dataset. Since one transcription factor is defined by a group of genes, the outliers have smaller effect on the data and the activity is not biased towards the outliers, that are minorities, but is derived according to the expression that the majority of the genes of a set exhibit.

1.4.2 Signaling Pathways

A signaling pathway is a group of molecules in a cell, which work together with a common goal. This goal usually refers to the stimulation of one or more cell functions. Right after the first molecule in the pathway receives a signal and is activated, a chain process begins with molecules activating or deactivating other ones that are part of the pathway. The process will finish once the goal of the pathway is achieved. A possible malfunction of a signaling pathway may result to disease or even cancer. Signaling pathways provide a more structured form of genes. In contrast to gene expression that takes account of every gene separately, signaling pathways group together genes that take part in specific functions in the cell, thus providing a wider and more causal perspective to the gene expression topic. This method can be also considered as a higher biological level method, since it uses prior knowledge regarding the signaling pathways of the cell to derive the ones that are activated or deactivated under certain circumstances.

1.4.3 Gene Ontology (GO) Terms

Gene Ontology (GO) is a structured vocabulary that is widely used for gene products annotation [4]. In fact, a GO annotation reveals a link between a gene product type and a molecular function, biological process or cellular component type. This means that the annotation contains information about the capabilities of a gene product in terms of contribution to biological processes or cell's functions [5]. The genes that are present in an experiment are being classified, according to the Gene Ontology system, which is presented as prior knowledge. This knowledge is utilized to categorize the genes according to the functions that they contribute to. These functions are the GO terms. One gene can and is usually expected to be part of more than one GO term, since genes generally contribute to many biological functions. Using this approach, a new dataset is produced, which consists of the GO terms' activity, based on the gene expression analysis. If for example the most of the genes that belong to a GO term are over expressed, then this term's activity will be high. By grouping the genes according to the biological functions that they contribute to, outliers are once more absorbed and their effect on the final result is lower. In this way, a higher level biological characteristic is derived from the initial low-level gene expression data by using Gene Ontology as prior knowledge.

1.4.4 Protein Networks

Protein-protein interactions (PPIs) [6] are crucial to the majority of a cell's processes. According to this, in order to gain knowledge and understanding of the cell's processes, the study of PPIs is really important. Moreover, in a reverse way of thinking, one can use and analyze the activity in the PPI level to generate insights into the underlying processes of the cells. When the PPIs are linked, the protein-protein interaction networks (PPIN) are built. These are directed networks that visualize the way that the proteins interact with each other, which proteins are linked and as a result which is the effect of the proteins on the network. Knowing the above mentioned PPIN and the gene expression or the transcription factor activity of an experiment, the protein network of a specific cell can be constructed, by combining these two characteristics. This network describes sufficiently the activity that takes place in the cell and is unique for every experiment, provided that the gene expression varies each time that an experiment is carried out under different circumstances. Thus, PPIN are a higher level biological attribute that is defined using a known PPI as a prior knowledge factor and the gene expression or the transcription factor activity of an experiment. This attribute includes some topological knowledge, based on the PPI, and also the biological experimental results. Their combination can lead to characteristic protein networks for the cells that are subjected to research each time.

1.5 Gene expression in drug perturbations

A perturbation in terms of system biology is the alteration of a biological system's function caused either by internal or external factors. On the one hand, internal factors may be linked to biological processes that are damaged or changed in the system. On the other hand, external stimuli can be caused by environmental changes or a drug inhibition for example.

Once a perturbation occurs, it has an apparent impact on the gene regulatory networks. This means that some gene expression levels are higher or lower than what they are under normal circumstances. Also, transcript levels are affected, since they are directly linked to the gene expression levels. All of these changes can be described as perturbations in the cell's transcriptome. The transcriptome is the total of all the messenger RNA (mRNA) molecules that are expressed from the cell's genes.

The perturbations studied in this report are caused by the drug inhibition. When a drug is inserted to a biological system, it interacts with its targets, but also with off-targets. These interactions trigger signaling alterations that cause the above mentioned perturbations in the cell's transcriptome. The perturbations result to differentiations of the cell's gene expression. By monitoring and measuring these differentiations, some conclusions can be derived regarding the effect of each drug.

A great deal of research has been carried out on the field of drug-induced perturbations. The most significant and well-known is the construction of Connectivity Map (CMap). Two version of CMap exist. The first one, CMap 02 (build 02) [7]. uses an Affymetrix-based dataset that is the result of 1300 compounds tested on four cell lines. The latest version [8] utilizes the L1000 assay to generate over one million gene expression profiles. These two gene expression calculation methods will be analyzed in the Materials and Methods section.

Both of the CMap versions are utilized through the report, for reasons that will be clarified in the following chapters.

1.6 Comparing drugs from gene expression

The reason of the CMap's existence is the potential that is provided through the gene expression profiles in terms of drug comparison. The researchers' community is using gene expression as a means to uncover similarities between virtually completely different drugs. Two drugs could cause similar gene expression to a cell, although their target genes differ. This phenomenon might be related to the off-target genes that each drug affects and are usually neglected, since drugs are defined and used according to their target genes. In addition, a drug's target gene is possible to affect other signaling pathways along with the one that is considered significant for the drug. These two possibilities can explain the foundations of the concept of using gene expression to study drug similarity.

To sum up, up to this point, gene expression is defined as a metric of approaching the similarity between two drugs. This similarity is completely based on the biological effect (gene expression) that each drug has on the cells and consequently this is a purely biological metric.

1.7 Motivation

Since the establishment of the biological metric through CMap, a lot of research is being carried out with respect to the correlation of the CMap data with other possible drug similarity metrics. A significant part of this research investigates the possibility of relating the chemical structure, and similarity by extension, with the biological similarity acquired from the CMap database.

One of the milestones in this field is the research paper published by Fransesco Sirci et al. entitled as "Comparing structural and transcriptional drug networks reveals signatures of drug activity and toxicity in transcriptional responses" [9]. This publication deals, among others, with the subject of whether a correlation between the chemical structure and the biological distance of two drugs can be found. The methods that are used to compute the chemical distance of two drugs will be analyzed in the upcoming paragraphs. Regarding the biological (or transcriptional, as it is such named in the publication) distance, the CMap 02 database is used and the distance is based on the gene expressions provided from the specific database.

The pairwise distances for the drugs are plotted against each other and the following figure is the result provided by the publication.

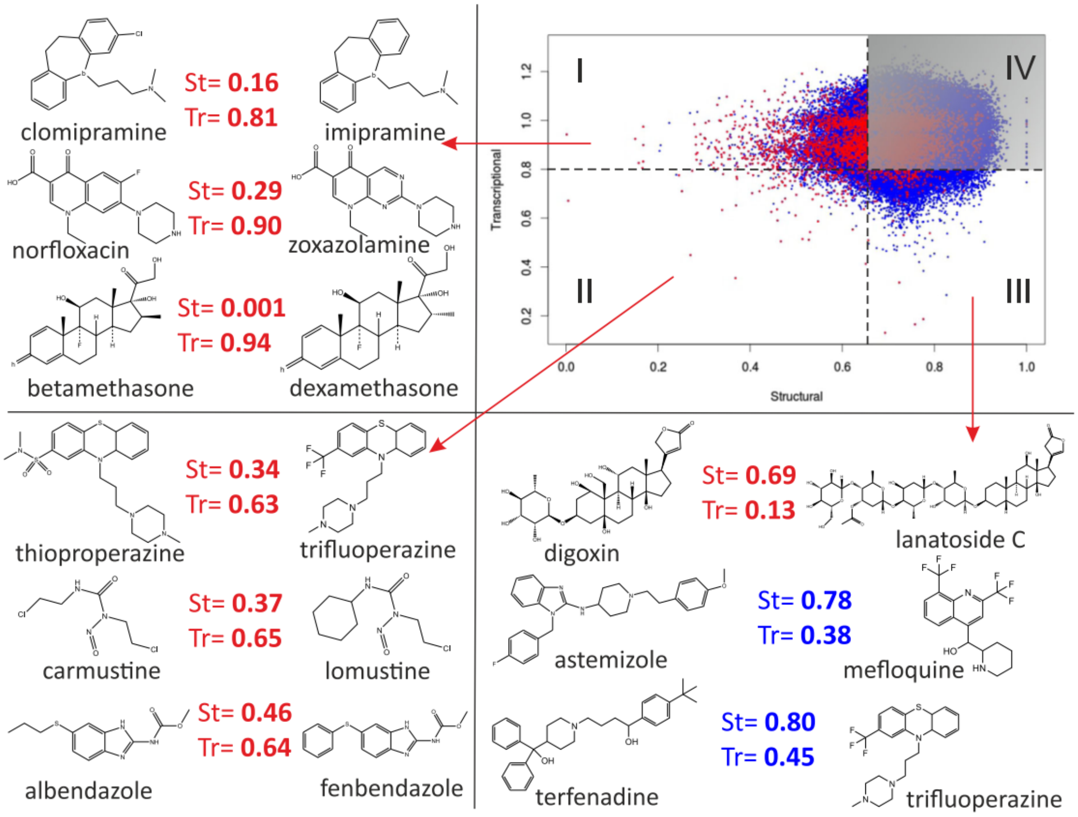


Figure 4 Correlation between structural and transcriptional distances (Source: NPJ Systems Biology and Applications 3.23 (2017))

It is obvious from the plot format that no apparent correlation exists between the structural and the transcriptional distance of the drugs. There are some findings presented at the publication regarding the toxicity and similarity of some drugs. Apart from these, no correlation can be established regarding the two metrics that are under investigation. This lack of correlation can be explained either by errors or simplifications in the distance computations or by the acknowledgement that no correlation can exist between these two distances. In order for further research to be carried out on this topic, the first option is considered as the outcome of the upper mentioned publication.

So, is it possible to improve the calculation of the distances presented in the publication, or even define new distances that will provide a better insight into the field of similar drugs?

This is the main question dealt in this report and the motivation behind it.

2 Materials and methods

2.1 Initial data information

Through the whole report's tests and results two datasets take part. The first one consists of the gene expression data that come from CMap 02 (build 02) and the second one comes from the latest version of CMap. From now on the first will be mentioned as CMap 02 and the second (latter) one as CMap. As mentioned in the introduction (see 5.Gene expression in drug perturbation), each dataset is constructed using a different method. CMap 02 uses an Affymetrix-based method, while CMap uses the L1000 assay. The two methods are explained below.

The Affymetrix-based dataset (CMap 02) is built based on microarrays. Microarrays are used to determine which genes are present in a sample and at what quantities. Microarrays contain oligonucleotide probes that are capable of binding specific parts of mRNA from a sample. There can be many probes from the coding regions of any given gene. At the next step, mRNA is fluorescence labeled and visualized as an image. This image's intensity is correlated to the amount of mRNA in the sample and according to this density, the expression of the genes is defined.

When it comes to the newest CMap version, the data are generated by the L1000 assay. This is a high-throughput gene expression assay which measures the mRNA transcript abundance of 978 "landmark" genes from human cells. The "L" in L1000 refers to the Landmark genes that are measured. The measurement is made using 500 colors of Luminex beads. With this method, two transcripts are identified by a single bead color. Moreover, there are 80 control transcripts whose expression is measured. These are specifically chosen due to their invariant expression across all of the cell states.

Apart from the different methods that are mentioned above, the two datasets contain different cell lines as well. CMap 02 contains four cell lines, MCF7, PC3, HL60 and SKMEL5. Their characteristics are presented below.

MCF7 is a human breast epithelial adenocarcinoma cell line derived from pleural effusion (ATCC# HTB-22). MCF7 cells are cultured in DMEM supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin-glutamine.

PC3 is an epithelial cell line established from human prostate adenocarcinoma (ATCC CRL-1435). PC3 cells are cultured in RPMI supplemented with 10% fetal bovine serum, 1% sodium pyruvate and 1% penicillin-streptomycin-glutamate.

HL60 is a human promyelocytic cell line established by leukopheresis from promyelocytic leukemia (ATCC CCL-240). HL60 cells are cultured in RPMI supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin-glutamate.

SKMEL5 is a human malignant melanoma cell line derived from a metastatic axillary node (ATCC HTB-70). SKMEL5 cells are cultured in DMEM supplemented with 10% fetal bovine serum, 1% penicillin-streptomycin-glutamate, 1% non-essential amino acids and 1% sodium pyruvate.

On the other hand, from the new CMap five cell lines are used. They are MCF7, PC3, A375, HEPG2 and NPC. The first two are also contained at CMap 02, so they won't be explained, while the other three are presented below.

A375 is a human epithelial cell line established from human malignant melanoma (ATCC CRL-1619). A375 cells are cultured in DMEM supplemented with 10% fetal bovine serum.

HEPG2 is a human liver epithelial cell line derived from hepatocellular carcinoma (ATCC HB-8065). HEPG2 cells are cultured in EMEM supplemented with 10% fetal bovine serum.

Finally, NPC is a human nasopharyngeal carcinoma cell line.

The two datasets differences can be spotted in another final point. While the CMap 02 provides around 7,000 expression profiles that represent 1,309 compounds, the latest version of CMap provides over 1.5 million gene expression profiles that represent around 5,000 small-molecule compounds and around 3,000 genetic reagents.

The data that are described are the initial ones of the tests. They are all downloaded by the respective online sites of the CMap versions. All the perturbations are available in CEL file format, with each CEL file corresponding to a specific perturbation. Also, a list accompanies each one of the two datasets with information about the perturbations. This information can be the cell line that each perturbation is applied, the CEL file that corresponds to each perturbation, the applied drug's full name or ID, the quality of the experiment (only for newest CMap version) or even the control sample of each perturbation.

These data sources will be processed and combined in order for the experimental data from CMap to be introduced to the algorithms and functions that are utilized in the following methods in the proper format.

2.2 Data acquisition

The initial format of the data used as input to the analysis is the results of biological experiments that measure how much the genes of the dataset are transcribed (mRNA levels). The experimental design followed uses the typical approach of case vs. control. This means that the measured genes of a sample can either be considered as perturbations (case) or as control. Even though both CMap versions follow a case vs. control format, the data acquisition process of the two versions differs. These processes are presented below.

2.2.1 Biological data

2.2.1.1 CMap 02

Affymetrix genechips

In this version of CMap (CMap 02), Affymetrix genechips [10] are used. In these chips, each gene is represented by 11 to 20 'probe pairs'. The probe pairs are 3' biased. Each probe pair consists of the Perfect Match (PM) and the MisMatch (MM) probes. MM probes have altered their middle (13th base) base, as shown in the below figure. In this way, the non-specific binding (NSB) can be measured.

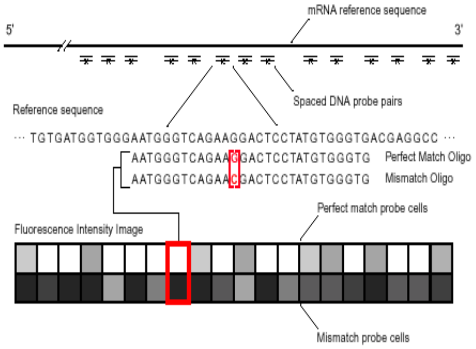

mRNA reference sequence
5'
3'
Spaced DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC ···
AATGGGTCAGAAGGACTCCTATGTGGGTG
Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTGGGTG
Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

Figure 5 Perfect Match (PM) and MisMatch (MM) probes characteristics

The genechips are scanned according to a specific process. The RNA sample is prepared, labeled and then hybridized to a chip. The chip is afterwards fluorescently scanned to provide a pixelated image (.DAT file). A grid is applied to the image in order to separate pixels that are related to single probes. The intensities measured are averaged to compute a unique intensity for each one of the probes (.CEL file). At the last step, the probe level intensities are combined in terms of probe sets to give a final single intensity value for each gene. Probe sets are made up from 20 to 25 unique probes. The CMap 02 dataset consists of multiple CEL files that are produced according to the above described method.

2.2.1.2 CMap (newest version)

ELISA

ELISA (Enzyme-linked Immunosorbent Assay) [11] is a plate-based assay technique. It is used to detect and identify substances, such as peptides, proteins, antibodies and hormones. The elementary enzyme-linked immunosorbent assay (ELISA), or enzyme immunoassay (EIA), is different from the other available antibody-based techniques in a way that in a polystyrene (96 or 384) multi-well plate (which is considered good for protein binding) the binding to solid surface sequentially results in the identification of specific reactions excluding the non-specific ones. This results in the generation of quantitative results. In this method, an antigen has to be immobilized on a solid surface and afterwards complexed with an antibody that is linked to an enzyme. The detection of the assay is carried out by evaluating the enzyme activity of the conjugate which is implemented by incubation with a suitable substrate for the enzyme resulting in a generation of a quantifiable product. Generally, the ELISA technique results in a colored end product which absorbs at a particular wavelength and can be correlated to the quantity of analyte in question present in the sample. An extremely specific antibody-antigen interaction is the utmost critical component of the entire process. This capability of the process to wash away non-specific unbound reactants makes the ELISA technique an influential and reliable means for providing precise information on the analytes even when present within a crude and impurified sample.

This CMap edition provides the data from ELISA, which is considered to be superior to microarrays.

2.2.1.3 Other biological data acquisition methods

RNA sequencing

RNA sequencing (RNA-seq) [12] is a technique that can examine the quantity and sequences of RNA in a sample using next generation sequencing (NGS). It analyzes the transcriptome of gene expression patterns encoded within our RNA. RNA-seq enables the investigation and discover the transcriptome, the total cellular content of RNAs including mRNA, rRNA and tRNA. Understanding the transcriptome is key to the connection of the information on the genome with its functional protein expression. RNA-seq can provide an insight regarding which genes are turned on in a cell and what their level of expression is.

RNA sequencing pipeline consists of three major processes. At first, the RNA isolation happens. Through this process, the RNA is isolated from tissue and is mixed with deoxyribonuclease (DNase). DNase reduces the amount of genomic DNA. The amount of RNA degradation is checked with gel and capillary electrophoresis and is used to assign an RNA integrity number to the sample. This RNA quality and the total amount of starting RNA are taken into consideration during the subsequent library preparation, sequencing, and analysis steps. Then, the RNA selection/depletion takes place. In order to analyze signals of interest, the isolated RNA can either be kept as is, filtered for RNA with 3' polyadenylated (poly(A)) tails to include only mRNA, depleted of ribosomal RNA (rRNA), and/or filtered for RNA that binds specific sequences. The RNA with 3' poly(A) tails are mature, processed, coding sequences. Poly(A) selection is performed by mixing RNA with poly(T) oligomers covalently attached to a substrate, typically magnetic beads. Poly(A) selection ignores noncoding RNA and introduces 3' bias, which is avoided with the ribosomal depletion strategy. The rRNA is removed because it represents over 90% of the RNA in a cell, which if kept would drown out other data in the transcriptome. The final step is the cDNA synthesis. There, RNA is reverse transcribed to cDNA because DNA is more stable and to allow for amplification (which uses DNA polymerases) and leverage more mature DNA sequencing technology. Amplification subsequent to reverse transcription results in loss of strandedness, which can be avoided with chemical labeling or single molecule sequencing. Fragmentation and size selection are performed to purify sequences that are the appropriate length for the sequencing machine. The RNA, cDNA, or both are fragmented with enzymes, sonication, or nebulizers. Fragmentation of the RNA reduces 5' bias of randomly primed-reverse transcription and the influence of primer binding sites, with the downside that the 5' and 3' ends are converted to DNA less efficiently. Fragmentation is followed by size selection, where either small sequences are removed or a tight range of sequence lengths are selected. Because small RNAs like miRNAs are lost, these are analyzed independently. The cDNA for each experiment can be indexed with a hexamer or octamer barcode, so that these experiments can be pooled into a single lane for multiplexed sequencing. A schematic representation of the RNA-seq is the following.

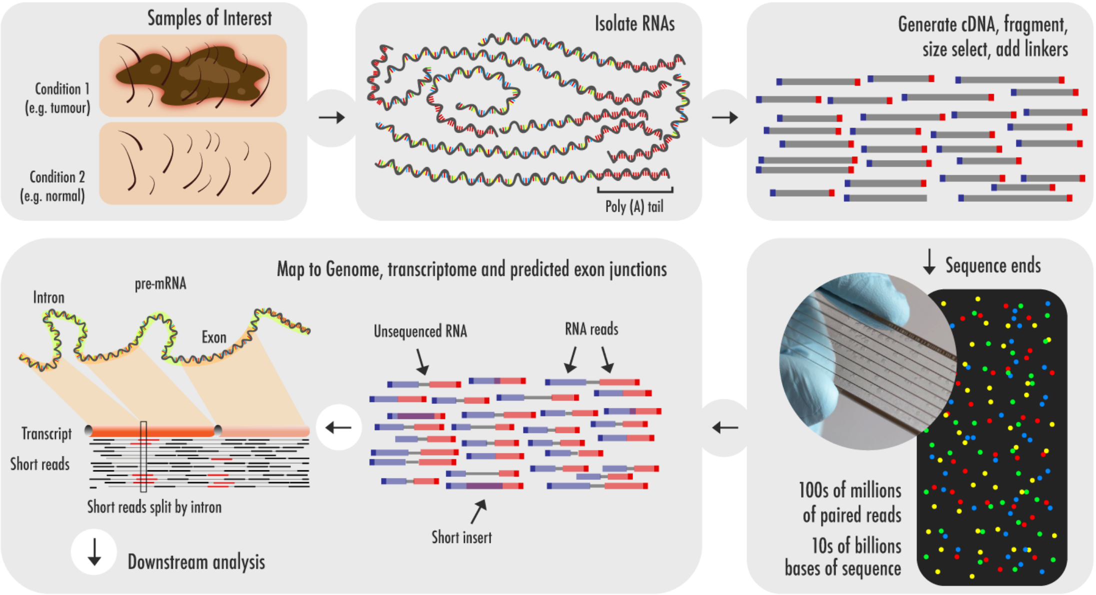


Figure 6 RNA sequencing data generation overview (Source: PLoS Computational Biology 11.8 (2015))

RNA-seq is another method that is widely used and its results are considered to be significantly reliable.

2.2.2 Chemical structure data

At the basis of this project, the investigation of the correlation between the chemical structure and the biological effect of various compounds is found. The initial biological data are the ones mentioned above. Regarding the compounds chemical structure data, they can be distinguished in two major categories. They are either 3-dimensional (3D) or 2-dimensional (2D) structural data.

2.2.2.1 3-dimensional structural data

The 3-dimensional structural data of the compounds are retrieved from the available data of Fransesco Sirci et al. article [9]. What matters is the compounds' distance based on these structural data. For these distances to be computed, at first the canonical SMILES (Simplified Molecular Input Line Entry Specification) that correspond to these compounds and describe their chemical structure were collected. A SMILES is a specification for unambiguously describing the structure of chemical molecules using short text strings. Then, the pairwise chemical similarities were computed using two methods that can be applied to SMILES. The first one is based on a definition of distance between molecular electrotopological states, whereas the second one is based on comparisons between extended-connectivity fingerprints and, making use of a software tool from SciTegic©, computes a Property Distance inversely proportional to chemical similarity.

The final 3D structural distance is equal to 1 minus the electrotopological states (ESF) similarity between the SMILES of the two drugs.

2.2.2.2 2-dimensional structural data

At another approach that is tested, the 2D structural distance of the compounds is calculated. This is also done using the SMILES of the compounds. SMILES can be used to derive the compound fingerprints through the programming language Python. Processing the SMILES with the RDKit library of Python, the compounds' chemical fingerprints are found. Chemical fingerprints can be expressed through multiple types. The five chemical types calculated with this method are the circular, the topological, the Maccs keys and the atom pair fingerprints, along with the topological torsion descriptors. These five chemical metrics are made available for every compound. At the next step, the pairwise similarity of all five metrics is calculated for the total of the perturbations. Then, a weighted average of the five similarities is computed for every perturbation pair and this is how the pairwise similarity is defined. The average has to be weighted, since some of the metrics are correlated with others. The similarity values are in the range of zero to one. Finally, the distance is calculated by subtracting the similarity values from 1, e.g. in case that a pairwise similarity between two compounds is 0.3, then their pairwise distance will be 0.7. Following this method, the pairwise distances of the compounds regarding their 2D chemical structure are defined.

2.3 Data background correction and normalization

Considering that in both of the versions, the initial data are available in intensity values for the genes of each perturbation, the background correction and normalization method [13] that is applied is the same for both of the versions.

The main problem lies to the fact that the subtraction of MM data (see 2.1.1) can be used to correctly measure the NSB, but it introduces noise. On this basis, a method that will provide with positive intensity values is needed. Moreover, carrying out the normalization process at probe level can lead to the avoiding of information loss.

For the above reasons, the method used is Robust Multi-array Average (RMA) [14].

It consists of four stages:

1. Background correction

2. Normalization (across arrays)

3. Probe level intensity calculation

4. Probe set summarization

2.3.1 Background correction

The data of each probe intensity (PM) are formed by its signal intensity (s), combined with its background noise (bg) in a form of simple addition as exhibited below.

$$PM_{ijn} = bg_{ijn} + s_{ijn}$$

The subscripts references are as follows, i refers to the array or sample (i), j refers to the probe (j) and n refers to the probe-set (n) of probe (j).

In this case, the real average noise would be: E $(bg_{ijn}) = \beta_i$

So, the noise present in the analysis is approached as $bg_{ijn} \sim N(\beta_i, \sigma_i^2)$.

The background correction is carried out by utilizing the following convolution model.

$$B\left(PM_{ijn}\right) = E\left[s_{ijn} \middle| PM_{ijn}\right]$$

For the model to be applied, a strictly positive distribution for the intensity signal (s) is assumed $(s_{ijn} > 0)$. This distribution is also assumed as exponential with a parameter κ $(s_{ijn} \sim Exp\ (\kappa_{ijn}))$. That, combined with the noise approach, means that also the background corrected signal is positively distributed.

For example, let's consider the model as O = S (signal) + N (noise), where O is the measured PM intensity. Then, let S follow an exponential with parameter α and let N follow a normal with μ,σ. In this case, the background corrected values are as follows.

$$E(S|O=o) = a + b\,\frac{\varphi\left(\frac{\alpha}{b}\right) - \varphi\left(\frac{o-\alpha}{b}\right)}{\Phi\left(\frac{\mathrm{\alpha}}{\mathrm{b}}\right) + \Phi\left(\frac{\mathrm{o}-\mathrm{\alpha}}{\mathrm{b}}\right) - 1}$$

Where $a = o\ - \mu\ - \sigma^2\alpha$, $b = \sigma$ and ϕ,Φ are the normal distribution and density functions.

2.3.2 Normalization (across arrays)

Non-biological factors can have an effect on the variability of the data. In order for the data to be reliable across different arrays, the differences caused by non-biological factors have to be minimized. Normalization is a process of reducing the unwanted variance across arrays by using information from multiple chips. Moreover, quantile normalization is a method to make the distribution of probe intensities the same for every array.

An example where the normalization need can be explained is when two light sources of different strength are used for the image acquisition of two arrays. It is obvious that the signals produced by the stronger light source will be higher. This is an undesired characteristic, since it is caused by technical reasons. The removal of such technical variability is the aim of normalization.

A schematic explanation of the quantile normalization is presented below.

Let's assume three samples that are processed, measuring four probes at each one of them. The probes are distinguished by different colors and the results are as show in the following image.

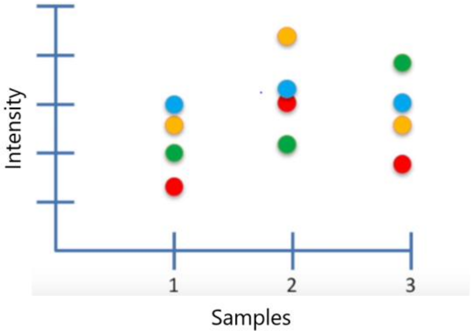


Figure 7 Raw experimental intensities

According to these results, the probe intensity distribution of each sample is different. To achieve similar distribution between all samples, the probes with the highest intensity from each sample are used. Their mean is calculated and this is their intensity value at the quantile normalized form. According to this, the same process is applied to the next probes as presented in the following image, where the method is applied for the highest and the second-highest intensities. At the left are the raw intensity measurements and at the right are the quantile normalized intensities.

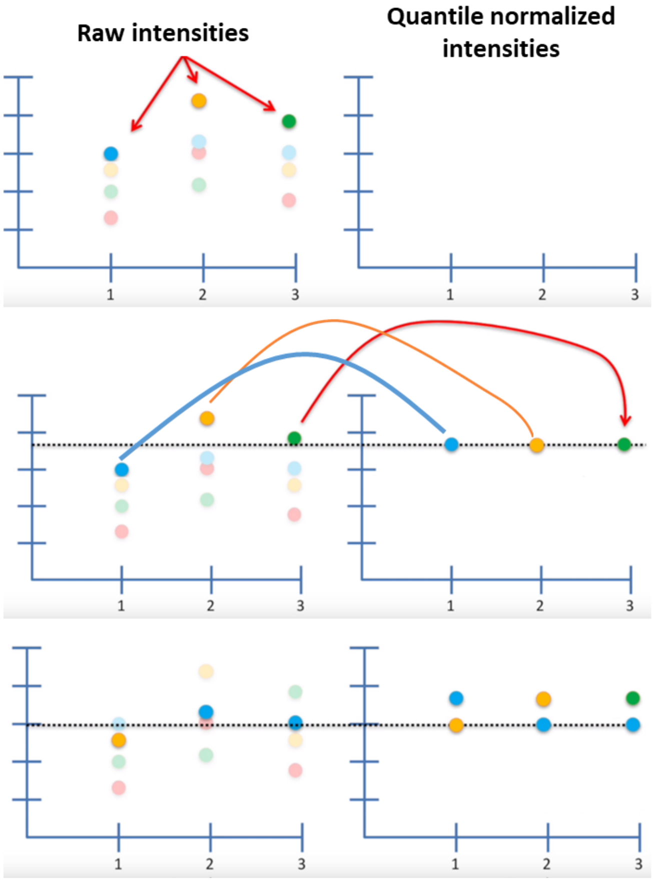


Figure 8 Quantile normalization process

With the completion of the quantile normalization, the intensity values across the samples are the same, but the original ranks are preserved, as demonstrated at the following image. The normalized data have identical quantiles.

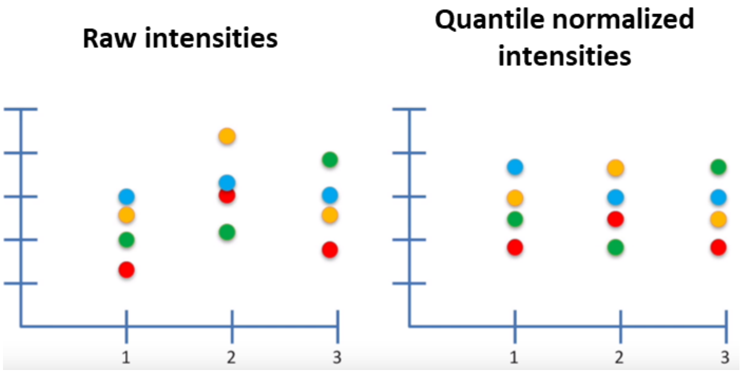


Figure 9 Quantile normalization final results

2.3.3 Probe level intensity calculation

For each probe, a final intensity value is calculated after the background correction and the normalization process. This value is transformed to a $\log_2$ scale to provide the probe level intensity. That means, if for example the calculated intensity of a probe after the background correction and the normalization is A, then the value that will be used for the further process when referring to the intensity will be the $\log_2(A)$. This transformation happens so that the numbers that the further process deals with are in a manageable and convenient scale that $\log_2$ provides.

2.3.4 Probe set summarization

According to the experiment's set up, each probe is part of a larger probe-set that is linked to a specific gene. For this reason, the probe intensities have to be summarized in terms of each probe-set to acquire an expression value for each gene. In RMA, which is the method used, the probe-set intensities are calculated by utilizing a robust linear model that takes into account probe and array effects.

For each probe-set (gene) k, the $\log_2$ transformed probe intensities are modeled as follows.

$$Y_{i,j} = \mu + \delta_i + \alpha_j + \varepsilon_{i,j}$$

The symbols of this equation are explained as:

$Y_{i,j}$ is the signal intensity of probe j and array i in the probe-set

δ_i is the parameter of the model for the array effect (row effect)

α_i is the parameter of the model for the probe effect (column effect)

μ is a constant

$\varepsilon_{i,j}$ are the residuals

Finally $\mu_i = \mu + \delta_i$ is the expression value of gene k for the array i.

The model is fit using median polish until the residuals converge. Median polish is an exploratory data analysis technique. It provides a way to quantify a two-way table using a fixed value μ along with a column and row effect. According to median polish method, the row and column medians are alternately removed (substracted) until the summary of absolute residuals converges. This process is carried out for one probe-set (gene) k at a time. An illustration of the method is shown below.

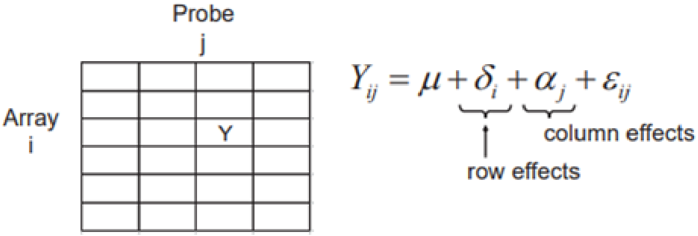


Figure 10 Median polish method

After the model converges, the final calculated μ_i is the expression value of gene k for the array i.

2.4 Enrichment methods

Through the above described methods (3. Background correction and normalization), the gene expression is calculated for all the genes in each array. In order to acquire data about other biological factors like transcription factors, signaling pathways, GO terms or protein networks, some enrichment methods have to be applied to the gene expression results. For the enrichment calculations, the programming language R is used. R provides some useful libraries and commands that aim to the above mentioned enrichment calculations.

2.4.1 Transcription Factors (TFs) enrichment

As described at the introductory part (4.1 Transcription Factors), the transcription factors activity can be acquired by gene expression data. In order for this to happen, the programming language R is utilized. More specifically, its VIPER [15] (Virtual Inference of Protein-activity by Enriched Regulon analysis) algorithm is used. This algorithm allows computational inference of protein activity, on an individual sample basis, from gene expression data. It uses the expression of genes that are most directly regulated by a given protein, such as the targets of a transcription factor (TF), as an accurate reporter of its activity.

The R library that contains this algorithm is called "*viper*" and the command used is also named *viper()*. The minimum set of parameters the algorithm needs are a gene expression matrix (or "*ExpressionSet*" object as named in R programming language) and an appropriate regulatory network. Having these, a VIPER analysis can be carried out with the *viper()* function. This analysis effectively transforms a gene expression matrix to a regulatory protein (TF) activity matrix. The simplest implementation of VIPER is based on single-sample gene expression signatures obtained by scaling the probes or genes – subtracting the mean and dividing by the standard deviation of each row. In this way, the TF enrichment is calculated.

2.4.2 Signaling pathways enrichment

Once again according to the theoretical background provided in the introduction (4.2 Signaling Pathways), it is possible to calculate a signaling pathway activity based on the gene expression values. The enrichment takes place in R programming language, by utilizing the library "*pathfindR*" [16]. This library provides a tool for pathway enrichment analysis utilizing active subnetworks. It identifies gene sets that form active subnetworks in a protein-protein interaction network using a list of genes provided by the user. It then performs pathway enrichment analyses on the identified gene sets. The user-defined genes are the results of the gene enrichment after the normalization that is mentioned above (3. Data background correction and normalization). The function used is *run_pathfindR()* with the only necessary input parameter being the gene expression matrix. Its output is the signaling pathways enrichment. The function's workflow is presented below.

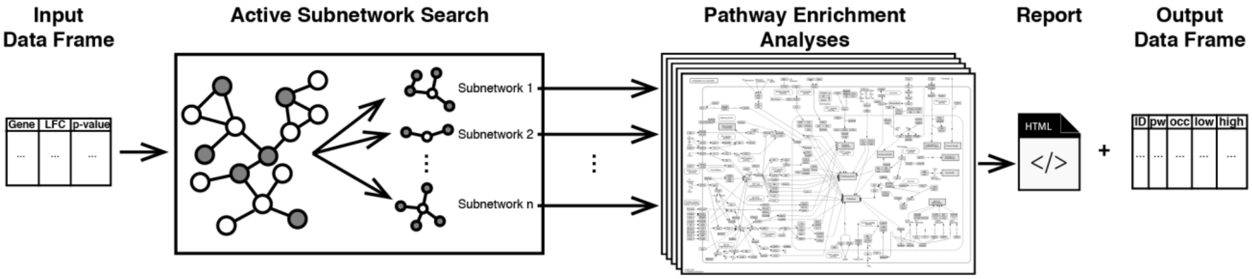


Figure 11 Signaling pathway enrichment workflow

2.4.3 GO terms enrichment

Gene Ontology (GO) terms enrichment is also possible to be conducted based on the gene expression results. For this enrichment, the R package "*topGO*" [17] is utilized. This package is designed to facilitate semi-automated enrichment analysis for GO terms. The process input is again the normalized gene expression measurements, but this time along with the gene identifiers, the GO annotations and the GO hierarchical structure. These data make up an object named "*topGOdata*" in R, which is used as input to the enrichment command. This command is the *runTest()* and contains three inputs, the above mentioned topGOdata object, the algorithm that will be used and the statistic that is used to test the enrichment result. The final object's class is *topGOresult* and this is the GO terms enrichment result.

2.4.4 Protein networks enrichment

Finally, protein networks can be constructed by using gene expression as the starting point. This is achieved through the R package CARNIVAL (Causal Reasoning for Network Identification with integer VALue programming) [18]. This package provides a framework to perform causal reasoning to infer a subset of signaling network from transcriptomics data. Transcription factors' (TFs) activities and pathway scores from gene expressions can be inferred with *DoRothEA* and *PROGENy* tools, respectively. TFs' activities and signed directed protein-protein interaction networks (+/-) drug targets and pathway scores are then used to derive a series of linear constraints to generate integer linear programming (ILP) problems. An ILP solver (CPLEX) is subsequently applied to identify the sub-network topology with minimized discrepancies on fitting error and model size. The result of this processing is a protein network enrichment analysis that has information not only about the interactions of the nodes, but also about each node's activity.

2.5 Distance calculations

All of the above biological ways to describe the activity that a perturbation induces in an organism provide some final data for each biological metric that is under investigation. These data have to be examined in order to get the results regarding their biological similarity. One way to define the similarity between two objects is to compute their pairwise distance. In case their distance is low, they are similar, and if not, they are obviously dissimilar. The definition of whether a distance is low or high depends on the threshold that is selected. The threshold in its turn depends on the range that the distance values belong to.

Two distance calculation methods are applied to the data constructed. The first one is based on the Gene Set Enrichment Analysis (GSEA) [19] and is part of the "*GeneExpressionSignature*" package, which is available in R. The function used to calculate these distances is the *scoreGSEA()* and will be further discussed in the next paragraph. The second method is derived by adding prior knowledge information to the *scoreGSEA()* function by alternating some parts of the algorithm, as well as its inputs. This approach is based on a method proposed by Junghwan Kim et al. which uses representation learning in signed directed networks. The approach is called SIDE and that's why the second distance calculation method is called *SIDE_scoreGSEA* and will also be presented in the following parts.

Data Preprocessing

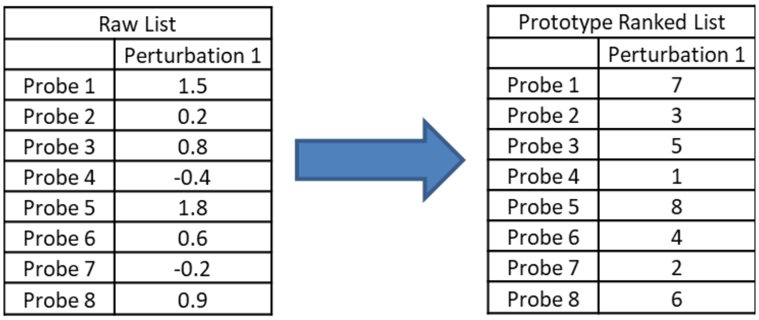
For the application of both of the distance calculation methods, the data have to be in a specific format. This format is Prototype Ranked List (PRL). The raw data that are constructed with the enrichments are in a form of matrices. The number of matrices' columns is the same when a specific array is examined. That's due to the fact that the columns refer to the perturbations that are applied to the array. So, no matter whether one looks the TFs for example or the gene expression, the number of columns will be the same. On the other hand, the number of matrices' rows varies even in the same array from the one biological metric to another. This variation occurs because the number of rows connected with the number of the features that are measured each time. For instance, the gene expression results might have around 20,000 rows, since this is the number of probes that are measured, while the TFs have around 180 rows, since this is the number of TFs that can be calculated. The form of the raw data for each array is as below, regarding the gene expression for example, where $E_{i,j}$ is the expression measured at the probe i for the perturbation j.

	Perturbation 1	•••	Perturbation (m)
Probe 1	$E_{1,1}$		$E_{1,m}$
Probe 2	$E_{2,1}$		$E_{2,m}$
•••			
Probe (n-1)	$E_{n-1,1}$		$E_{n-1,m}$
Probe (n)	$E_{n,1}$		$E_{n,m}$

Table 1 Gene expression raw data

According to these, the expression results can be considered as a group of lists that each one of them contains the expression for a single perturbation. So, the upper example is a group of m lists, with each one of them having n elements, which correspond to the number of probes.

Based on this approach, in order for the expression data to be used at the GSEA and provide the distances needed, the lists that refer to the perturbations have to be transformed to PRLs. As already mentioned, at their raw format the lists contain the expression levels that each perturbation exhibits. A PRL will contain the ranks of the expression values of the raw data, in terms of each perturbation list. A simplified example of how a PRL of a list of numbers should seem is presented below.



Raw List	
	Perturbation 1
Probe 1	1.5
Probe 2	0.2
Probe 3	0.8
Probe 4	-0.4
Probe 5	1.8
Probe 6	0.6
Probe 7	-0.2
Probe 8	0.9

Prototype Ranked List	
	Perturbation 1
Probe 1	7
Probe 2	3
Probe 3	5
Probe 4	1
Probe 5	8
Probe 6	4
Probe 7	2
Probe 8	6

Figure 12 Transformation of a raw data list to a PRL

As the example shows, each value is compared to the rest that belong to a specific perturbation and that is how its rank is calculated. The lowest value has rank 1 and the rest of them follow in in increasing order up to the highest value, whose rank should be same with the number of rows that the list has. This process is carried out for each one of the perturbations of the data set imported to the algorithm so that all of the perturbation lists are in PRL format.

GSEA

GSEA considers experiments with genome-wide expression profiles from samples belonging to two classes, labeled 1 or 2. These two classes are the two perturbations in the present approach whose pairwise distance is computed. Genes are ranked based on the correlation between their expression and the class distinction by using any suitable metric. Given an a priori defined set of genes S, the goal of GSEA is to determine whether the members of S are randomly distributed throughout another gene set L or primarily found at the top or bottom.

Up to now, GSEA was only used in terms of gene expression data. An innovative approach is proposed, according to which GSEA is applied to TF activity data, to signaling pathways activity data and to GO terms activity data as well. The principles of GSEA are suitable for these types of data in addition to the gene expression. This fact enables its use on these data and the investigation of the results that it is capable of providing.

2.5.1 scoreGSEA

The first distance calculation is carried out by using the *scoreGSEA()* function. This function computes the pairwise Enrichment Score (ES) of two perturbation PRLs and does so for all the RPLs included in the input dataset. The output of this function is a square matrix, which contains the pairwise distances of all the perturbations of the dataset and its number of rows, as well as columns is the same with the number of perturbations present in the dataset. Its diagonal is zero and the matrix is symmetric in regard to its diagonal. The method of this matrix construction is explained in the next paragraphs.

The algorithm, as mentioned above, calculates a pairwise enrichment score of two perturbations in PRL format. The list elements are put in descending order and in this way the top and bottom elements are the extremes. The algorithm in fact checks whether the extremes (top or bottom) of the one PRL are similar to the corresponding ones of the other PRL.

This ES reflects the degree to which a set S is over-represented at the extremes (top or bottom) of the entire ranked list L. The score is calculated by walking down the list L, increasing a running-sum statistic when a gene in S is encountered and decreasing it when genes not in S are encountered. The magnitude of the increment depends on the correlation of the gene with the phenotype. The enrichment score is the maximum deviation from zero encountered in the random walk, which corresponds to a weighted Kolmogorov–Smirnov-like statistic.

The length of the extremes is defined each time by the user based on the total length of the list and the needs of each experiment.

In fact, the algorithm builds two running sums internally for each ES calculation. One of them is about the data that are located at the top extreme of the first list and checks where they are located, by walking down the second list, while the other running sum is for the bottom data and has the same purpose. In this way, two enrichment scores are calculated and finally, their average is used as the ES of the two PRLs under investigation. An overview of the running sum function and a schematic representation of the ES value are shown below. In the case of *scoreGSEA()* the ES shown in the figure is calculated once for the top elements and once for the bottom as already mentioned and its final value is their average. In this way both of the top and bottom elements are taken into consideration with the same weighting factor, which is apparently 50% for each, since their average is the result.

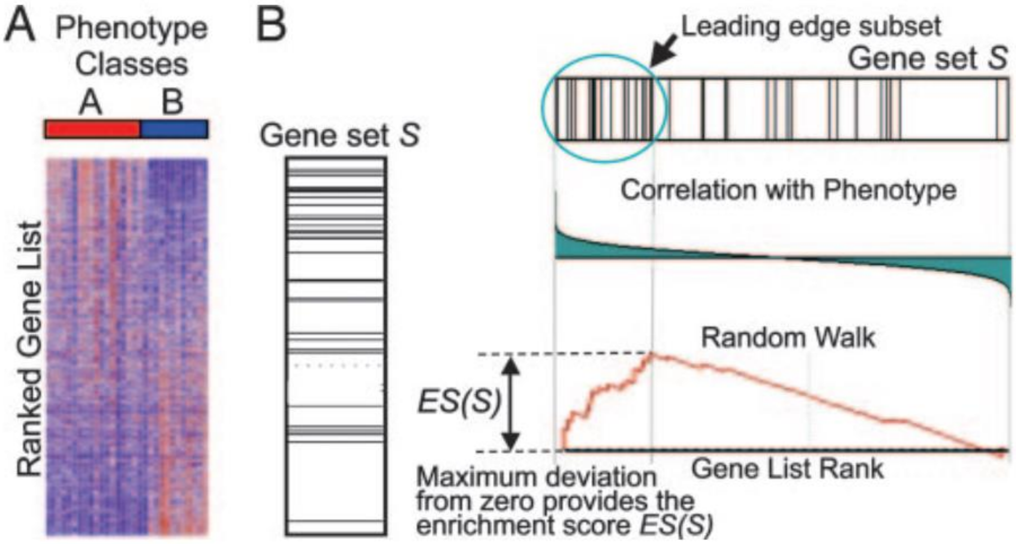


Figure 13 A GSEA overview consisting of (A) heatmap of an expression dataset correlation and (B) plot of the running sum and the ES calculation (Source: PNAS October 25, 2005 102 (43) 15545-15550)

As mentioned before, the ES is calculated pairwise for all the perturbation pairs included in the input dataset. While the algorithm progresses, a matrix is being built, which contains the pairwise ES. The mathematical aspect of the ES calculation is as follows.

Given the signature of n elements of a drug (d), the up-regulated genes will have a signature

$p = \{p_1, \dots, p_n\}$,

while the down-regulated genes will have a signature

$q = \{q_1, \dots, q_n\}$.

Then the distance between the drug (d) and a drug (x) will be the Inverse Total Enrichment Score (TES) of the drug (d) signature {p,q} with respect to the PRL of the drug (x) and is defined as follows.

$$TES_{d,x} = 1 - \frac{ES_x^p - ES_x^q}{2}$$

With ES, the enrichment score is annotated with respect to the PRL of drug (x). ES ranges in [-1,1] and quantifies how much a set of a genes is at the top or the bottom of a ranked list. The closer the ES is to 1, the closer is the set at the top, and the closer to -1, the closer is the set at the bottom. As a result, the TES ranges in [0,2].

The average Enrichment Score Distance between two drugs A and B that is used in this approach is defined as

$$D = \frac{TES_{A,B} + TES_{B,A}}{2}$$

Every time a pairwise ES distance (D) is calculated, it is put in a specific place of the matrix that corresponds to the annotation of the perturbations compared. Within every algorithm's loop, a new column and a new row are added to the matrix. The resulting matrix will be of dimensions [N, N], where N is the number of unique perturbations in the dataset. For example, in the case of an input with three perturbations (A, B and C), the ES distance result will be as follows, with the N being three.

	A	B	C
A	0	ES_{AB}	ES_{AC}
B	ES_{BA}	0	ES_{BC}
C	ES_{CA}	ES_{CB}	0

Table 2 Enrichment Score distance matrix (GSEA)

This ES distance matrix displays some characteristics. The elements that are symmetrical regarding the diagonal are equal. This means that ES_{AB} is the same with ES_{BA}, as well as ES_{AC} with ES_{CA} and ES_{BC} with ES_{CB}. This fact points out that no matter which one of the two PRLs that take part in one ES calculation is selected and tested against the other, the pairwise ES will be the same. Moreover, the diagonal elements are all zero, since these are the cases of an ES calculation between two identical PRLs (e.g. $ES_{AA}=0$). With ES_{IJ}, the ES distance between compound I and J is annotated.

SIDE

SIDE [20] is a general network embedding method, proposed by Junghwan Kim et al. This method represents both sign and direction of edges in the embedding space. SIDE formulates and optimizes likelihood over both direct and indirect signed connections. In fact, a vector representation of each node of a network is attempted to be learnt while encoding information on the network topology. Most existing network embedding methods only focus on modeling basic symmetric link structure, thus failing to exploit additional useful information in negative links and link directions. SIDE utilizes these characteristics to provide better encodings that take into account more network information. By leveraging both sign and direction in the network, SIDE achieves a remarkable accuracy in link sign prediction task.

A really appealing implementation of SIDE is the protein networks. These are signed and directed networks, just like the ones that SIDE is formulated to be used on.

These protein networks are called prior knowledge networks or protein-protein interaction (PPI) networks. This means that their signs and directions are defined by prior biological knowledge and that is why they are used as the foundation of many tests and algorithms, since they are established. A widely known database of such a network is OmniPath. These data contain a network with causal interactions.

The input to the SIDE algorithm is a file with all the interactions known for a network. So, OmniPath data can be utilized as input for SIDE. SIDE's result is as mentioned before, a representation in an embedding space of the network's nodes. SIDE also gives the user the ability to define how many embedding parameters are assigned to each node. These parameters are in form of numbers and their values are defined by the algorithm. They are basically the mapping of each node to the d-dimensional embedding space, where d is the number of the parameters.

In the presented case, 64 parameters are selected to represent the 2313 nodes that make up the input network in a 64-dimensional embedding space. The result is a matrix which consists of 2313 rows and 64 columns. Each row responds to the embedding vector of a protein, whose names are also available as the matrix row names.

At this point, the way that this embedding is utilized has to be clarified. As mentioned before, GSEA looks for the exact matches of the top and bottom elements in each perturbation regarding the others. Although this is an effective way of carrying out the analysis, it is possible that it could be improved. A way for improvement is the aggregation of more information in the algorithm.

It is believed and also proved for some examples that proteins with similar wiring in the PPI networks have similar functions. So, topology can provide an insight into the protein functionality. This is where SIDE comes in use. By comparing the proteins based on their embedding space distance, information about their topological distances can be extracted. Taking into consideration that topological similarity imposes functional similarity, the network's topology can be integrated in the GSEA algorithm to define a new approach. This is the *SIDE_scoreGSEA* that is explained below.

2.5.2 SIDE_scoreGSEA

The second method regarding the definition of a distance between biological characteristics is *SIDE_scoreGSEA*. It combines the PPI network embedding with *scoreGSEA*. At this point, it is important to highlight that this method can be applied in the case of gene expression or TF data. These are the two biological characteristics that are also present in the PPI networks.

The network embedding is not utilized at the form of the node attributes, but in fact only the nodes' distances are used by the algorithm. These distances are defined by applying the pairwise Euclidean distance metric at the nodes based on the 64-dimensional representation. By this calculation, a distance matrix that contains all the pairwise Euclidean distances is constructed. Its dimensions are [2313, 2313] as the number of the nodes in the network is 2313. Only part of these nodes is also present in the dataset under investigation (gene expression or TFs). In order for the Euclidean distance matrix to be useful, it has to contain each time the proteins that exist in both the PPI network and the dataset processed. To achieve this, the Euclidean distance matrix, that is constructed only once for the PPI network, is filtered through the dataset to include the desired proteins. Right after that, any protein of the dataset that is not in the PPI network is also removed, since no topological information exist for it.

Following this process, a distance matrix is available for the proteins that part of the gene expression or TF activity dataset and of the PPI network as well. This matrix is scaled for its values to belong to the range of [0, 1]. This normalization will be helpful in the next steps. The distance matrix has a structure same to the enrichment score's one and for three proteins (A, B and C) is the following.

	A	B	C
A	0	D_{AB}	D_{AC}
B	D_{BA}	0	D_{BC}
C	D_{CA}	D_{CB}	0

Table 3 Enrichment Score distance matrix (GSEA with SIDE)

The Euclidean distance is D_{ij}, where i and j refer to proteins, with values between zero and one. The matrix is symmetric which means that D_{ij}=D_{ji} and also, as expected, the elements of the diagonal are zero, since they represent the distance of a node in regard to itself.

The way that these distances are taken into account in the algorithm is simple and is based on a threshold that the user defines. Instead of only looking for the exact same proteins in the PRLs as explained in paragraph 5.A., the algorithm in this case considers that proteins with a pairwise Euclidean distance below a specific threshold can be viewed as similar and thus take part in the running sum. The concept behind this method is to check whether the topological information of the nodes can be proved helpful and produce better results when take into consideration. Moreover, through this approach, the belief that proteins which demonstrate similar topological wiring can induce similar functionalities is under investigation. A crucial parameter in the whole task is the threshold that will be defined. A high threshold might lead to the majority of the proteins though as similar, while a low one could lead to the absence of similar proteins. One way to decide this threshold is the trial and error. By comparing the results of various threshold values one can understand how it affects the outcome of the algorithm and then select its value.

To sum up, this version of the GSEA score has more parameters and can be modified to provide a wide overview of how the data respond to different similarity thresholds. Due to this fact, it is possible that underlying relationships might be uncovered while changing the parameters. On the other hand, whether this algorithm is superior to the simple GSEA score can only be confirmed through tests with various parameter values.

2.6 Machine learning

Machine learning is the scientific study of algorithms and models that are used to perform specific tasks without utilizing strictly defined instructions, but based on patterns and inference instead. Machine learning belongs to the field of artificial intelligence. The algorithms that machine learning consists of, aim to the establishment of mathematical models. These models are used to describe relationships between the data under investigation. Afterwards, the derived models are used to make predictions or decisions without the need of any explicit programming. At their final state, these algorithms use as input some data, on which statistical analysis is carried out and a prediction is made in the form of the algorithm's output.

The working principle of machine learning algorithms is the following. An initial dataset is used as input, called "training data". At the next step, the algorithm attempts to establish a mathematical model that corresponds to characteristics of these data. This attempt may contain regression models, neural networks or other mathematical and programming methods that aim to fit a model over a specific dataset. After the model is constructed, it is put under testing by using unknown data as input and evaluate the way it responds. These data are the “test data”. Based on the predictions of the test data the algorithm is accepted or modified for further testing.

Machine learning algorithms can be distinguished as supervised or unsupervised. Supervised algorithms require a data scientist or data analyst with machine learning skills to provide both input and desired output, in addition to furnishing feedback about the accuracy of predictions during algorithm training. Data scientists determine which variables, or features, the model should analyze and use to develop predictions. Once training is complete, the algorithm will apply what was learned to new data (test data). Unsupervised algorithms do not need to be trained with desired outcome data. Instead, they use an iterative approach called deep learning to review data and arrive at conclusions. Unsupervised learning algorithms, also called neural networks, are used for more complex processing tasks than supervised learning systems, including image recognition, speech-to-text and natural language generation. These algorithms work by combing through millions of examples of training data and automatically identifying often subtle correlations between many variables. Once trained, the algorithm can use its bank of associations to interpret new data. These algorithms require great amounts of training data and are thus feasible in the age of big data.

Neural Networks

Neural networks are a set of algorithms, inspired loosely by the human brain, that are designed to recognize patterns. They interpret data through a kind of machine perception, labeling or clustering the raw input they get. The patterns they recognize are numerical, contained in vectors, into which all real-world data (images, sound, text or time series) must be translated. This translation is almost always feasible and that is why neural networks are so popular, since their applications are not quite limited.

Neural networks are mainly used to cluster and classify sets of data. They can be perceived as a clustering and classification layer on top of the data that are under investigation. They are proved to be helpful at tasks that need unlabeled data grouping according to similarities among the example inputs, and data classification when a labeled dataset to train on is available. Neural networks can also extract features that are fed to other algorithms for clustering and classification. From this aspect, deep neural networks can be thought as components of larger machine-learning applications involving algorithms for reinforcement learning, classification and regression.

Neural networks are composed by several layers that consist of multiple nodes (or neurons). A node is a place where computation happens, resembling a neuron in the human brain, which fires when it encounters sufficient stimuli. A node could apply a function on the input data or do a simple calculation and then passes through the output data to the following nodes. Moreover, a node’s so-called activation function exists, to determine whether and to what extent that signal should progress further through the network to affect the ultimate outcome, for example an act of classification. If the signal passes through, the neuron has been “activated”.

A node layer is a row of those neuron-like switches (nodes) that turn on or off as the input is fed through the net. Each layer’s output is simultaneously the subsequent layer’s input, starting from an initial input layer receiving the network’s initial data. An example of a neural network with three layers is shown in the following figure. The input layer has three nodes, the second (hidden) one has four and the output layer, which is the third one, has two nodes.

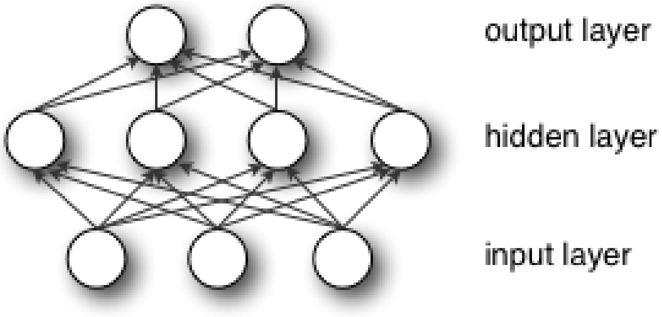


Figure 14 Simple Neural Network example

Multilayer Perceptrons (MLP)

The preceptor algorithm is an initial application of neural networks. It is a simple algorithm that is intended to perform binary classification; i.e. to predict whether an input belongs to a specific category of interest or not. In fact, a perceptron is a linear classifier. That means that it is an algorithm which classifies input by separating two categories with a straight line. Input is typically a feature vector x multiplied by weights w and added to a bias b. The function that gives the output y is $y = w \cdot x + b$. According to the neural network architecture that was previously presented, it is obvious that a perceptron, which is a single node, produces a single output based on several. The perceptron combines the input data with a set of coefficients, or weights, which either amplify or dampen that input, thereby assigning significance to inputs with regard to the task the algorithm is trying to learn (e.g. which input is most helpful is classifying data without error). These input-weight products are summed and then the sum is passed through a nonlinear activation function. This process can be described with the following mathematic expression.

$$y = \varphi(\sum_{i=1}^{n} w_i x_i + b) = \varphi(\mathbf{w}^T \mathbf{x} + \mathbf{b})$$

At this expression, **w** denotes the vector of weights, **x** is the vector of inputs, **b** is the bias and phi (ϕ) is the non-linear activation function. The activation function describes the relationship between the input and the output in a non-linear way. Its presence provides the model with the ability to be more flexible in describing arbitrary relations. Some popular activation functions are the sigmoid, the ReLU (Rectified Linear Unit) and the tanh.

A perceptron can be described with the following representation.

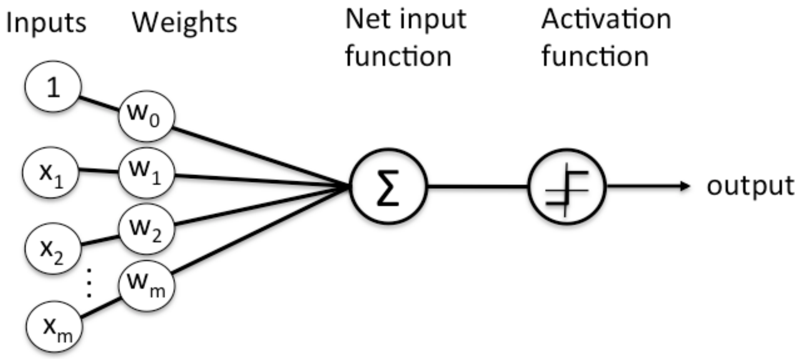


Figure 15 Perceptron representation

By pairing the model's adjustable weights with input features is how significance is assigned to those features with regard to how the neural network classifies and clusters input.

Multilayer perceptron (MLP) is, as its name suggests, a network with many perceptrons that make up several consecutive layers. The structure is similar to the three layer network that was presented above. MLP is the most typical neural network model. The layers that make it up are fully connected, which means that the output of each node is passed to all the nodes of the following layer and consequently each node's input is formed from all the previous layer nodes. The parameters of each node are independent of the rest, which makes each weighting that takes place at the nodes unique.

The performance assessment of the network is done by utilizing a loss function. The loss will be high in case that the predicted outcome does not correspond to the real one and low otherwise. Usually, the weights are adjusted according to the loss function reduction until a satisfactory loss value is achieved. A schematic example of a broad MLP network follows.

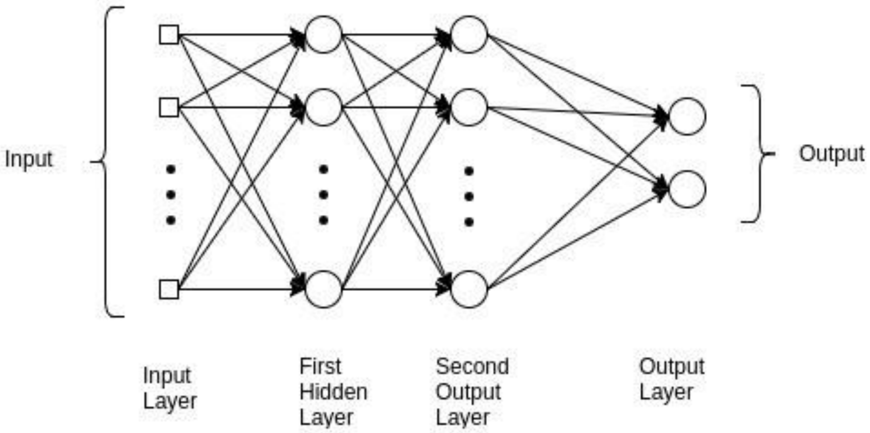


Figure 16 The Multilayer Perceptron (MLP) structure

The model

The model that is tested in the present project is called ReSimNet [21] and is a model that utilizes Siamese neural networks to predict drug response similarity. Siamese neural networks are a category of neural networks that utilizes two or more identical MLPs with shared weights between them that are simultaneously updated during the model's training. In the case of ReSimNet, two MLPs are present. The network's input is a pair of chemical compounds. They are represented by 2048-bit extended connectivity fingerprints (ECFPs). These are the initial input features. The ECFPs are then augmented with downstream trainable parameters until the network's function is determined. Let the x_a and x_b be the input ECFPs of two compounds. Their relationship with the outputs c_a and c_b is defined by the following functions.

$$c_a = w_2 \cdot f(w_1 x_a + b_1) + b_2$$

$$c_b = w_2 \cdot f(w_1 x_b + b_1) + b_2$$

Where w_1 and w_2 are trainable weights, b_1 and b_2 are trainable biases and f ( ) is an element-wise nonlinear activation function.

The model is trained to minimize the mean squared error between the true GO terms similarity (calculated as 1 – GO terms distance) (GO_{ab}) and the cosine similarity (s_{ab}) of the two outputs. Based on this, the loss and optimization function is defined as follows.

$$J(\boldsymbol{\Theta}) = \frac{1}{N} \sum\nolimits_{a,b} (s_{ab} - GO_{ab})^2$$

As N, the total number of training examples of input pairs is denoted, and as **Θ** the trainable parameters of the model are defined.

A schematic outline of the model consisting of how its training and validation are carried out, but also of its usage pipeline is presented at the following figure.

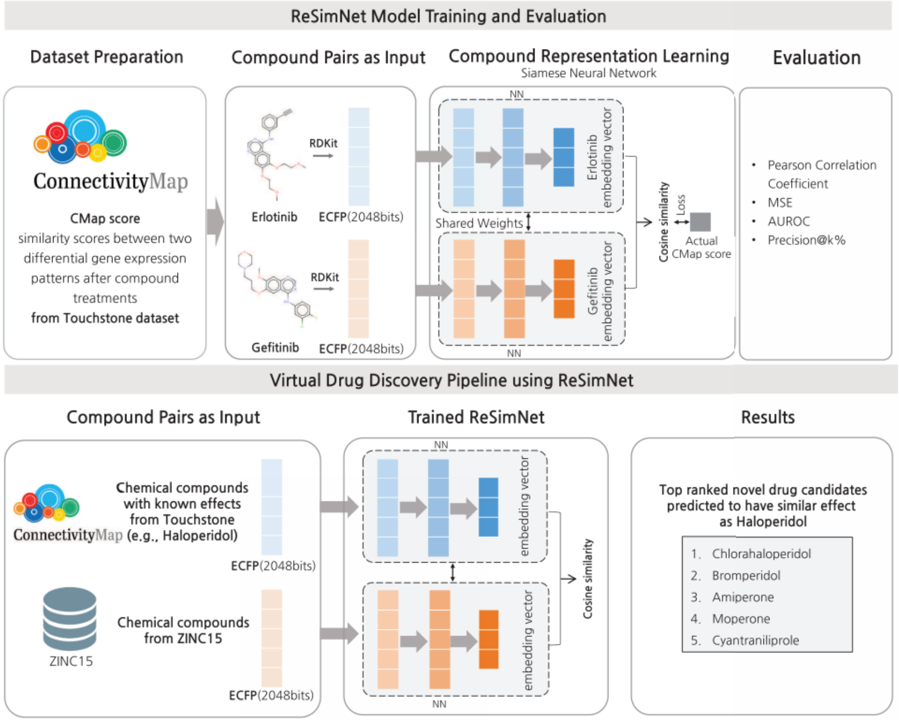


Figure 17 ReSimNet overview (Source: Bioinformatics, btz411)

In terms of structure, the Siamese neural networks consist of two layers each. The input is a vector with 2048 elements, as mentioned above. The first layer consists of 512 nodes, while the second one, which provides the output to be tested against the loss function, has 300 nodes. So, the vectors that are then compared have a 300 elements length.

Extended Connectivity Fingerprints

The extended connectivity fingerprints (ECFPs) of a compound is a circular topological fingerprint used for molecular characterization. In general, topological fingerprints were developed for substructure and similarity investigation. ECFPs are a recently developed fingerprint methodology explicitly designed to capture molecular features relevant to molecular activity. While not designed for substructure searching, they are well suited to tasks related to predicting and gaining insight into drug activity. They can be really fast calculated, they are not predefined and can represent an essentially infinite number of different molecular features (including stereochemical information). Moreover, their features represent the presence of particular substructures.

3 Results

3.1 Outline

Through the project, multiple methods and data were tested and various results were produced. As a starting point and motivation for this research, the correlation between drugs chemical structure and biological effect is investigated. Right after that, the establishment of new biological effect metrics is attempted and their use in terms of drug similarity is studied. All of the findings and results are then compared between one another to draw conclusions about their accuracy and quality. Finally, a machine learning model is implemented on the data in an attempt to learn drug representations and distances.

3.2 Extended motivation

The initial motivation of this project, as mentioned at the introduction part, is the question posed by F. Sirci et al. in their article named "Comparing structural and transcriptional drug networks reveals signatures of drug activity and toxicity in transcriptional responses" of whether a correlation between the chemical structure and the gene expression levels of a compound exists. The publication's findings are summed up at the plot presented in the Introduction part (see Figure 4).

As a first approach, the results of this article are reconstructed with more compounds taken into account. At the article, 784 CMap compounds are used to generate the plot, while at the recreation 883 compounds are present. This happens because the article's compounds are selected based on their ATC annotation existence, while at the recreation, the most compounds from the pool whose 3D structural distance is available are used without any filtering. The only that matters in the plot reconstruction data is whether these compounds' gene expression is available. The plot that occurs with these data is similar to the article's plot, as expected, and verifies the publication's findings and also the reconstruction methods followed. The reconstructed plot with the pairwise 3D structural distance plotted against the pairwise gene expression distance of 883 compounds is the following.

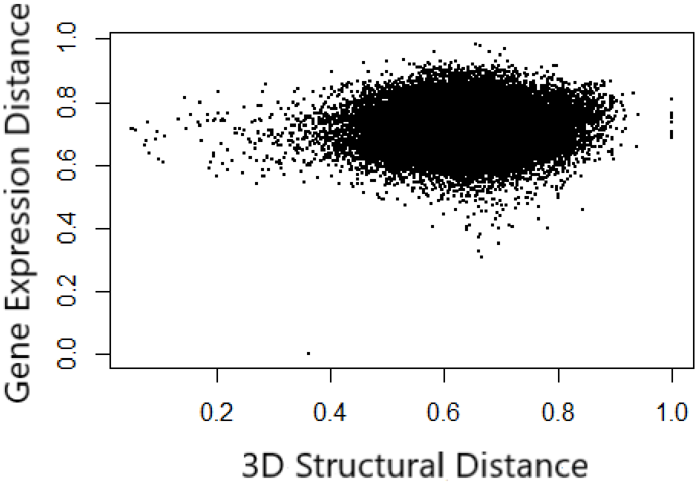


Figure 18 Gene expression distance vs 3D structural distance

Apparently, the metrics under investigation at this plot don't exhibit any correlation at all. Based on this fact, the question whether the utilization of different metrics can provide better results arises.

3.3 Biological - structural distances correlation

The first alternative that can be applied is changing the biological similarity metric. This can be done by using the pairwise transcription factors (TF) distances between the compounds under investigation. The TF distance is considered to contain less error than the gene expression one (since it is a higher level biological characteristic) and due to this fact is expected to provide more distinct pairwise distances. The GSEA score is calculated for all the pairs of compounds and the results are plotted at the y-axis, where before the gene expression distance could be found. The x-axis is kept as before, with the 3D structural distances of the compound pairs. The resulting plot is presented below.

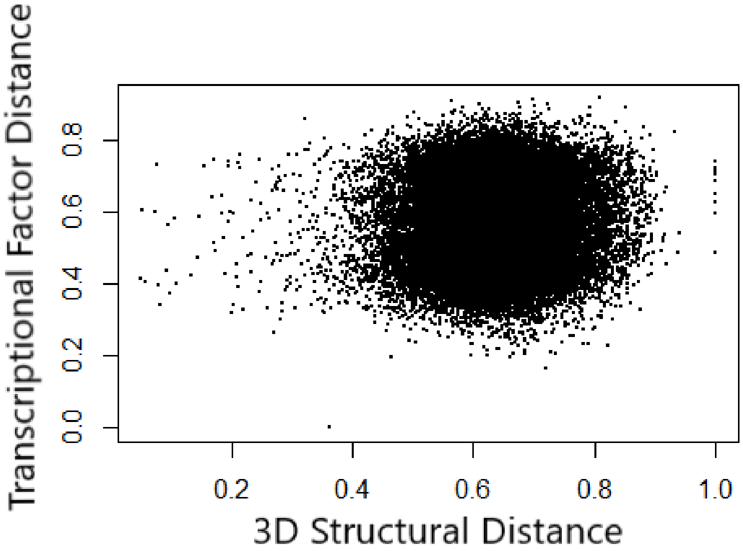


Figure 19 Transcriptional factor distance vs 3D structural distance

This plot is more scattered than the one with the gene expression distances, which means that the TFs as a biological effect metric provide a wider distance distribution. This happens due to the fact that the TFs are a higher level biological metric that contains less error than the gene expression. This plot enables a better clustering of the drug distances which is always depended on the distance thresholds that are applied. Even though the use of TF distances exhibits some improvements, still no correlation between the two metrics is revealed.

Up to this point, the results did not prove to be satisfying and no correlation is observed. Since the change in terms of the biological metric didn't make any significant difference at the plot, another approach is applied. Instead of using different biological metric (TFs where gene expression was used), a different chemical structure metric is used. This metric is the compounds' 2D chemical structure, according to which the structural distances are computed.

Both of the biological distance metrics (gene expression and TF activity) are available and are plotted against the 2D chemical structure distance of the compounds. The results are the following plots.

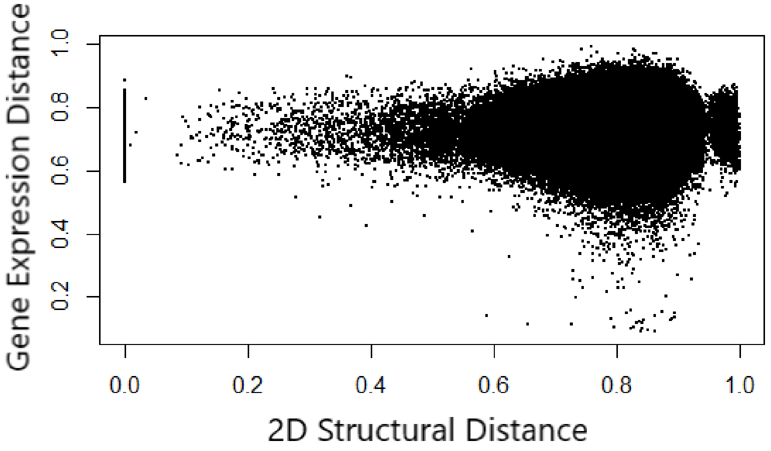


Figure 20 Gene expression distance vs 2D structural distance

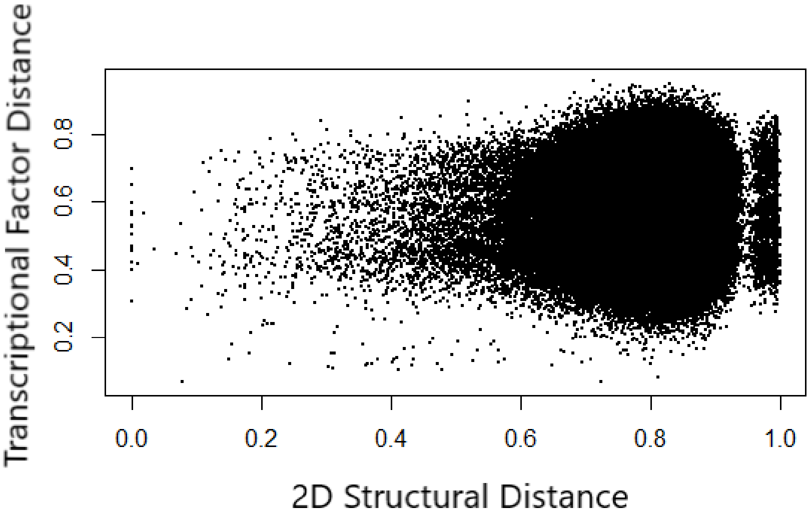


Figure 21 Transcriptional factor distance vs 2D structural distance

As posed by these plots, the morphology and the distribution of the pairwise distance points are considerably different from the 3D structural distance ones. This indicates a strong variation of the structural distances from the 3-dimensional level to the 2-dimensional level, since the biological pairwise distances are the same in the two levels respectively. The 3D to 2D distance differences can be result of the different algorithms that are followed in order to be extracted at each dimensional level. Both of the methods (2D and 3D) have the SMILES as the initial input and are based on them but still exhibit the upper mentioned variation. Moreover, the plots that refer to the 2D distances seem to distinguish a number of pairwise structural distances near to 1, which is the highest value that can be achieved. This means that they are totally different in chemical structure and their grouping is a positive characteristic of the 2D distance algorithm. On the other hand, this might be a flaw of the algorithm, because almost no values with structural distance around 0.9 exist. Overall, the results don't present any correlation even at the 2D level. The structural distance values are wider, which is a positive element but in terms of correlation, no new insight is available. The algorithm that computes the 2D structural distances can't be modified and thus these are the only results that can be achieved.

At this point, a comparison between the 2D and 3D structural distances would make sense in order to provide a final and complete evaluation of the whole process. This comparison can help towards the clarification of the differences that are exhibited between the two dimensions in terms of the chemical structure distances. So, the 3D chemical structure distances are plotted against the 2D chemical structure distances and the result is the following.

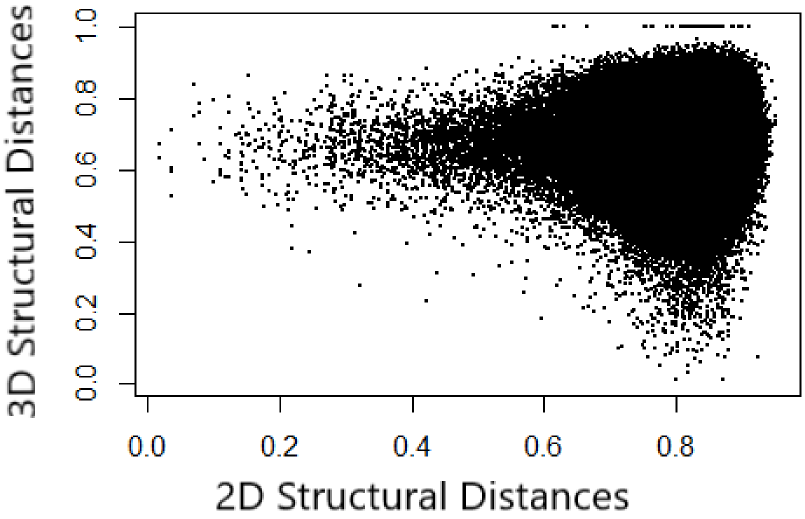


Figure 22 3D structural distances vs 2D structural distances

The two structural distances show many differences, as expected from the previous results. Despite the differences, the points that correspond to distances higher than 0.6 according to both of the metrics seem to form a dense region on the plot. These points correspond to pairs that can be considered as significantly different in terms of structure. This formation indicates that the majority of the chemically different pairs are considered as different in both of the approaches (2D and 3D). Apart from this dense area, there are also points at the upper left and lower right quarters. These are pairs that are considered chemically similar with the one approach and different with the other. This is where the variation between the two approaches lies and that is the reason why the results of structural distances compared to the biological distances vary, depending on the structural level used (2D or 3D). These points, along with the plot's total formation, indicate that no correlation exists between the two methods of chemical structure distance definition. Each one seems to provide different results that are not related with the others, even though the input data are the same (compound SMILES). The pipelines used to create the distances are the cause for these results and the lack of correlation between 2D and 3D distances, while unexpected, can be explained by the distinct algorithms that each approach (2D and 3D) utilizes. These algorithms are not related with respect to each approach and subsequently, their calculated outputs differ in the ways shown above.

3.4 New biosimilarity metrics definition and assessment

As the upper results indicate, no relationship between the chemical structure and the biological effect (gene expression or TFs) can be extracted. Moreover, the attempt to compare and correlate the chemical structure with the biological effect of the compounds seems to be a hard task that might not have any solution, at least while using the current available data.

On this ground, it makes sense to look for other methods that can be utilized for the compound similarity investigation. Compounds can be viewed through two aspects, the chemical and the biological.

Regarding the chemical characteristics of the compounds, many limitations are posed against them, since the pipelines to extract and utilize these characteristics are complicated and need many algorithms. This complexity makes the results of the chemical characteristics hard to modify and reproduce. Moreover, one needs advanced chemistry and molecular biology knowledge to assess the results of these algorithms. That's why this field is not suitable for extensive computational research under the current circumstances.

The second available field that can be investigated is the biological. In this case, many algorithms exist through which initial gene expression data can be processed and provide more data, based on biological prior knowledge (see Introduction paragraph 4, "Establishing new metrics based on prior knowledge"). The whole process can be carried out by a single programming language that comes with a free software environment as well. This language is R, which through its libraries and packages makes the research on the biological effects of compounds accessible to anyone. The only prerequisite is the initial gene expression data that are calculated from experiments with various compounds and are available in public databases. So, the data needed are available to anyone, as well as the means to process them and reach some results with computational methods. Due to this fact, the biological aspect of the compounds' similarity is the field to which further research regarding the compound similarity can be applied.

The approach consists of the construction of multiple biological level data (see Materials and Methods paragraph 4, "Enrichment calculations") that are compared with each other based on multiple methods (see Materials and Methods paragraph 5, "Distance calculations") and their outcomes are the results of it.

The biological levels that are investigated are the Transcription Factors (TFs), the signaling pathways, the Gene Ontology (GO) terms and the protein networks. They are all constructed by combining gene expression data with prior biological systems knowledge. Then, the distances of the data within each level are calculated using two methods.

The validation of the algorithms' results is based at the comparison of the distances that are calculated for the replicates of some compounds with the distances that are calculated between unique non-replicated compounds. At each dataset, there can be compounds that exhibit one or more replicates. The replicates are the results of a compound that is present in the experiment more than one times and they are expected to show low pairwise distances, since they correspond to the same compound.

The comparison of the distances is carried out through their density distributions that are presented throughout this paragraph.

The tests are applied to five cell lines present in the latest CMap version. These are the A375, the HEPG2, the MCF7, the NPC and the PC3 cell lines. The results of each cell line are independent, in order for the errors and variances to be minimized. All the data used are of the highest quality available, according to the information of the data. That means that only the high quality (quality 1) data are filtered and used from the datasets, in an attempt to come up with reliable results that contain the least possible uncertainty. The quality of the data is defined by the organization that conducts the experiments and is not further investigated.

3.4.1 Parameter investigation

All of the distance calculation methods are applied to each cell line in order to compare the results that each one provides. These results may vary according to the parameters of the distance calculation algorithms. These parameters are the signature length that is selected for the score GSEA algorithm and the topological similarity threshold for the custom score GSEA algorithm that utilizes SIDE. In order to estimate the effect of each parameter's value, tests have to be carried out and assessed according to the results. The tests are carried out on one cell line for brevity reasons. To select a cell line, the ratio of replicated compound samples to the total unique compounds of each cell line is calculated. The results are the following.

	Unique	Replicated	Ratio
A375	241	3	0.012
HEPG2	106	4	0.038
MCF7	503	14	0.028
NPC	249	13	0.052
PC3	416	6	0.014

Table 4 Cell lines information

From these ratios, the biggest one is selected in order to have sufficient replicated compounds in terms of the total ones in the cell line that the algorithms are tested. According to the results, the selected cell line is the NPC, with a ratio of replicates to unique compounds around 5%. So, the tests that will define the parameters of the distance algorithms are carried out on the NPC cell line. The total sample number of NPC is not the highest between the cell lines, but this is not a problem, since the number of total samples is satisfying and in fact it enables the algorithms to perform faster, while testing many cases.

The performance of the distance algorithms is defined by how well they manage to distinguish the replicated samples distances from the unique sample distances. The density plots that follow exhibit how each algorithm performs.

3.4.1.1 GSEA score on Gene Expression

At first, the score that is applied to the gene expression data is under investigation, since this is the initial data available and they are considered as the one that contain the most noise as well. The GSEA score algorithm is applied four times, with the signature length being 10, 30, 50 and 70. This means that the each time the top and bottom genes are 10, 30, 50 or 70, according to the length selected. All the genes that are available for processing are 978 for each perturbation. So, the signature length is 1%, 3%, 5% and 7% of the total genes for the upper selections respectively. The results are in the form of density plots, as mentioned before and are the following ones. The black density lines refer to the replicated samples pairwise distances, while the red ones to the unique samples pairwise distances, annotated as normal.

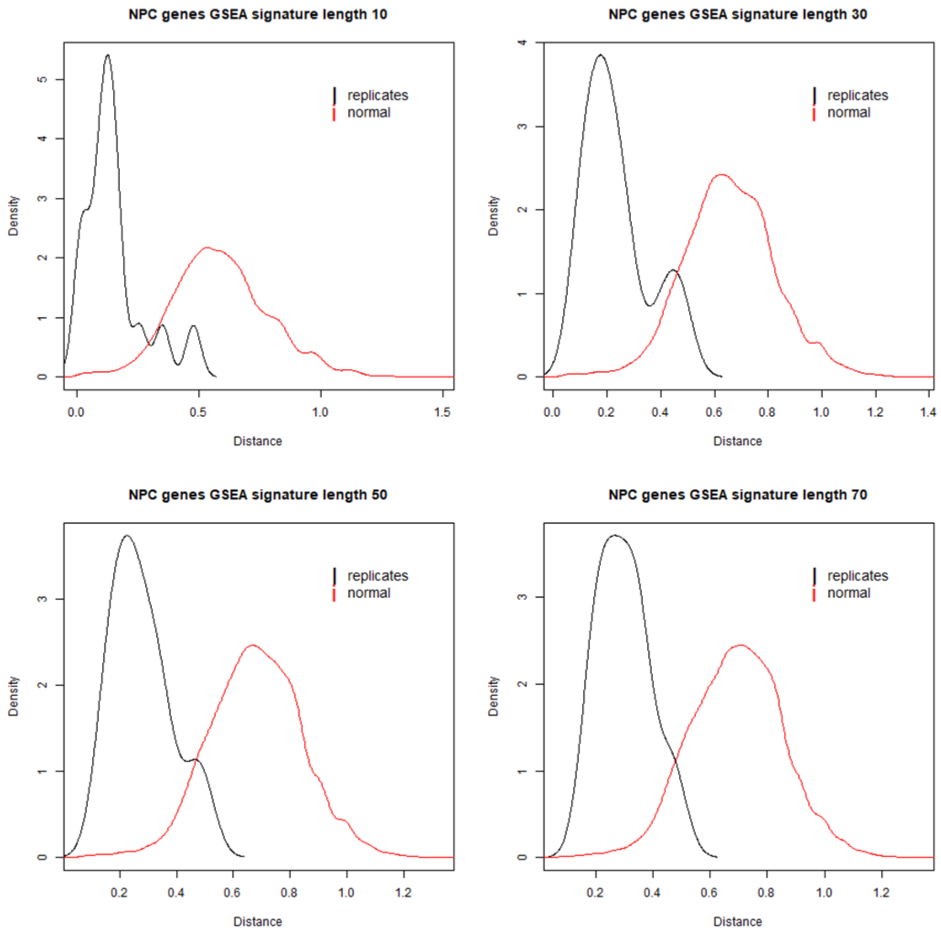


Figure 23 NPC cell line genes GSEA distances density plots

According to the results, the best separation between the replicated sample pairwise distances and the unique samples pairwise distances is achieved for a signature length of 10 genes.

3.4.1.2 GSEA score on Transcription Factors (TFs)

For the TFs GSEA score, the signature length is tested for the values 5, 10, 20 and 30. The total TFs available from the data are 175 for each perturbation. So, the signature length is 3%, 6%, 12% and 18% respectively. By applying the score GSEA at the TF data, the next density plots are constructed.

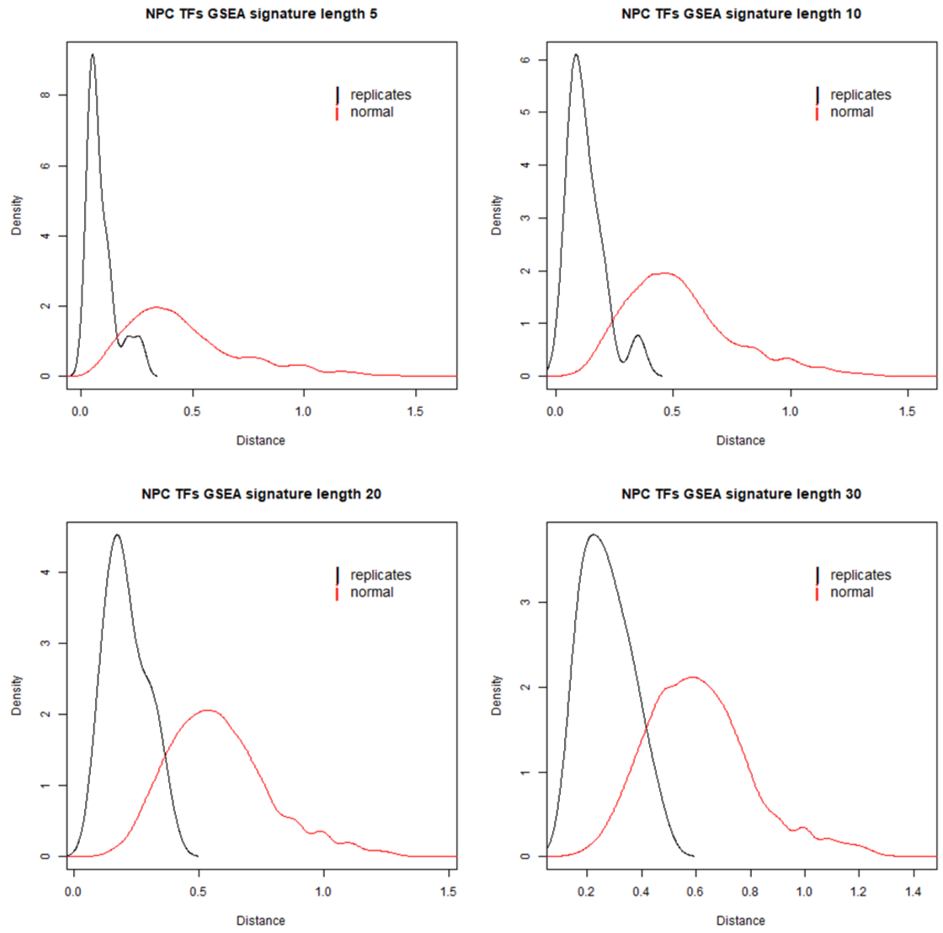


Figure 24 NPC cell line TFs GSEA distances density plots

The results indicate that at this case as well, a signature length of 5 TFs is capable of providing the best separation of the two distance categories.

3.4.1.3 GSEA score on Signaling Pathways

The GSEA applied at the signaling pathways is tested for signature length of 5, 10, 20 and 30 pathways. The total pathways present for each perturbation are 125, so the signature length is 4%, 8%, 16% and 32% respectively. The density plots are the following.

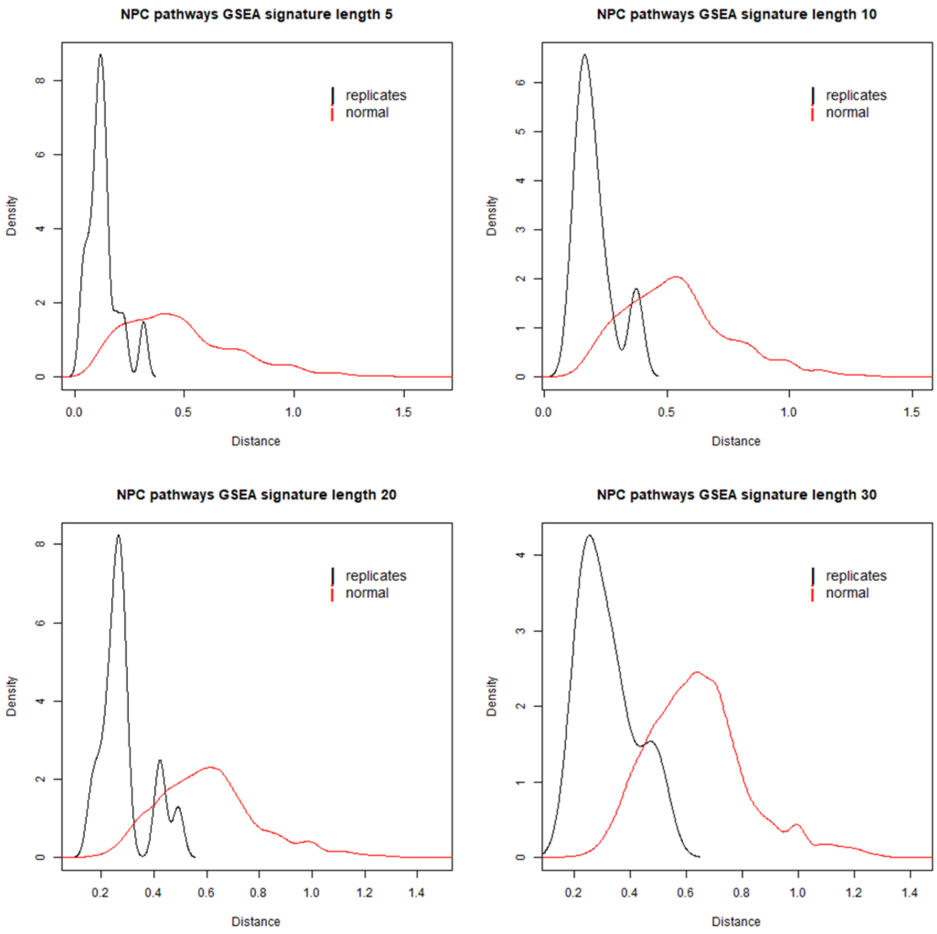


Figure 25 NPC cell line pathways GSEA distances density plots

Once more, the best result seems to be provided by a signature length value of 5 pathways.

3.4.1.4 GSEA score on Gene Ontology (GO) Terms

Regarding the GO terms, for each perturbation there are 2,105 GO terms available. The GSEA score signature length is set at 25, 50, 100 and 150 GO terms, which are the 1%, 2%, 5% and 7% of the total GO terms of each perturbation. The pairwise density plots are presented below.

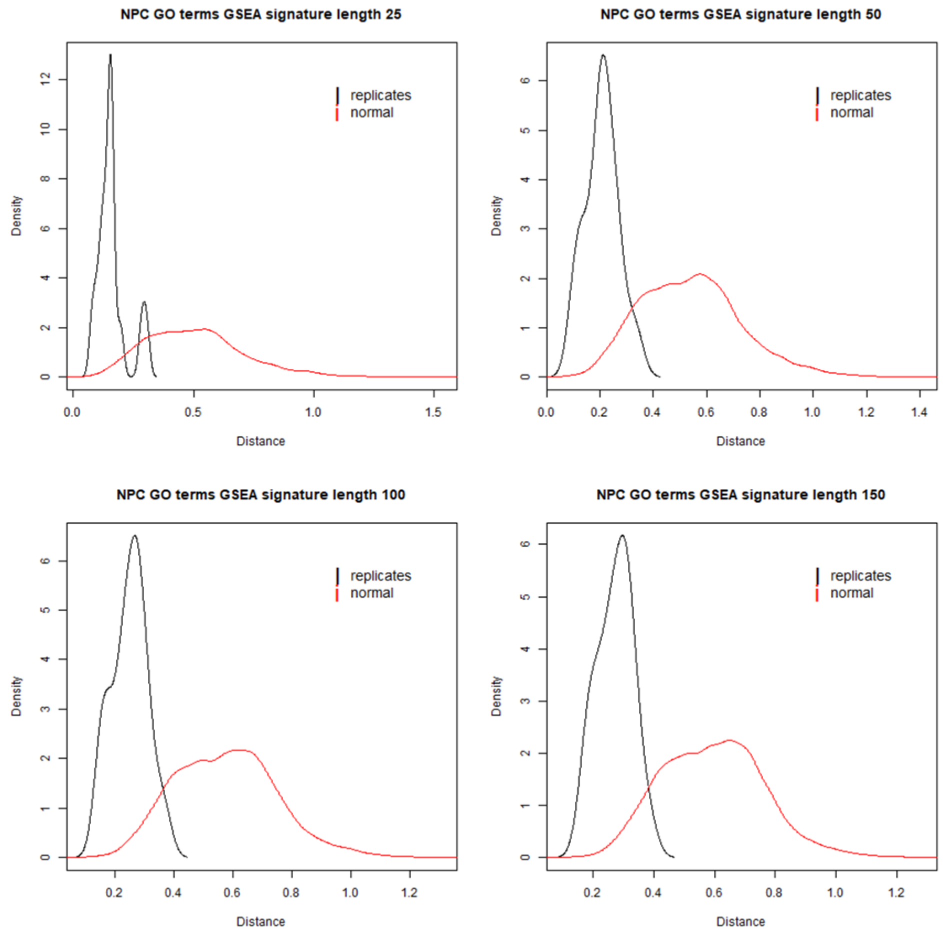
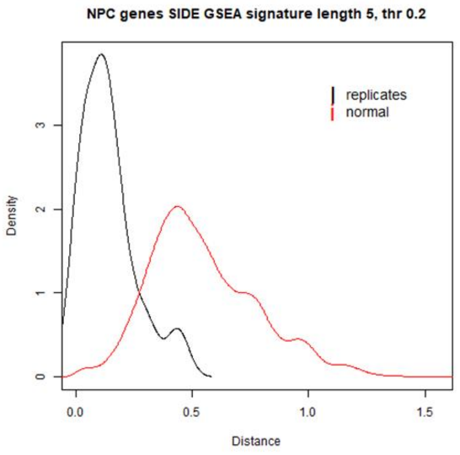


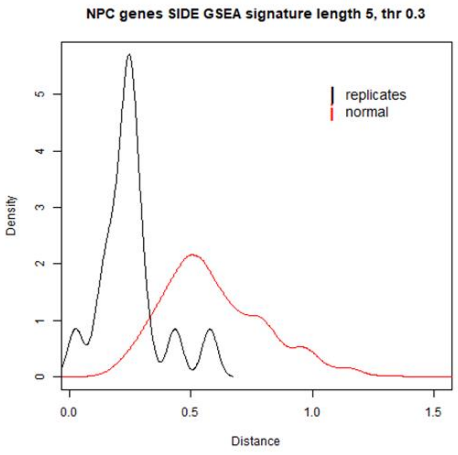
Figure 26 NPC cell line GO terms GSEA distances density plots

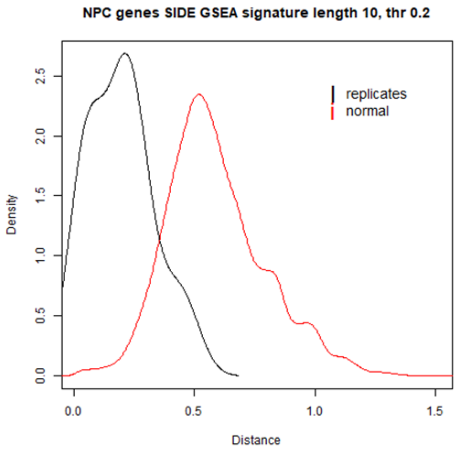
The most satisfying signature length from the ones mentioned above for the GO terms GSEA score is 25.

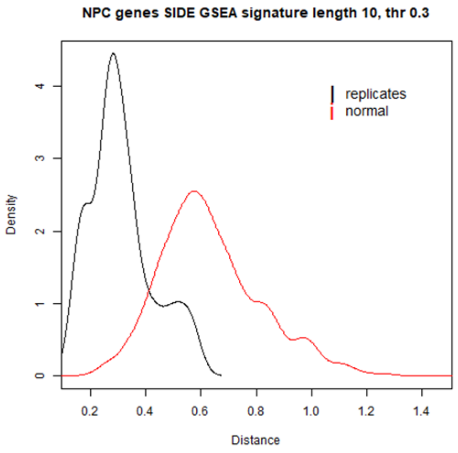
3.4.1.5 SIDE GSEA score on Gene Expression

The next distance algorithm whose parameters have to be selected is the SIDE GSEA score algorithm. Due to the fact that this approach utilizes a PPI network, the genes that are used from the gene expression data have to be present also in the PPI network. This filtering makes only 331 genes out of the 978 present in the initial data available for SIDE distance calculations. Although the number of genes processed with SIDE is almost the 34% of the available ones, the use of the PPI network relationships could provide a better insight into the pairwise distances. The SIDE GSEA score algorithm, which was explained in Materials and Methods, uses two parameters in order to decide the pairwise distance of the compounds. The first one is same as the previous mentioned in this section and refers to the signature length that the algorithm uses. The second one is about the distance threshold that is applied to the PPI networks distances. This threshold is used to decide the pairwise distance, regarding PPI network, below from which the compound pairs are considered to be topologically similar. The distances in the PPI network are normalized to take values from the space [0, 1]. The PPI network thresholds that are tested are 0.2 and 0.3. Regarding the signature length, the values tested are 5, 10 and 20. These correspond to the 1.5%, 3% and the 6% of the available genes for the SIDE GSEA score algorithm. The resulting density plots are presented below.









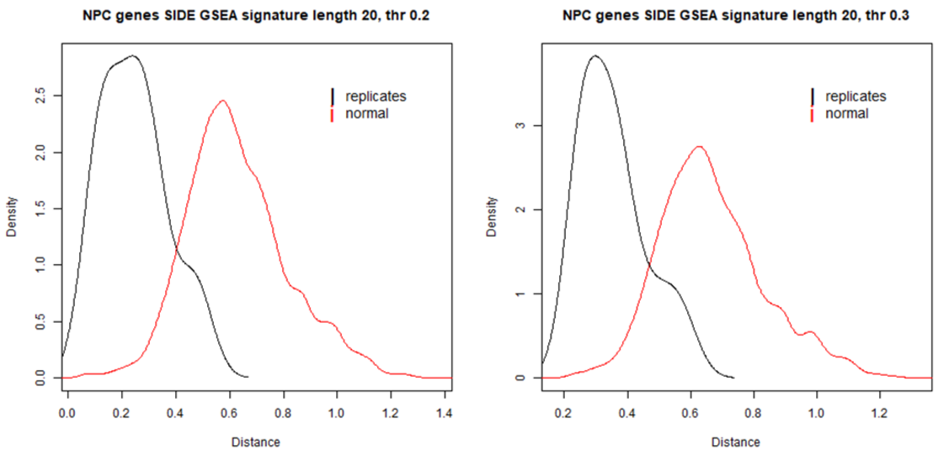
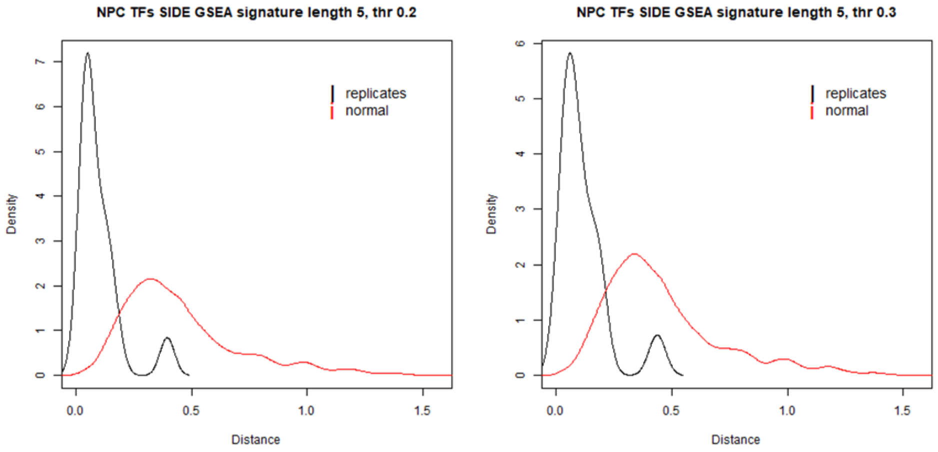


Figure 27 NPC cell line genes SIDE GSEA distances density plots

Based on the plots, the best selection of parameters is a signature length of 5 genes with a topological distance threshold of 0.2.

3.4.1.6 SIDE GSEA score on Transcription Factors

In terms of SIDE GSEA score applied on TFs activity data, the same filtering as above happens. The TFs present in the PPI network is 112. The topological distance thresholds tested are 0.2 and 0.3 again. Also, the signature lengths are 5, 10 and 20 as well. These correspond to the 4%, 9% and 18% of the total TFs available for each perturbation. The following plots correspond to the TFs SIDE GSEA score tests.



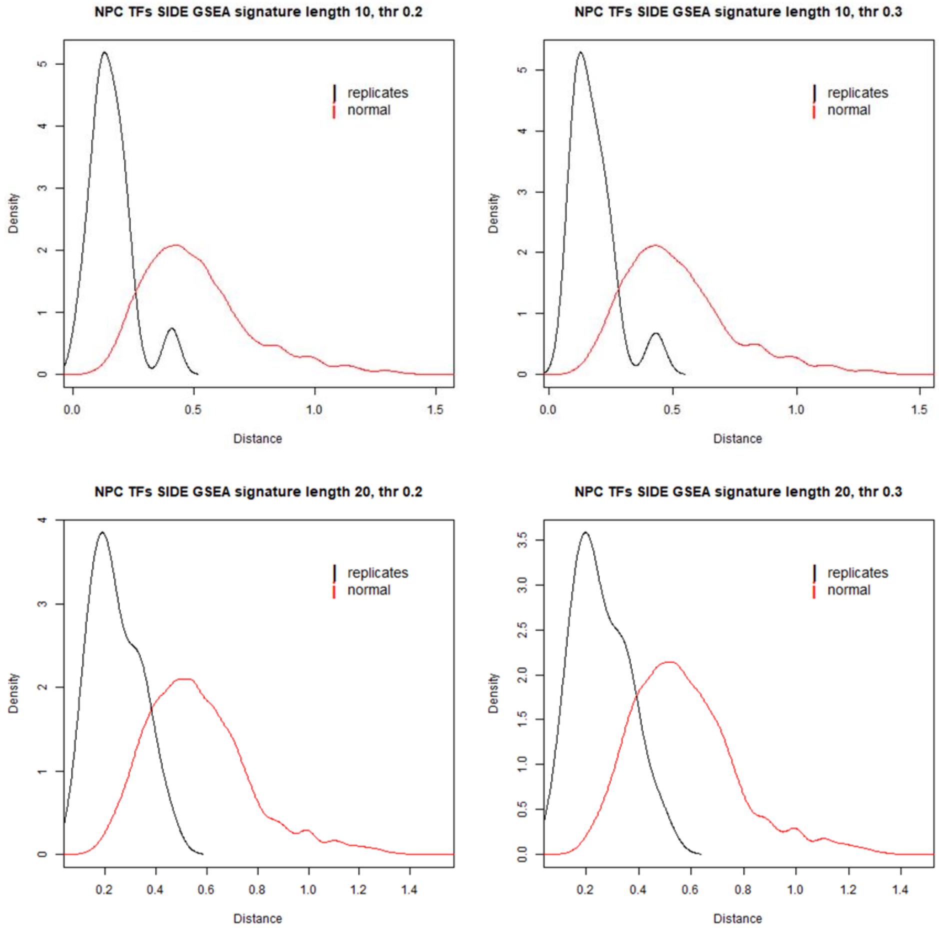


Figure 28 NPC cell line TFs SIDE GSEA distances density plots

The plots indicate that the case of a signature length of 5 TFs and a topological similarity threshold of 0.2 provides the best separation of the two sets of distance data.

3.4.1.7 Parameter investigation results

From the metrics investigated, the GO terms GSEA score exhibits the best separation of the pairwise distances, since it manages to achieve a density of 12 values at the lowest distance.

According to the upper results, the distance algorithms provide the best results when the signature length is the lowest at each test. Moreover, it is obvious that as the signature length increases, the separation of the replicated and the unique compound distances gets worse. The same effect is exhibited by the topological similarity threshold increase at SIDE GSEA.

3.4.2 Metrics comparison across cell lines

The parameters that provided the best plots at the previous paragraph are used to test all six algorithms regarding the different metrics and approaches across each cell line and get a general overview of their performance. The reason why they are applied to each cell line separately for the assessment is that even the same metrics can exhibit different results when applied to another cell line. This variation stems from the fact that each cell line contains different tissue (and cells by extension) but more importantly a different disease. So, even the same perturbations behave in different ways according to the cell line they are tested on.

As an example of this variation, the pairwise distances of the same compound pairs, based on gene expression, at the cell line VCAP (human prostate cancer cells) and the cell line MCF7 (human breast cancer cells) are exhibited.

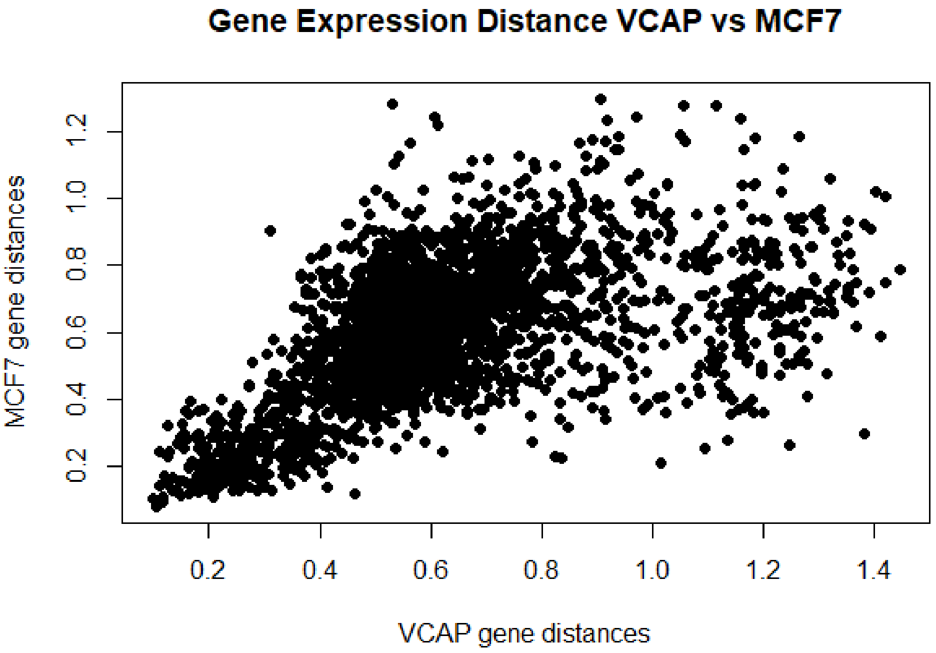
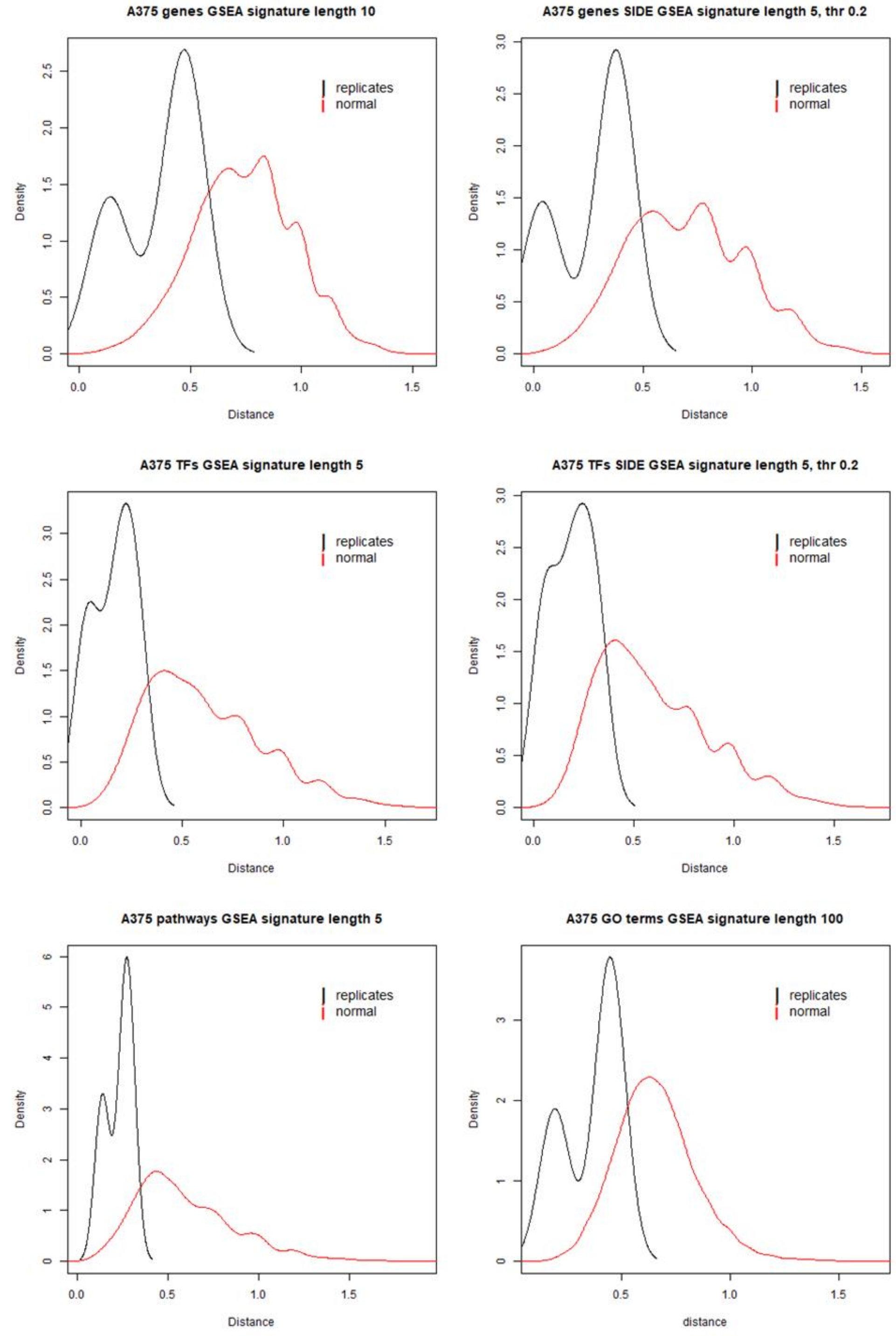


Figure 29 MCF7 gene expression distances vs VCAP gene expression distances

As the plot indicates, the distances vary significantly between the two cell lines, despite the fact that they refer to the same pairwise compound distances. Based on this fact, the pairwise compound distance algorithms are tested on every single cell line and the results are presented through the next plots in the form of density plots, as it was explained earlier in this section.

3.4.2.1 A375 cell line



A375 genes GSEA signature length 10

A375 genes SIDE GSEA signature length 5, thr 0.2

A375 TFs GSEA signature length 5

A375 TFs SIDE GSEA signature length 5, thr 0.2

A375 pathways GSEA signature length 5

A375 GO terms GSEA signature length 100

Figure 30 A375 cell line all distance calculation metrics density plots

3.4.2.2 HEPG2 cell line

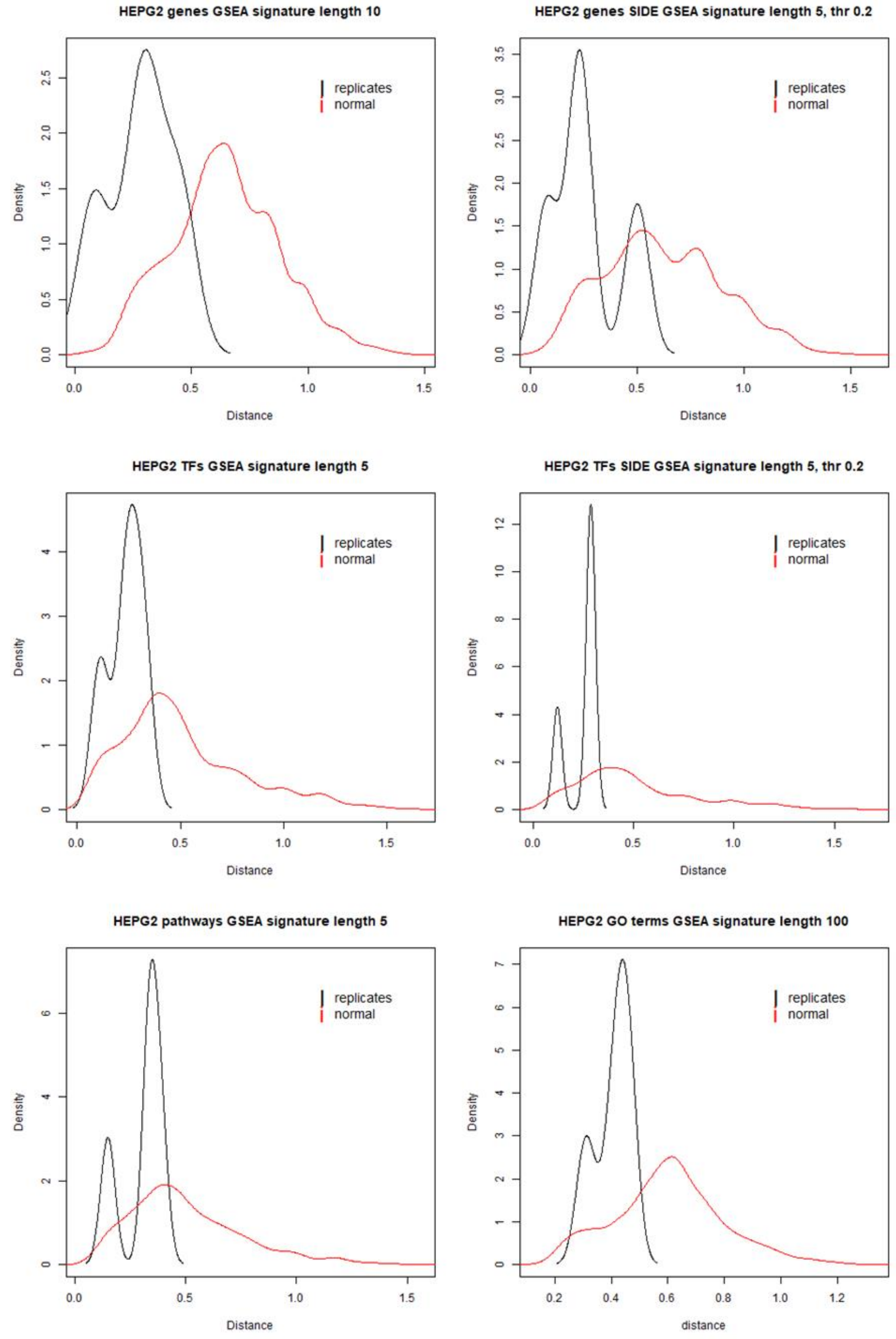


Figure 31 HEPG2 cell line all distance calculation metrics density plots

3.4.2.3 MCF7 cell line

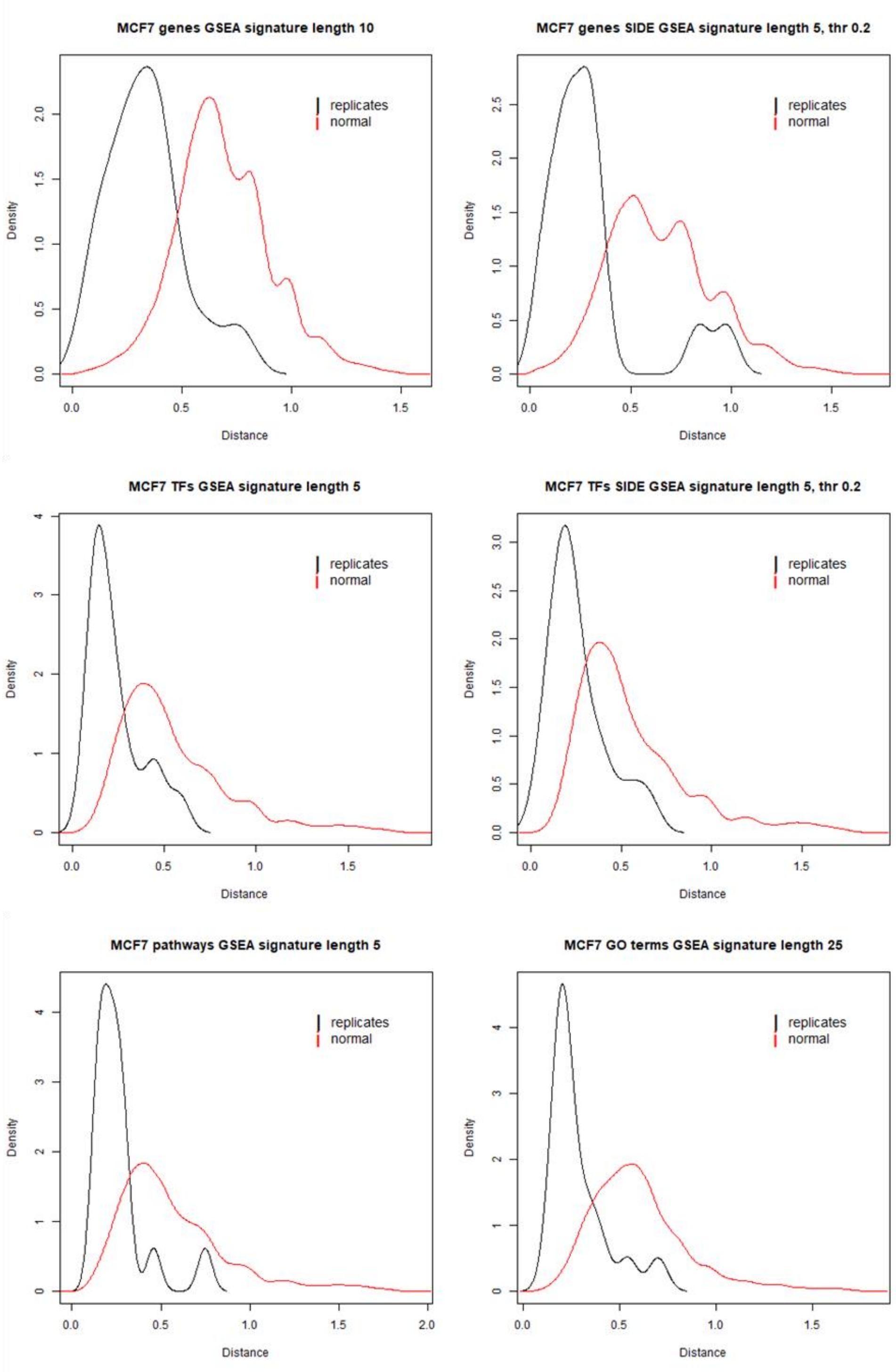
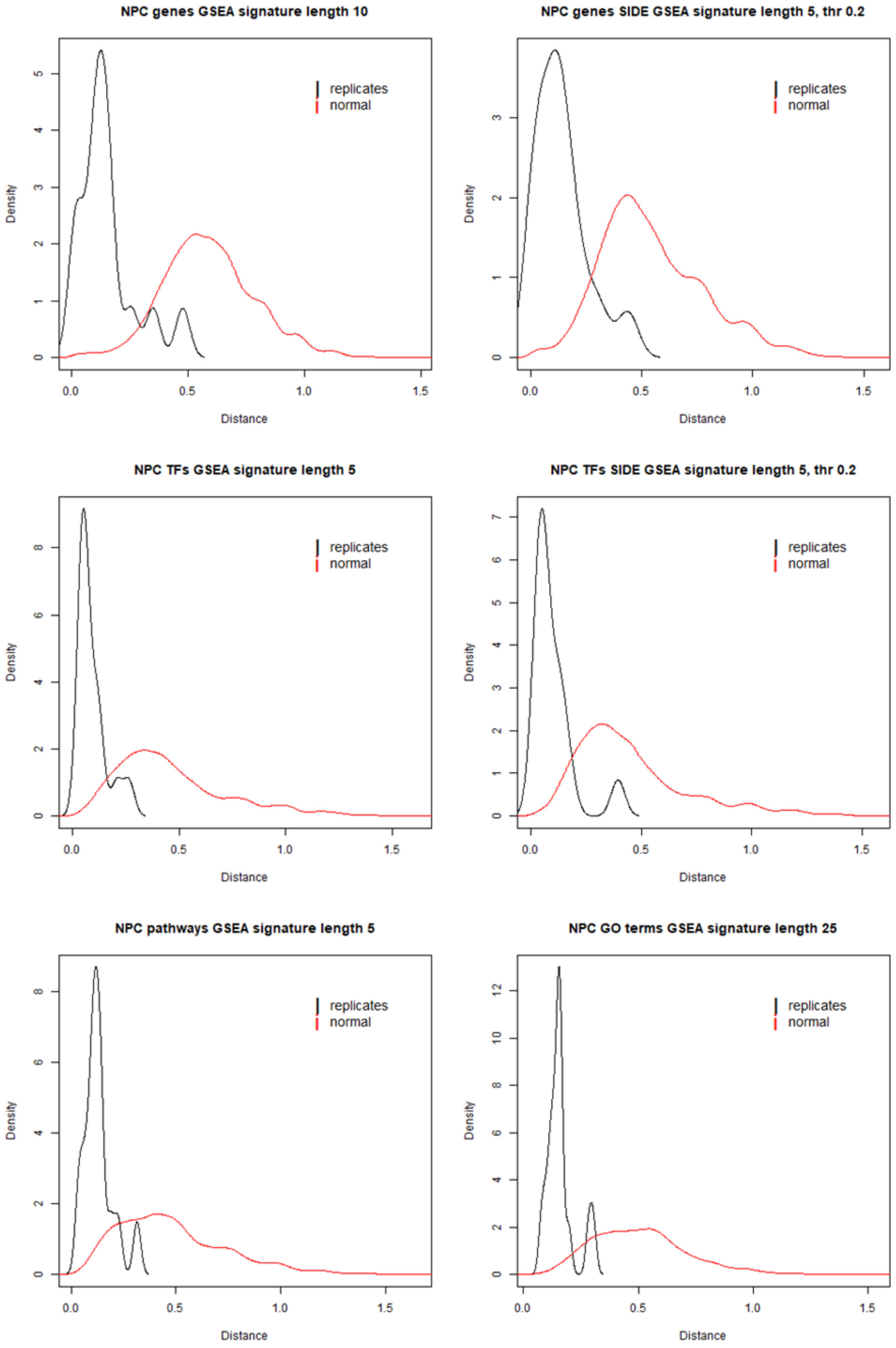


Figure 32 MCF7 cell line all distance calculation metrics density plots

3.4.2.4 NPC cell line


NPC genes GSEA signature length 10

NPC genes SIDE GSEA signature length 5, thr 0.2

NPC TFs GSEA signature length 5

NPC TFs SIDE GSEA signature length 5, thr 0.2

NPC pathways GSEA signature length 5

NPC GO terms GSEA signature length 25

Figure 33 NPC cell line all distance calculation metrics density plots

3.4.2.5 PC3 cell line

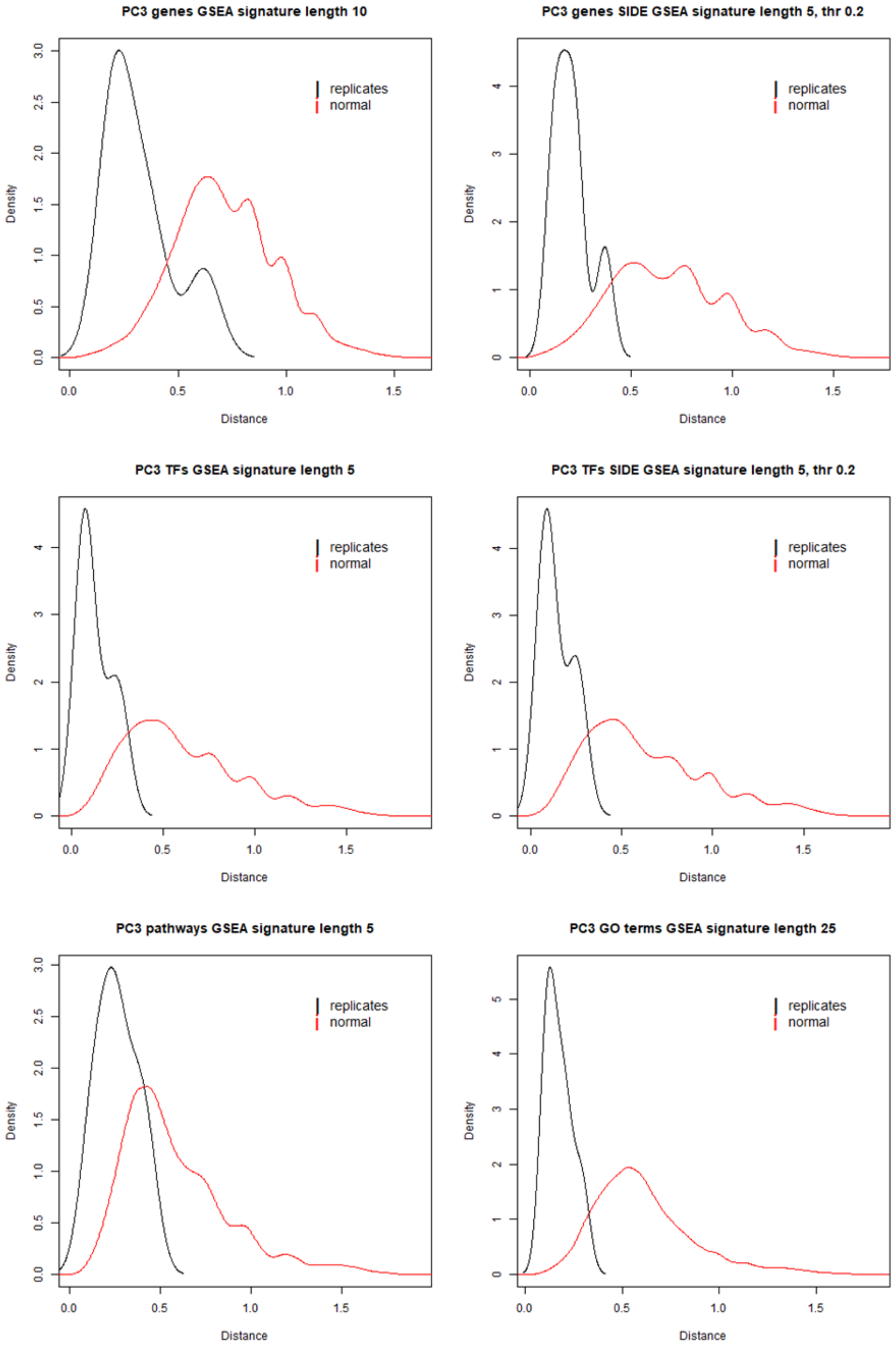


Figure 34 PC3 cell line all distance calculation metrics density plots

3.4.2.6 Results assessment

The distance calculation algorithms exhibit similar characteristics regarding each cell line. At some cases a distance metric may perform better or worse at a cell line when compared to the other cell lines. This has to do with each cell line's expression data, which vary between different cell lines due to the disease that is present at each one or even for experimental reasons (errors, noise). As a general remark, GO terms, when used as input at the GSEA score algorithm, provide at most of the cases better separation of the two distance categories than any other biological metric. Moreover, slightly better distances are observed at the replicated data, when SIDE GSEA score is applied, instead of normal GSEA score to the gene expression and the TFs. SIDE GSEA pairwise distance results are considered better for the replicated samples because they exhibit a bit lower distances than the normal GSEA distances for the same input. Finally, all the biological distances that are tested perform better than gene expression distance. This fact indicates that it is significantly meaningful to apply GSEA score at the rest of the biological metrics (TFs, signaling pathways, GO terms), despite that it has never been tested before.

3.4.3 Precision test on MCF7 data

As mentioned above, it is possible that some biological metrics provide different distance results when applied to different cell lines. Apart from this, the distance results can and do vary even in terms of the same cell line when different signature lengths are used. The distance results are linked to the number of differentially expressed genes that exist in the pair of perturbations under examination. This applies also to the TFs, the signaling pathways and the GO terms, since they are calculated using the gene expression. On this ground, it is highly possible that by using different signature lengths and topological similarity thresholds for SIDE, different distance results are produced.

Due to this fact, another approach is tested on the MCF7 cell line. This cell line is selected because it contains the most total perturbations than any other available. It has 517 samples consisting of 503 unique ones and 14 replicated.

3.4.3.1 Calculation of MCF7 average distances

This approach uses many signature length thresholds and calculates the average value of the distances computed each time to provide a more robust result that contains more information about the pairwise distances than an algorithm with one signature length. Regarding the GSEA score algorithm, for the genes and the GO terms, the signature lengths used are 10, 20, 30, 40 and 50, while for the TFs and the signaling pathways, the signature lengths are 4, 8, 12, 16 and 20. For the SIDE GSEA score, the topological similarity threshold is kept constant at the value of 0.2 since its increase, with the current PPI network used, introduces worse distances. The signature lengths used in this case for the genes and the TFs are 5, 10, 15 and 20. The results with the average pairwise distances computed for the above signature lengths for each metric are exhibited below. Then, a precision test is applied to the results in order to add one more insight into the methods tested and their effectiveness.

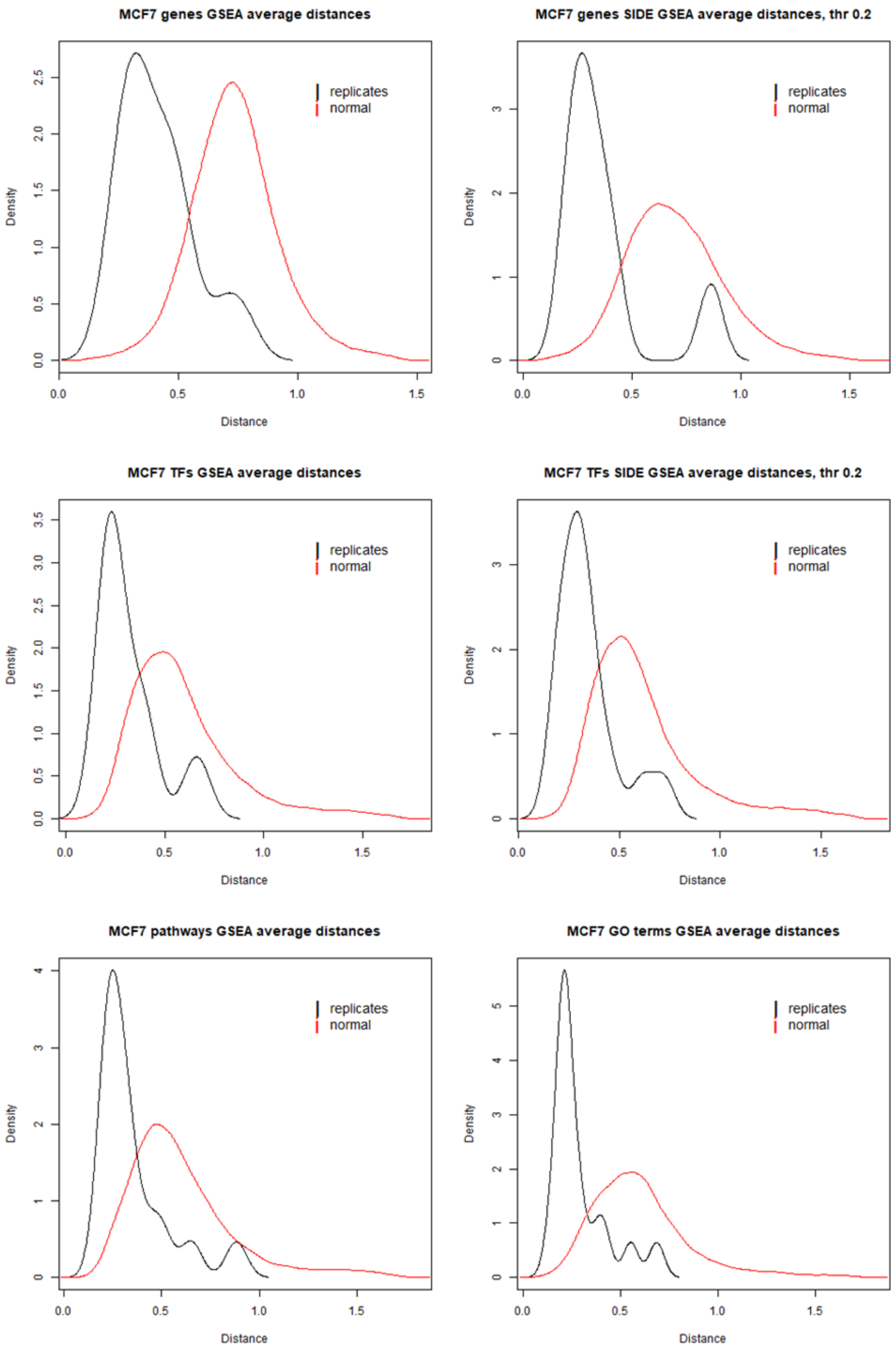


Figure 35 MCF7 cell line all average distance calculation metrics density plots

As the plots indicate, GO terms are again the best characteristic used for the separation of the two distance groups. It is also clear that while going from one metric to another (from gene expression to TFs, from TFs to pathways and from pathways to GO terms), the distance categories are more distinguished, with the main change being spotted at the replicated samples pairwise distances that are more and more gathered in low distance values.

3.4.3.2 Correlation between the various biological metrics

Based on the gene expression data the new biological data are constructed as mentioned before. The algorithms that are used to derive them are using prior knowledge about protein networks to provide the new data. In this paragraph, the correlation between the metrics under investigation is examined. The distances of each metric are plotted against another one and the results are shown below. Only the GSEA distances are used in order for the same distance calculation algorithm to be applied to all metrics and be comparable.

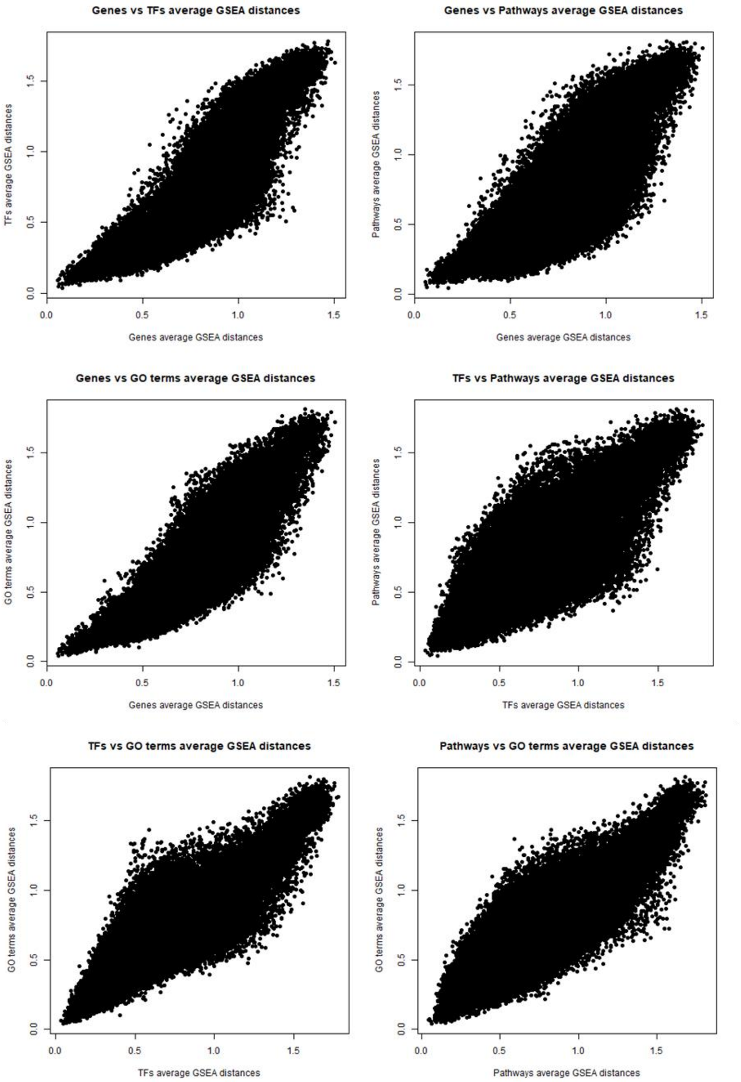
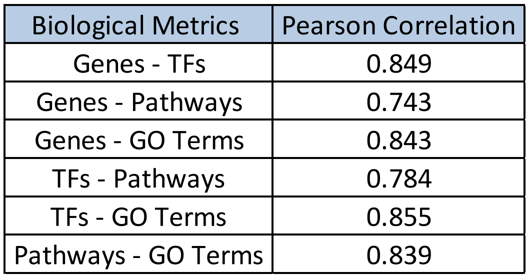


Figure 36 Biological metrics comparison plots

As the plots morphology indicates, the metrics are all highly correlated. This is an expected characteristic, since they are all based on prior knowledge that comes from protein networks. More specifically, the Pearson correlation of each pair of biological metrics is found as follows.



Biological Metrics	Pearson Correlation
Genes - TFs	0.849
Genes - Pathways	0.743
Genes - GO Terms	0.843
TFs - Pathways	0.784
TFs - GO Terms	0.855
Pathways - GO Terms	0.839

Table 5 Biological metrics pairwise Pearson correlations

The above correlations are high and prove the existence of a strong relationship between the biological metrics that are studied. The further investigation that follows is carried out on these metrics and regards the precision that each one achieves when compared with the 2-dimensional structural distance of the chemical compounds.

3.4.3.3 Precision test between 2D and the average biological distances on MCF7

As precision of the biological distances with regards to the 2D structural distances, the following value is defined.

$$Precision = \frac{TP}{TP + FP}$$

As TP, the number of the true positive pairwise distances is annotated and with FP, the number of false positive pairwise distances is annotated.

For the precision test to be carried out a threshold is calculated at the biological distances of the replicated samples and a varying 2D structural distance threshold is used to inspect the precision that is achieved. The biological distance threshold value defines the value below which the samples can be considered as biologically similar. It is calculated as the average of the replicated samples distances that are equal or less than 0.5.

Then, the moving 2D distance threshold is utilized to evaluate the number of pairwise distances that are correctly considered as similar based in 2D structural terms and the ones that are wrongly. The pairwise distances that are below or equal to both of the thresholds are the true positives, while the ones that are only below of the 2D structural threshold are considered as the false positives. A schematic representation of this method is shown below on the gene expression data. The green line corresponds to the gene expression distance threshold, while the blue one at a value of the moving 2D structural distance threshold. The points (pairwise distances of compounds) that are found in the green area are the true positives according to the thresholds that are tested, while the ones that belong in the red area are the false positives, meaning that they are considered as similar based on the 2D structural distance threshold under investigation but are not similar according to the biological distance threshold. The total positives are the sum of the true and the false positives. The precision of the 2D distance threshold is then calculated as presented above.

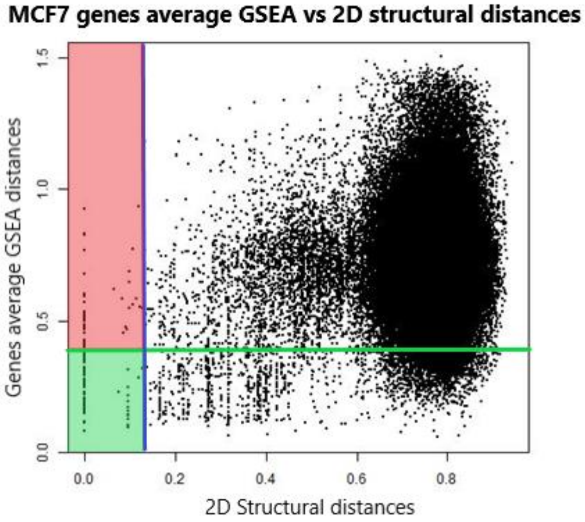


Figure 37 MCF7 cell line genes average GSEA distances vs 2D structural distances precision example

The biological distance similarity thresholds are calculated for each metric based on the average of the replicated samples distances that are below 0.5. Every biological pairwise distance that is below each category's threshold is considered as a biologically similar pair. Their values are the following.

Metric	Genes	TFs	Pathways	GO terms	SIDE Genes	SIDE TFs
Threshold	0.380	0.257	0.275	0.340	0.294	0.295

Table 6 Biological metrics similarity thresholds based on the replicated samples distances

Based on these values, the precision tests are carried out. The moving 2D structural distance threshold is set to take six values (0.05, 0.10, 0.15, 0.20, 0.25 and 0.30). The precision tests results are presented in a form of matrix.

2D threshold	Total Positives	Genes Prec.	TFs Prec.	Pathways Prec.	GO terms Prec.	SIDE Genes Prec.	SIDE TFs Prec.
0.05	89	71.91%	66.29%	62.92%	84.27%	60.67%	61.80%
0.10	112	70.53%	66.07%	64.28%	82.14%	61.61%	63.39%
0.15	141	62.41%	58.86%	55.31%	75.18%	54.61%	56.77%
0.20	275	53.09%	48.36%	48.73%	66.54%	47.64%	46.91%
0.25	411	52.80%	46.96%	46.23%	65.20%	44.52%	44.28%
0.30	670	49.25%	44.47%	44.18%	62.38%	41.49%	40.60%

Table 7 Precision test results

The precision values indicate the superiority of GO terms with respect to the rest of the biological metrics. GO terms achieve more than 10% higher precision at all 2D thresholds than the rest of the metrics. This result confirms also the rest of the investigations that also point to the GO terms as the best biological metric available between the four studied.

On the other hand, the TFs and the pathways seem to underperform. They are thought to provide better results than the gene expression but in this case it doesn't happen. The low precision values of these two metrics may be linked to the fact that the biological distance thresholds of the TFs and the pathways are lower than the ones of genes and GO terms. That's why fewer values than the ones according to the other metrics, are considered as true positives, and as a result lower precisions are achieved. Finally, SIDE GSEA algorithms exhibit low precisions and biological similarity thresholds as well. Their results will be assessed in the following paragraph.

3.4.4 Investigation of SIDE distance algorithm results

The results of SIDE GSEA score on MCF7 cell line are further investigated in order to get a broader view of the way that this scoring algorithm performs. The investigation is carried out by calculating the resulting averages of the replicated and the unique samples pairwise distances of the various thresholds used at the SIDE GSEA. Then, the difference of these averages is used to assess whether the two groups of distances are well-distinguished from each other. As mentioned before, this algorithm is applied on gene expression and TF activity data, since these are the only biological metrics present in the PPI networks that is uses. Moreover, most of the tests are characterized by a topological similarity threshold of 0.2, because this is the value that leads to the most satisfying results of replicate and unique samples separation.

3.4.4.1 SIDE distance algorithm performance on MCF7 genes

The resulting averages of the SIDE GSEA distances for various signature lengths when applied to the genes of MCF7 cell line are presented. The 2D structural similarity threshold is set at 0.2.

Signature length	Average of uniques	Average of replicates	Difference
5	0.6474	0.2992	0.3482
10	0.6894	0.3637	0.3257
15	0.7141	0.3896	0.3245
20	0.7309	0.4094	0.3215

Table 8 Genes SIDE GSEA distances with different signature lengths

As the signature length increases, the difference between the two groups of pairs (uniques and replicates) decreases. While this happens, the average value of the distances increases for both of the groups. This means that the algorithm considers the pairs as more dissimilar for both of the groups. Also, for the genes the effect of the 2D structural similarity threshold is investigated, while the signature length is kept constant, and provides the following results.

Structural threshold	Signature length	Average of uniques	Average of replicates	Difference
0.2	20	0.7309	0.4094	0.3215
0.3	20	0.7626	0.4730	0.2896
0.4	20	0.8507	0.7262	0.1244

Table 9 Genes SIDE GSEA distances with different 2D structural similarity lengths

The distance difference decreases as the structural similarity threshold increases and this is expressed as the two sample group distances being less distinguishable.

The simple GSEA score that was presented above, for a signature length of 10 and for the average of the signatures lengths (see 3.3.1) against which the algorithm overall performance should be tested has the following results on genes.

Signature length	Average of uniques	Average of replicates	Difference
10	0.6879	0.3360	0.3519
average	0.7367	0.4131	0.3235

Table 10 Genes simple GSEA distances

The SIDE GSEA distances differences are close the normal GSEA ones. This fact validates the performance of SIDE GSEA scoring distance, but also indicates that more research has to be done on this method to test whether significantly better results than the simple GSEA can be achieved.

3.4.4.2 SIDE distance algorithm precision on MCF7 TFs

The resulting averages of the SIDE GSEA distances for various signature lengths when applied to the TFs of MCF7 cell line are presented. The structural similarity threshold is set at 0.2 again.

Signature length	Average of uniques	Average of replicates	Difference
5	0.5507	0.2581	0.2926
10	0.6045	0.3317	0.2728
15	0.6339	0.3781	0.2558
20	0.6534	0.3978	0.2556

Table 11 TFs SIDE GSEA distances with different signature lengths

Once more, the increase of signature length induces a decrease of difference between the two distance groups. Additionally, the averages of both groups increase while signature length increases.

Regarding the simple GSEA distance algorithm, the results that it provides are the ones shown below for two signature thresholds (5 and 10) and the average of the thresholds (see 3.3.1.).

Signature length	Average of uniques	Average of replicates	Difference
5	0.5472	0.2330	0.3142
10	0.5987	0.3123	0.2863
average	0.5910	0.3168	0.2741

Table 12 TFs simple GSEA distances

The SIDE GSEA results are close to the simple GSEA ones at the TF case as well. When compared to the genes distance results, both the unique and the replicated samples exhibit lower distance values. This indicates the superiority of TF distances over the gene expression distances.

In conclusion, SIDE GSEA algorithm's results are validated according to the normal GSEA ones and more investigation on modifications of the SIDE GSEA algorithm should be done in order to find out whether other variables combinations are capable of performing better.

3.5 Implementation of a machine learning model

The machine learning model that is implemented, as described in the Materials and Methods section is validated by a set of 62,000 compound pairs. In this dataset, 80 unknown compounds are included and almost 1,000 known compounds. The term unknown refers to compounds that are not present in the training set.

The model attempts to predict whether a pair of compounds is biologically similar, in terms of GO terms distance, or not. GO terms are selected as the biological metric due to their overall better performance than the rest of the biological metrics investigated. The input is the two compounds' ECFPs, as described in the Materials and Methods section.

The model is tested in three set ups and its performance is assessed according to the following presented metrics.

The first set up (normal) is a single execution of the model. These results of this model come from a single run of the siamese neural networks.

The other two tested models use an ensemble method. According to this method, in both of the augmented and random case, the model is run 50 times with random initialization and the average distance values of these trials are calculated as the results.

Regarding the second model ("augment"), augmentation takes place because the dataset of the MCF7 cell line consists of two categories of samples based on their quality. The quality 1 samples are the best ones and they make up to almost 250,000 pairwise distances that are calculated from 713 compounds. The quality 2 samples provide around 5 million pairwise distances that are defined based on 3,300 compounds. In order to have better training set and for the model to gain an insight in a broader spectrum of the data, each time supplementary to the quality 1 data (250,000 pairwise distances), another 250,000 pairwise distances are used that come from the quality 2 data. These distances are differentially selected each time, but they must comply with a rule. The distribution of the input data has to remain the same with quality 1 data distribution each time. So, if a distance value was present one time at the quality 1 data, it will have two instances at the augmented dataset and so on, since the total number of input data is double of the quality 1 data. Along with the augmentation, the 50 models used for the final ensemble are randomly initialized, as mentioned above. In this way, the augmentation is different each time and the results are more representative, since the augmentation differs every time the model is tested.

When it comes to the "random" model, only the quality 1 data are utilized as input, with the calculated values being the result of an ensemble of 50 trials of the model, by taking the average distance value for each compound pair. Each one of the 50 runs of the model is characterized by a different, random initialization.

Model	Prec. @0.5%	Prec. @1%	Prec. @1.5%	Prec. (all)	Positives (Predicted)	MSE	MSE of FPs (1%)	Cor.	MSE 1%
ReSimNet (normal)	0.703	0.629	0.574	0.347	21,102	0.0156	0.0627	0.55	0.0288
ReSimNet Ensemble (augment)	0.803	0.734	0.709	0.494	6,882	0.0135	0.0382	0.48	0.0164
ReSimNet Ensemble (random)	0.742	0.674	0.591	0.345	21,331	0.0154	0.06	0.57	0.026

Table 13 Learning model results

As the results indicate, the model performs well, especially when the precision of the top 1% or less of true positives is investigated. As true positives, the compound pairs that are predicted to have low GO terms distance and they truly have, according to the GO terms GSEA score, are annotated. For instance, the top 1% of the true positives consists of the 1% most confident predictions of the model. These are the predictions that according to the model are the most certain to be biologically similar, and they truly are. The distance threshold for a pairwise distance to be considered as low, meaning that a pair of compounds is similar in terms of GO terms distance, is set at 0.22. Below this GO terms distance value, a pair of compounds is considered as similar and above that as dissimilar. This threshold is applied to the predicted GO terms distances, but also to the true ones, in order for the predictions to be defined as right or wrong.

Regarding the most confident values of the right predictions (top 1% or less), it would be expected that they refer to compound pairs with low structural and also biological distance, since the neural network model predicts them correctly similar with such confidence. Also, it makes sense for compounds with similar chemical structure to exhibit similar biological effects. Surprisingly, these distances refer to compounds that are not structurally similar at their majority, while being biologically similar. The histogram of the 2D structural distances of the compounds that belong to the top 1% of the model's true positives is as follows.

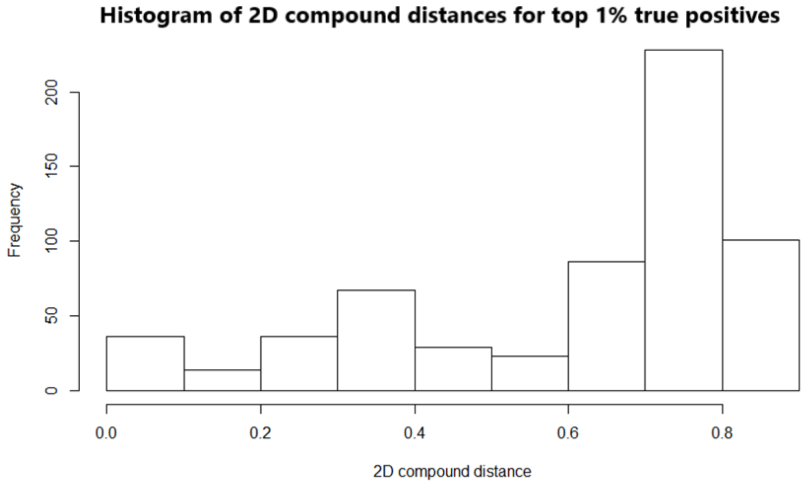


Figure 38 Histogram of 2D compound distances for the top 1% of the model's true positives

In fact, most of the compound pairs that are correctly predicted as biologically similar, based on GO terms, do not exhibit similar chemical structures. This finding uncovers the ability of the neural network model to learn and predict pairs of compounds that have similar biological effects but dissimilar chemical structures. This is the ultimate aim of this research and the model's results pose a confirmation of the hypothesis that the investigation of various biological levels can provide satisfying results and that a learning model can be useful in order to overcome the barrier that the strictly defined structural similarity algorithms raise.

The overall results of the learning model are presented below, where the predicted GO terms pairwise distances are plotted against the true ones.

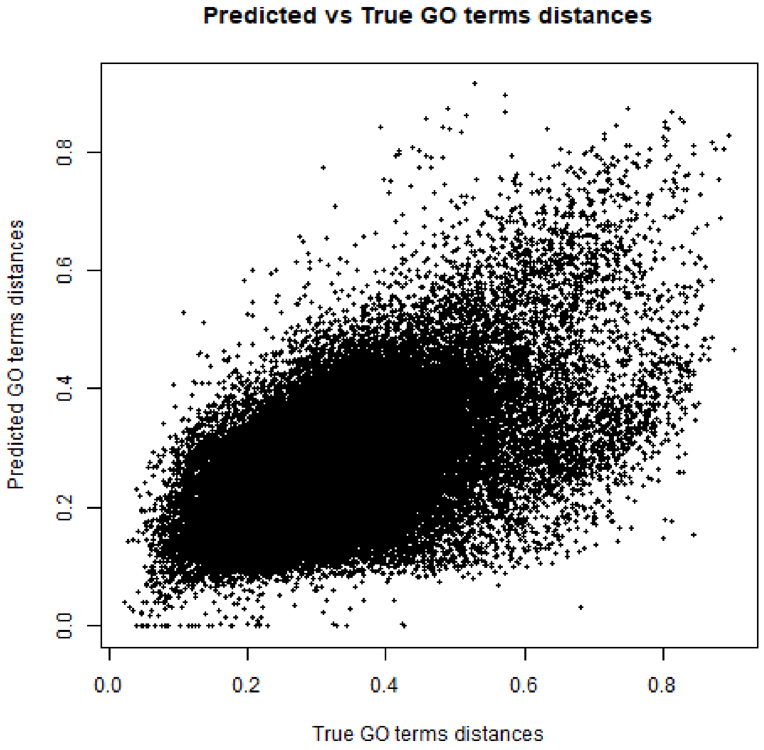


Figure 39 Model's predicted GO terms distances vs true GO terms distances

The Pearson correlation between the two categories of distances is 0.566, but as shown above when the model is confident about the predictions regarding biologically similar pairwise distances, they can be meaningful and provide results that are worth for further investigation to be carried out.

4 Discussion

The presented work dealt with the pharmaceutical compounds repurposing problem. The main goal of this work was the investigation of better and faster compound screening methods than the existing ones, in order to carry out more specific and targeted experiments in drug research.

Through this work, a detailed and innovative pipeline is defined for the investigation and validation of many new biological distances. Existing distance calculation methods are applied on specific biological data, on which they have never been applied before and are assessed. Moreover, a new method of biological distance definition (GSEA score using SIDE) is established, validated and investigated. This method uses prior protein network knowledge to calculate biological distances, which is something that hasn't been attempted in the literature yet.

Additionally, a machine learning model is implemented aiming to learn compound representation and biological distances according to the data that are created through the above mentioned pipeline, while having structural compound data as input.

The results of the new biological distances perform better than the existing baseline and the learning model provides satisfying predictions, with some of its top correct predictions being remarkable.

5 Further work

All of the findings presented are potential source for further work on the subject of compound distance investigation. First of all, an aggregation of the biological metrics (GSEA score on genes, TFs, signaling pathways and GO terms, as well as SIDE GSEA score on genes and TFs) could provide a new insight into the biological distances of the compounds. By testing various factors, the aggregation has the potential to surpass the other metrics when they are used separately.

Moreover, a lot of space for investigation exists regarding SIDE GSEA score. The tuning of the thresholds (topological similarity and signature length) followed by modifications of the algorithm could exhibit significant results. The modifications can be made on the scoring method, the possibility of adding penalties or variable similarity values at the topologically similar elements can be subject of further research. Also, the establishment of new PPI networks or the testing of other distance metrics on the PPI network is capable of improving the algorithm.

Last but not least, altering of the learning model's structure and functions to achieve better predictions is a major part of the further work based on this project. Modifying the learning model and investigating its capabilities could lead to a breakthrough regarding the compounds biological distance estimation.

6 References

[1] Rainer Breitling. Front Physiol. 1.9 (2010), *What is Systems Biology?*

[2] Carolyn R. Cho, Mark Labow, Mischa Reinhardt, Jan van Oostrum and Manuel C. Peitsch. Current Opinion in Chemical Biology 10.4 (2006), pages 249-302, *The application of systems biology to drug discovery*

[3] David S. Latchman. The International Journal of Biochemistry & Cell Biology 29.12 (1997), pages 1305-1312, *Transcription factors: An overview*

[4] Louis du Plessis, Nives Skunca and Christoplhe Dessimoz. Briefings in Bioinformatics 12.6 (2011), pages 723-735, *The what where, how and why of gene ontology - a primer for bioinformaticians*

[5] David P. Hill, Barry Smith, Monica S. McAndrews-Hill and Judith A. Blake. BMC Bioinformatics 9(Suppl 5).S2 (2008), *Gene Ontology annotations: what they mean and where they come from*

[6] Nahid Safari-Alighiarloo, Mohammad Taghizadeh, Mostafa Rezaei-Tavirani, Bahram Goliaei and Ali Asghar Peyvandi. Gastroenterol Hepatol Bed Bench. 7.1 (2014), pages 17-31, *Protein-protein interaction networks (PPI) and complex diseases*

[7] Justin Lamb et al. Science 313.5795 (2006), pages 1929-1935, *The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease*

[8] Aravind Subramanian et al. Cell 171.6 (2017), pages 1437-1452, *A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles*

[9] Sirci F., Napolitano F., Pisonero-Vaquero S., Carrella D., Medina D.L., di Bernardo D. NPJ Systems Biology and Applications 3.23 (2017), *Comparing structural and transcriptional drug networks reveals signatures of drug activity and toxicity in transcriptional responses*

[10] Dennise D.Dalma-Weiszhausz et al. Methods in Enzymology 410 (2006), pages 3-28, *The Affymetrix GeneChip® Platform: An Overview*

[11] Suleyman Aydin. Peptides 72 (2015), pages 4-15, *A short history, principles, and types of ELISA, and our laboratory experience with peptide/protein analyses using ELISA*

[12] Malachi Griffith, Obi L. Griffith et al. PLoS Computational Biology 11.8 (2015), *Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud*

[13] Rafael A. Irizarry et al. Biostatistics 4.2 (2003), pages 249-264, *Exploration, normalization, and summaries of high density oligonucleotide array probe level data*

[14] Rafael A. Irizarry et al. Nucleic Acids Research 31.4 (2003), *Summaries of Affymetrix GeneChip probe level data*

[15] Alvarez M.J., Shen Y., Giorgi F.M., Lachmann A., Ding B.B., Ye B.H. and Califano A. Nature Genetics 48.8, pages 838-847, *Functional characterization of somatic mutations in cancer using network-based inference of protein activity*

[16] Ege Ulgen, Ozan Ozisik, Osman Ugur Sezerman. bioRxiv*, pathfindR: An R Package for Pathway Enrichment Analysis Utilizing Active Subnetworks*

[17] Adrian Alexa, Jorg Rahnenfuhrer. *Gene set enrichment analysis with topGO*

[18] Anika Liu, Panuwat Trairatphisan, Enio Gjerga et al. bioRxiv, *From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL*

[19] Aravind Subramanian et al. PNAS 102.43 (2005), pages 15545-15550, *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*

[20] Junghwan Kim, Haekyu Park, Ji-Eun Lee, and U. Kang. (2018), *SIDE: Representation Learning in Signed Directed Networks.* In Proceedings of ACM International Conference on World Wide Web, Lyon, France, April 2018

[21] Minji Jeon et al. Bioinformatics, btz411 (2019), *ReSimNet: drug response similarity prediction using Siamese neural networks*

7 Appendix

The R programming language codes that are used to produce the results presented can be found at the BioSysLab Github repository regarding Biosimilarity, at the following link.

https://github.com/BioSysLab/Biosimilarity