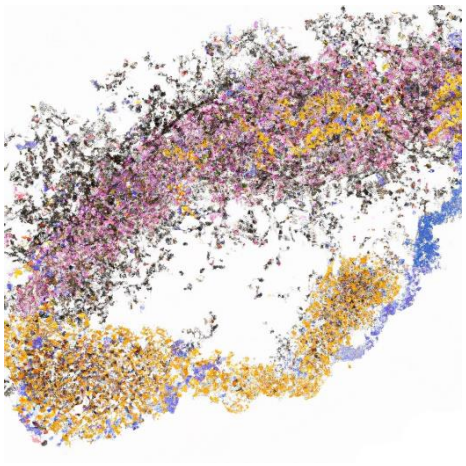
From Department of Microbiology, Tumor and Cell Biology
Karolinska Institutet, Stockholm, Sweden

METHODS, TOOLS, AND COMPUTATIONAL ENVIRONMENT FOR NETWORK-BASED ANALYSIS OF BIOLOGICAL DATA

Iurii Petrov



Stockholm 2023

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2023

ISBN 978-91 -8016-947-9
Cover illustration: Network of protein-protein interactions drawn by DALL·E 2 neural network

Methods, tools, and computational environment for network-based analysis of biological data
Thesis for Doctoral Degree (Ph.D.)

By

Iurii Petrov

The thesis will be defended in public at Ragnar Granit, Solnavägen 9 (Biomedicum), Solna, 22nd of March, 9:00-12:00

Principal Supervisor:
Andrey Alekseenko
Karolinska Institutet
Department of MTC

Co-supervisor(s):
Erik Aurell
KTH
Department of Computational Science and Technology

Ingemar Ernberg
Karolinska Institutet
Department of MTC

Opponent:
Professor George Spyrou
The Cyprus Institute of Neurology & Genetics
Department of Bioinformatics

Examination Board:
Mika Gustafsson
Linköping University
Department of Physics, Chemistry and Biology

Volker Lauschke
Karolinska Institutet
Department of FyFa

Lars-Gunnar Larsson
Karolinska Institutet
Department of MTC

To professor Galina Ivanova

Popular science summary of the thesis

Cancer is one of the leading causes of death, causing millions of deaths worldwide. Despite being considered one of the top-priority problem in medicine and great progress in treatment during previous several couples of decades, cancer is still relatively poorly studied. Cancer medicine gradually progresses from surgery and aggressive chemotherapy to the concept of precision medicine. Precision medicine requires discovery of cancer driver genes and potential therapeutic targets. In order to achieve this goal, we use network analysis algorithms. We demonstrate their applicability to a number of research problems in oncology as well as cancers of different types and origin. Based on that, we also offer public interactive tools for usage by researchers with no computational analysis skills.

Abstract

Cancer currently affects more than 18 million persons world-wide annually. It is a leading cause of death and so far, only 60% cure rate can be reached within the most developed health care systems. The nature of cancer has been a mystery for centuries, until discoveries during recent decades shed light on the underlying molecular events. This depended on the progress in understanding cell and tissue biology, developments of molecular technologies and of -omics technologies. Cancer has then emerged as a highly heterogeneous disease, however with some very basic mechanistic features common to all cancers. To deal with the complexity of causes and consequences of pathological changes in the molecular machinery, methods and tools of network analysis can be helpful. Complexity of this task requires easy-to-use tools, which allow researchers and clinicians with no background in computer science to perform network analysis.

Paper I describes a web-based framework for network enrichment analysis (NEA), using previously developed algorithm and code. The developed platform introduces functionality for a researcher to use data pre-downloaded from various popular databases as well as own data, perform NEA and obtain statistical estimations, export results in different formats for publications or further use in research pipeline.

Paper II presents development of another web server, which provided vast opportunities for exploration and integrated analysis of multiple public cancer datasets that describe *in vitro* and *in vivo* sample collections. The web server linked molecular data at the single gene level, phenotype and pharmacological response variables, as well as pathway level variables calculated with NEA and connected to the framework presented in Paper I. Researchers can use the platform for creating multivariate models based on raw or pre-processed data from various sources, visualize created models, estimate their performance and compare them, export models for further usage in own research environments.

Paper III demonstrates NEAdriver, a practical application of NEA to probabilistic evaluation of driver roles of mutations reported in ten cancer cohorts. NEAdriver results are compared with cancer gene sets produced by other, both network analysis and network-free methods. The paper demonstrated ability of NEA to be used directly for discovering novel driver genes as well as being used in combination with other methods.

In order to demonstrate benefits of using NEA, some rare cancer types and types with low mutation burden were used.

Paper IV is a manuscript evaluating performance of most representative methods of network analysis across methods' parameters, functional ontologies and network versions. This study emphasizes discovery of novel functional associations for known genes, as opposed to previous tests dominated by a few "gold standard" genes which were well characterized previously. We performed the analysis in the context of various topological properties of networks, pathways of interest, and genes. It employed both existing, widely used topological metrics and a number of novel ones developed for this analysis.

List of scientific papers

I. Jeggari A, Alekseenko Z, Petrov I, Dias JM, Ericson J, Alexeyenko A. EviNet: a web platform for network enrichment analysis with flexible definition of gene sets. Nucleic Acids Res. 2018 Jul 2;46(W1):W163-W170. doi: 10.1093/nar/gky485. PMID: 29893885; PMCID: PMC6030852.
II. Petrov I, Alexeyenko A. EviCor: Interactive Web Platform for Exploration of Molecular Features and Response to Anti-cancer Drugs. J Mol Biol. 2022 Jun 15;434(11):167528. doi: 10.1016/j.jmb.2022.167528. Epub 2022 Mar 5. PMID: 35662462. (first author)
III. Petrov I., Alexeyenko A. Individualized discovery of rare cancer drivers in global network context. eLife 11:e74010, 2022.(first author) https://doi.org/10.7554/eLife.74010 PMID: 35593700
IV. Petrov I., Alexeyenko A. Comparative performance of network versions, algorithms, and topological properties while discovering novel disease genes (manuscript, first author)

Scientific papers not included in the thesis

I. Bozoky B., Szekely L., Alexeyenko A., Ernberg I., **Petrov I.** Identification of novel protein markers of prostate basal cells by application of deep learning to more than a hundred thousand images from the Human Protein Atlas (manuscript)

Contents

List of abbreviations

NEA	Network Enrichment Analysis
AGS	Altereg gene set
FGS	Functional gene set
mRNA	Messenger RNA
miRNA	Micro RNA
CNA	Copy number alteration
PPI	Protein-protein interaction
GBA	Guilt-by-association
RWR	Random Walk with Restart
PR	Page Rank
PPR	Page Rank with Priors
GO	Gene Ontology database
DO	Disease Ontology database
FDR	False Discovery Rate
MB	Medulloblastoma
AUC	Area under curve
ROC	Receiver-operator characteristic
MCC	Matthew's correlation coefficient
MGS	Mutated gene set
API	Application Programming Interface

1 Introduction

1.1 Brief history of cancer

There are numerous evidences that humanity has been encountering cancer since the ancient times, from documents of Ancient Egypt and Greece to paleontological research. Medicine is a relatively young science, but oncology is even younger, people started to discover true nature of cancer only in the middle of 20th century. Of course, some attempts were made even in the 19th century, when the first carcinogens have been discovered, but mechanisms beyond tumorogenesis remained unknown for a very long time. Molecular biology gave the first deep insight into the nature of cancer. But mid-century discoveries were only the first step: the subtle molecular machinery was a big puzzle yet to be solved.

But, as history proves, man does not have to understand molecular biology to cure diseases. Humanity discovered properties of certain plants tens of thousands of years ago, later humans learned to make remedies out of them without knowing any chemistry. As many other diseases, cancer for a long time was believed to have supernatural roots, it was only logical for people back then to assume that it was some kind of divine punishment or curse. Yet even ancient physiology attempted to explain cancer in more scientific terms: Hippocrates produced his humoral theory, describing diseases as an imbalance between four cardinal fluids (blood, phlegm, yellow bile, black bile), determining cancer as an excess of black bile and viewing tumours as localizations of it. For the most of human history, the only way to cure cancer was surgical intervention. Without knowledge of molecular background, by the late 19th century physicians agreed that cancer can somehow “poison” the surrounding flesh, leading to operations such as radical mastectomy, which was not only removing breast, but the surrounding muscles, sinews and sometimes even some bones in attempt to eradicate cancer, often leaving patient crippled and deformed. While surgery was becoming more and more radical, diminished returns were noted soon: radical surgery led only to minor improvements of overall survival in cohort, while drastically reducing quality of life of all patients. In the beginning of the 20th century, it was understood that surgery cannot be a silver bullet in cancer treatment.

Cancer “drugs” likely started to occur even before the proper discovery of cancer itself, most probably in neolith. Some of them were plants to relieve the pain (such as chamomile), some of them were some kind of “magical” remedies, none of them really

worked. The breakthrough in medicine occurred with the discovery of cytotoxic drugs. Cytotoxic effects of mustard gas were discovered during the First World War, when it was noted that it leads to suppression of haematopoiesis. The first attempt to use mustard gas as a treatment rather than a weapon has been made in 1942 in New Haven, Connecticut, when a group of Yale University researchers attempted to treat cancer with nitrogen mustard. One year later, in December of 1943, cytotoxic effects of mustard gas presented themselves on a larger scale. Due to infamous German attack on the port town of Bari, where dozens of ships were docked, at least 1000 people died. Many ships carried bombs filled with mustard gas. Survivors, exposed to carcinogenic agent, suffered from severe consequences. Yet, this tragic episode led to approval of the first chemotherapy drug – mustine. Chemotherapy damages actively dividing cells more, thus eliminating cancer cells – unless they obtain mechanisms of resistance to it.

The third cure came from physics. In 1895 X-rays were discovered by Wilhelm Röntgen. Pretty quickly it was noted that exposure to them can give skin burns. The idea of using radiation to cure skin lesions was born. The procedure underwent many changes and new forms of radiotherapy are being developed up to the date – such as the proton therapy. Just like chemotherapy mentioned earlier, radiotherapy produces breaks in DNA, forcing dividing cells to die. It can be localized, minimizing side effects, but cancer cells can still develop resistance to radiotherapy, and side effects can still be morbid.

While normal tissues have well-organized structure and certain life cycle, the cancer molecular phenotypes appear chaotic and unpredictable. Cancer cells emerge via the struggle to survive. In order to survive, they subvert several cellular mechanisms involved into control of cell life and behaviour(1). It was mentioned earlier that cancer cells can acquire resistance to chemotherapy and radiotherapy. Avoiding apoptosis is one of the hallmarks of cancer. But it is not the only trick cancer has in its sleeve. Metastasis is one of the most effective ways to evade extermination for solid tumours: cancer cells detach from tissue of origin, through mechanism of epithelial-mesenchymal transition invade blood stream and eventually find new “home” in another organ. This cancer cell dissemination allows tumour to create “colonies”, some of which could consist of just several cells, making them impossible to detect. To make things worse, cancer cells can become dormant, thus “populated” small colony can lie in wait for a long time, expanding rapidly(2). Normally, any abnormal cells should be killed by immune system, but cancer cells acquire immune evasion and ability to recruit immunosuppressive cells(3,4). Luckily, this does not mean that immune system cannot be activated “externally”: immunotherapy is an umbrella term for different treatments, activating native (to the host) immune cells or introducing specially produced immune cells(5–7).

Immunotherapy is considered to be safer, compared to chemotherapy, because this type of therapy is targeted, it attacks only cells introducing certain markers and features, while chemotherapy is non-targeted. It is “milder” for a patient, producing less side effects.

Surgery, chemotherapy, radiotherapy, immunotherapy – these four methods of cancer treatment are approved nowadays. Surgery and radiotherapy are the most straightforward since they can be used “physically” and on something visible. Hormonal therapy is also well established for some cancers, such as breast and prostate cancer, while targeting therapies are being developed starting twenty years ago. Immunotherapy and targeted therapy require understanding of cancer biology. The main questions are: why do these tumours arise? How can we distinguish them on molecular level to provide the most effective treatment with minimal side effects?

1.2 Cancer drivers

Cells can accumulate mutations during lifetime. Most of the mutations are “silent” mutations, which means they don’t affect phenotype, cell functions or behaviour at all. However, some of the mutations can change cell’s fate drastically. In nature mutations can be harmful or giving an organism a certain selective advantage – this is how evolution works. But in human organism mutations which gave cells this advantage are harmful, since such “selfish” cells give rise to cancer. Mutations, which lead to cancer initiation or progression, are called driver mutations. Other mutations, which does not affect cancer progression, even if they are not silent, are called passenger mutations(8).

Each type of cancer has its own characteristic signature, indicating most often mutated genes for the certain type of cancer. Pretty often cancer type can be broken into several subtypes based on mutational signatures. This is called a cancer landscape(9,10).

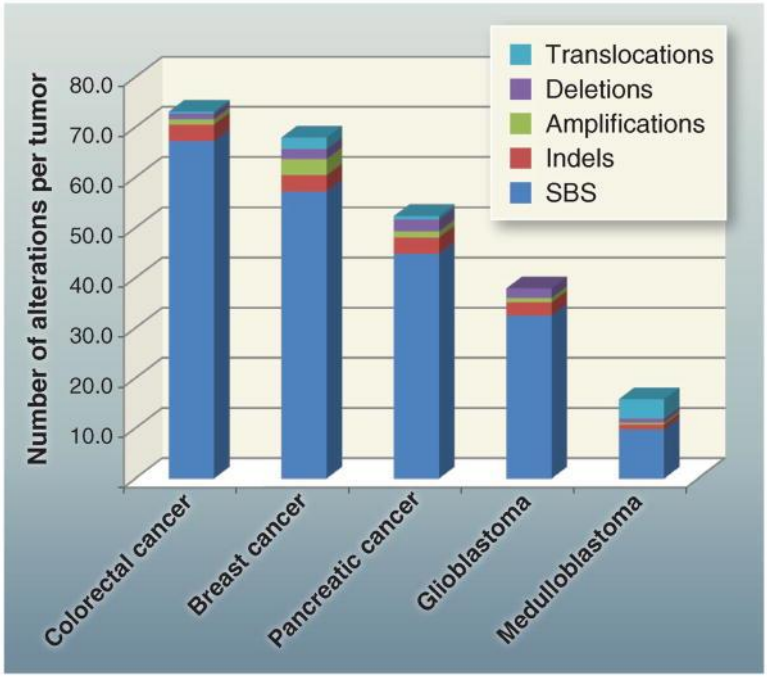


Figure 1. Cancer landscape for several types of cancer (Vogelstein et. Al. Cancer Genome Landscapes. Science, 2013).

Mutations are individual in every cell. Once a cancer cell obtains an advantage, it "overpowers" other cells and establishes its cancer cell line in patient. There can be tumours "founded" by only one cell line – monoclonal tumours. But it is a common situation when oncologists observe several competing cancer cell lines in a patient. Cancer cells are chaotic and ever evolving, so we cannot even describe each cancer cell type present in patient(11–15). This heterogeneity exists on different levels: intra-tumour heterogeneity(12,16) and, if patient has several tumours, intra-patient heterogeneity. Naturally, set of mutations is unique in each patient, leading to inter-patient heterogeneity.

Point mutations are not the only reason of cancer initiation and progression. There are also copy number alterations (CNAs)(17,18), different chromosomal events (such of kataegis(19) or chromothripsis(20)), and epigenetic alterations(21,22). These events are called "driver events". Patient familial history, which predispose to cancer, such as Li-Fraumeni syndrome and exposition to certain carcinogenic substances play an important role. All these factors increase complexity even further.

These points lead to the conclusion that cancer treatment should be personalized(23). This requires a mathematical framework which would allow researchers to describe interactions of different molecular agents and tools to discover previously unknown drivers and their interactions, suggest relevant therapeutic targets. One of the promising models for this purpose is network model. First, it allows naturally represent relations in molecular machinery. Second, it offers many powerful methods ready to be adopted for *in silico* discovery of novel drivers and treatment targets.

1.3 Networks in biology and medicine

Network theory evolved from graph theory and provided researchers with convenient ways to represent relationship between different entities. Entities (or agents) are represented as nodes (or, in graph theory terms, vertices) and their relations are represented as links (edges). There are different modifications of this basic model:

1. Links can be directed – for example, such link can mean "protein A affects protein B".
2. Links can have different type – for example, "protein A suppresses protein B".
3. Links can be weighted – for example, to represent probability of existence of such link or strength of established interaction.
4. There can be multiple links between two nodes ("multigraph" in graph theory terms) – for example, representing different types of interactions simultaneously existing between entities.
5. One link can connect multiple nodes ("hypergraphs" and "ultragraphs").
6. There could be different types of nodes – for example, some nodes represent mRNAs, others represent miRNAs etc.

As can be clearly seen from the features described above, networks are a flexible tool for a wide range of biological and biomedical research tasks(24–27). However, such a powerful mathematical model does not automatically guarantee good results. Among the most widely used types of networks in biomedical research are protein-protein interaction networks (PPI networks). Many networks were constructed for biological and biomedical needs(28–31). Despite the wide variety of possibilities offered be network science, most networks are simple graphs, either directed or undirected.

Previously described features of network models highlight representation of data, but in order to process data algorithms are required. Most of the used algorithms either came from the graph theory or were introduced in other scientific fields, such as in computer science. Importantly, many mathematical constructs appeared to have the meaning in biological context.

Node degree is a number of nodes connected with the given one: number of outgoing connections is called out-degree, while number of incoming links is called in-degree. Biological networks are assumed to possess a number of topological properties:

- Node degree distribution is often assumed to be "power-law": very few nodes in the network are hubs, while the vast majority of nodes are low-level nodes(32–35). However, this point is challenged by some researchers(36,37).
- Biological networks are "small world" networks: distance between any two reachable nodes is relatively short compared to many other types of networks, such as transport networks.

Nodes with very high degrees are called hubs. In Internet, hubs represent very popular websites, such as Google and Amazon. In biological networks they may represent powerful, global-level regulators or so-called "essential genes"(38,39), removal of which leads to severe consequences, including cell death. But even low-degree genes may represent "disease genes" – mutations or dysregulation of such genes can lead to abnormal phenotypes, although not necessarily cell death. Hubs tend to cluster together, producing "cliques" – densely-connected areas of high-degree nodes in network.

In order to identify genes linked to diseases different methods were adopted or designed (40–47). The most intuitive and popular idea was "guilt-by-association" approach (GBA), which assumes that genes strongly associated with known disease genes are themselves involved in relevant pathogenetic processes. Originally, only direct neighbours of known disease genes, but with further understanding of molecular biology more advanced methods were adopted and discovered.

The modern understanding of cancer raised from the level of major and universal causative genes to the paradigm of cancer pathway(48,49). Pathways (or disease modules in case of disease related genes) represent sets of functionally related genes

involved into some common biological process. In general, we call them functional gene sets (FGS). On the other hand, an important observation from the large-scale exome sequencing was that mutation patterns were not uniform even among patients of very similar cancer phenotypes. Rather, the mutation profiles of known cancer genes were sparse and disjoint. FGSs might be imagined as clusters of nodes connected in the network. However individual members of these clusters can still be relatively far apart from each other in the network. Distance between two nodes or a node and a whole FGS can be measured as a minimal uninterrupted path length in the network. Therefore studying network patterns of mutated genes could suggest cancer-related roles by associating them to known relevant pathways. The pathway-level view of cancer thus reduces the heterogeneity of cancer genomes: rather than single gene level of classification, cancer cases can be classified more systematically by affected pathways. Pathway analysis allows to better classify tumours(50–53) and offer more precise treatment based on altered pathways(54,55), e.g. two subgroups of medulloblastoma are named after deregulated pathways (SHH and WNT)(56). Pathway-level overview not only allows us to understand cancer biology better(57), but also opens new perspective for discovering novel driver genes and eventually cancer treatments by offering the drivers or their constellations as new potential therapeutic targets(58,59). A number of dedicated pathway-level resources have been created(60–62) for these purposes.

1.4 Network methods

It was previously mentioned that a number of network methods were adopted or specifically developed for biological tasks. Most of the previously cited methods are, in fact, a modification of one network algorithm.

Network methods can be divided into two categories according to their assumptions about network properties. The first class makes does not explicitly consider topological properties of the network, while methods belonging to the second class can, for example, account for node degrees of analysed genes.

One of the most abundant family of network methods are random walk methods(44,63). They are based on a simple physical analogue: there exists a walker in the network, who starts at some node and performs random walks along connections belonging to the node (outgoing connections, if network is directed), this process continues in iterative manner. In the end, assigned scores to the nodes represent probability of walker to land

on the respective node at any given step. The most popular modification of this simple model is random walk with restart (RWR). In this model, a walk starts from the set of chosen start nodes, called “seed nodes”. At each step it either continues along connections of the current node (if this is feasible) or – with a certain probability – jumps back to a randomly chosen seed node. Random walk methods are popular in biomedical research and produced a variety of modifications(45,64–67), some of which require but they have certain problems, which arise from the properties of both algorithms and networks. First, random walk methods tend to prioritize hubs, which is natural: for example, Google PageRank(68) was designed with the idea that the most important (i.e. best referenced externally) Internet pages must be of top priority for users. There are plenty of modifications of random walk models, some of them are designed for biomedical purposes solely(45,64,65,69). However as was mentioned before, although biological hubs could be essential or otherwise “powerful” genes, the disease research may concern less connected genes. A low node degree can be due to the gene being poorly studied or absence of a role in the global regulation. None of these precludes its pivotal role in a certain pathological condition.

Other methods, such as Network Enrichment Analysis (NEA)(70) employ distribution of node degrees for estimating confidence of functional relations between e.g. a set of seed nodes such as an altered gene set (AGS) versus another, functional gene set (FGS) . Unlike random walk methods, which estimate probability of landing on any given node, NEA actually measures excess of edges between two node sets or a gene and a set above a level expected by chance, i.e. in a random network (given its node degree distribution is the same as in the original network). NEA reports probabilistic estimates of significance in the form of p- and q-values for each pattern of interest.

It is assumed that network analysis results might be biased due to unequal topological properties of individual nodes (gene or protein level), node sets (pathway or AGS level), or even whole networks (network level) – as random walk methods might prioritize hubs etc.

On network level, such metrics as network size(71), diameter, density and motif distribution(72,73) can affect the results, though these effects were not systematically studied. Moreover, some of them cannot be really altered in a given network. For example, increasing or reducing a number of nodes and edges between them can be done by filtering edges by their weight (for example, confidence score), while density and motif distribution cannot be altered in a designed manner.

On pathway level, we study properties of certain AGSs in the given network. AGS (term) can consist of several disconnected modules in network (in some cases they can be called "disease modules"(38)). Number of modules, their sizes and placement in the network (including placement in relation to each other) can be important.

On gene level, except for node degree, a number of other characteristics should also be taken into account. Centrality measures can significantly affect results of random walk methods(74). There exists a number of centrality measures(75,76),but the most basic and popular are the following three:

1. Closeness centrality – meant to emphasise closeness of the given node to any other node in the network. For connected graph formula is the following:

$$C(v) = \frac{N}{\sum_t dist(v,t)}$$

where:

- N – number of nodes in network;
- $dist(v,t)$– distance from node v to node t.

2. Betweenness centrality – represents a measure of influence of individual node on the flow of information in network, the higher node betweenness centrality – the more important the given node is. It must be noted that betweenness centrality can be related to node criticality, i.e., critical nodes will receive high betweenness centrality score. Betweenness centrality for node v is:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where:

- σ_{st} – number of shortest paths between nodes s and t;
- $\sigma_{st}(v)$ – number of shortest paths between nodes s and t through the node v.

3. Eigenvector centrality is also known “prestige score” of the in network. This metric is designed to measure the node’s influence over network. PageRank algorithm calculates a variant of eigencentrality(76). Eigenvector centrality for node v is:

$$x_v = \frac{1}{\lambda}\sum_{t\in M(v)} x_t = \frac{1}{\lambda}\sum_{t\in V} a_{v,t}x_t$$

where:

- λ – constant, eigenvalue;
- v, t – nodes belonging to the node set V;
- $M(v)$ – set of neighbourhood nodes of node v;
- $a_{v,t}$ – element of adjacency matrix A indicating adjacency relationship between nodes v and t: 1 if v directly linked to t, 0 otherwise.

In addition to aforementioned metrics, we propose two metrics developed by us in order to capture local topological metrics. They are designed to assess centrality metrics of nodes in respect to seed nodes/AGSs. These metrics are described in detail in Paper IV. Below there are their short descriptions and formulas:

1. Modular closeness centrality – calculates closeness centrality for a node v in relation to a certain module:

$$C_m(v) = \frac{N_m}{\sum_{t\in m} d(v,t)}$$

where:

- N_m – number of nodes in module *m*;

For ontology split into n modules this formula transforms into:

$$C_M(v) = \frac{\sum_{i=1}^{n} N_{m_i}}{\sum_{i=1}^{n}\sum_{t\in m_i} d(v,t)}$$

where:

- $M = m_1 \cup m_2 \cup \ldots \cup m_n$

2. Betweenness centrality with respect to modules – calculates betweenness centrality for a node in relation to a distributed modular term. For a term split into 2 modules:

$$g_{m_1,m_2}(v) = \frac{\sum_{s \in m_1, t \in m_2, s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}}{N_{\sigma_{m_1,m_2}}}$$

where:

- $N_{\sigma_{m1.m2}}$ – number of shortest paths between all nodes belonging to module m_1 and all nodes belonging to module m_2.

General form for an ontology split into *n* modules:

$$g_M(v) = \frac{\sum_{m_i \in M, m_j \in M, i \neq j} \sum_{s \in m_i, t \in m_j, s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}}{\sum_{m_i \in M, m_j \in M, i \neq j} N_{\sigma_{m_i,m_j}}}$$

Additionally, we propose to incorporate centrality metrics in attempt to estimate "weight" of the given set of nodes in the network. We follow the assumption that network is uneven and consists of "core" (one big clique), containing densely connected hubs, and "outskirts", containing loosely connected relatively low-degree nodes. We developed the following metrics for these purposes:

1. Closeness weight: sum of closeness centralities for the given set of nodes in network divided by total sum of closeness centralities. In this approach the whole set is viewed as one "supernode", the score indicates proportion of closeness obtained by this "supernode", allowing to estimate place of the node set in the given network. In other words, this metric represents proportion of closeness centrality acquired the node set.

$$CW(M) = \frac{\sum_{v \in M} C(v)}{\sum_{t \in N} C(t)}$$

where:

- M – given module (set of nodes) in network;
- N – set of all nodes in the given network;

- $C(v)$– closeness centrality for node v (see formula for closeness centrality above).

2. Betweenness weight: similar to the previous metric but uses betweenness centrality instead.

$$BW(M) = \frac{\sum_{v \in M} g(v)}{\sum_{t \in N} g(t)}$$

where:

- M – given module (set of nodes) in network;
- N – set of all nodes in the given network;
- $g(v)$– betweenness centrality for node v (see formula for betweenness centrality above).

3. Closeness weight per node: closeness weight of the term divided by the term size. Since terms with more nodes included can naturally absorb more closeness centrality, this metric represents normalisation of closeness weight by term size.
4. Betweenness weight per node: betweenness weight of the term divided by the term size, reasons of correction are similar to the previous metric.
5. Relative closeness weight: average closeness centrality of the nodes in the group divided by average closeness centrality among all the other nodes. If value is greater than 1, than nodes of the given set have higher closeness centrality than average node in the remaining network and we assume that the given set of nodes tend to lie closer to the network core.

$$RCW(M) = \frac{\frac{\sum_{v \in M} C(v)}{\|M\|}}{\frac{\sum_{t \in N, t \notin M} C(t)}{\|N\| - \|M\|}}$$

where:

- M – given module (set of nodes) in network;
- N – set of all nodes in the given network;
- $\|N\|, \|M\|$ – power of set N (M) (amount of nodes in the given set);
- $C(v)$– closeness centrality for node v (see formula for closeness centrality above).

6. Relative betweenness weight: similar to the previous metric but uses betweenness centrality instead.

$$RBW(M) = \frac{\frac{\sum_{v \in M} g(v)}{\|M\|}}{\frac{\sum_{t \in N, t \notin M} g(t)}{\|N\| - \|M\|}}$$

where:

- M – given module (set of nodes) in network;
- N – set of all nodes in the given network;
- $\|N\|, \|M\|$ – power of set N (M) (amount of nodes in the given set);
- $g(v)$– betweenness centrality for node v (see formula for betweenness centrality above).

Last but not least, such methods of PR and RWR have tunning parameters (e.g. damping factor of PR), which can affect results significantly(77,78). Up to the date, there have been a few attempts to find a universal parameter value (or an algorithm for finding such a value), but no solid solution has been found(79). Some of the methods even claim that even within combination of one network and set of nodes parameter values should be set for individual nodes(66).

2 Aims of the thesis

The main purpose of this thesis is to develop a reliable network method for discovering novel cancer drives and therapeutic targets and implement is as a widely accessible framework for researchers. Secondary aim is to discover topological properties of networks affecting results of network methods and prove that the developed method is not biased towards any of them.

In Paper I we implemented web-platform utilizing NEA implementation as a backend. We demonstrated practical use cases of the created webserver, as well as explained advantages for researchers. This paper became a foundation for some functions of the novel web-framework described in Paper II.

In Paper II we presented a novel web server, which is designed to enable researchers with limited or no programming skills to explore vast public datasets (such as TCGA and CCLE) as well as perform certain computational experiments on them, such as creating regression models by combining variables of different types. This webserver has pre-downloaded data, both raw and pre-processed with different methods, including NEA. Previously developed framework is used for discovering novel drivers and potential treatment strategies (through exploration of drug correlations with molecular and clinical variables).

In Paper III we demonstrated application of the developed method for discovering novel potential drivers and therapeutic targets in pan-cancer cohort. We also combined the developed method with several previously known methods in order to achieve the better performance and minimize FDR. Results of the study demonstrated that NEA can be used for rare types of cancer and tumours with low mutational burden.

In Paper IV we compared several widely used "core" network analysis methods – such as Page Rank and Random Walk with Restart – with two NEA implementations. We estimated performance of different methods (employing different methods' parameters, when available) on various networks and ontologies, namely, on Gene Ontology (GO)(80,81) and Disease Ontology (DO)(82,83). We explore how different topological metrics, some of which were developed by us, affect results and attempted

to find the best combination of network version, method (and tunning parameter) for discovering novel genes, e.g. driver genes or novel members of some biological pathway. In this paper we paid more attention to low-degree nodes, since they often represent poorly studied genes.

3 Materials and methods

3.1 Data retrieval from public sources

The present study strongly relies on public datasets (ethical permits were not required). This is done due to the following reasons:

1. It allowed us to collect vast data cohorts, which is impossible to do under normal circumstances.
2. Using large publicly available data makes research results reproducible.

In our study, we used the following data sources:

1. The Cancer Genome Atlas (TCGA)(84) – we downloaded data of various types and used it in Papers II and III.
2. Cancer Cell Line Encyclopedia (CCLE)(85) – cancer cell line data was utilized in Paper II alongside with drug sensitivity data from CTRP v2.0(86) and GDSC1 and GDSC2 datasets(87).
3. PEME-CA and PBCA-DE medulloblastoma cohorts were downloaded from CDC website and used in Paper III.
4. Data retrieved from publications was used in Paper III.
5. Gene Ontology (GO) and Disease Ontology (DO) data was used in Papers I-IV in different forms.
6. KEGG pathways(88–90), data from BioCarta(91), Reactome(60), WikiPathways(92), MetaCyc(93) and MsigDB(94) was used in Papers I-III.
7. We used FunCoup (FC) v3(28), STRING v10.5(29,95), Pathway Commons (PWC) v9 networks in Papers I-III; we used FC1(96) FC2(97), Fclim(8) in addition to previously mentioned networks in different configurations and their merges in Paper IV.

Specific versions of retrieved dataset are specified either directly (if versioning is available) or by the date of information retrieval (we need to note that while data in publications is static, data in CDC could be updated). Some original data is available for download alongside with the processed data from our own platforms or other open public platforms.

3.2 Data preparations

The presented study relies on many different data types, each processed in a unique way in accordance with the demands of the specific paper. Below data preparation is described for each paper.

Paper I relies solely on network data (11 networks in total) as well as data from several pathway databases (such as GO, KEGG(88–90) etc.). All the provided data is present with minimal or no processing.

Paper II has some intersection in data with Paper I but adds pre-computed correlations and data from TCGA(84) and CCLE(85). Downloaded data was processed according to the following rules:

- Point mutations – type of mutation, treated as binary variable for analysis, representing is certain gene has a mutation or is a wild type; if mutation is present – its type is disregarded for statistical analysis (see below) but available for exploration. If there is no information on mutation of a certain gene, this gene is automatically considered to be a wild type.
- CNA – if not already done in original files, this variable is calculated as $\log_2(CN)$, where CN is copy number for the certain gene in the original file.
- Gene expression – log-transformed values obtained from Illumina RNA-seq and Affymetrix platforms, if not already done in the original file.
- Methylation – beta values were transformed to M-vlaues (a.k.a. logit units) as $\log_2 \frac{\beta}{1-\beta}$.

Pre-computed results of correlation analyses between omics variables and response to specific drugs, detected with univariate or covariate linear models are available for exploration and analysis. Table 1 (corresponds to Supplementary Table 2 of the original paper) presents amount of obtained significant correlations.

Table 1. Number of significant correlations discovered in different cancer cohorts in Paper II

Cohort	Mutation-based NEA profiles	Expression-based NEA profiles	Gene copy number	Point mutations	Gene expression	Protein expression
	MUT.NEA	**GE.NEA**	**CNA**	**MUT**	**GE**	**PE**
BRCA	10264	5939	86309	41251	1025989	5201
PAAD	544	324	6921	3922	11988	1164
PRAD	107	483	1401	0	26037	192
GBM	1173	12712	45423	440	362716	1491
OV	6516	12214	294428	3891	810621	0
LUSC	46	338	10657	52	93587	239
COAD	2405	3754	14739	29465	74142	0
LUAD	1504	342	3879	2576	116722	1169
SKCM	2697	1062	2549	4787	65747	1149
BLCA	4069	2160	0	0	0	0
CCLE	1984964	1946899	0	20033	3287447	0

Significance of correlation was estimated using functions from R package `base`. For CCLE source, 2 models were created:

- 1-way ANOVA: `anova(lm(Sensitivity ~ Feature))`.
- 2-way ANOVA: `anova(lm(Sensitivity ~ Tissue + Feature))`.

Where:

`Feature` – an original molecular variable MUT, CNA, GE, METH) or NEA score for the given gene or pathway;

`Tissue` – a covariate for organ or tissue of origin of the cancer cell line;

`Sensitivity` – resistance to the specified drug.

The p-values for `Feature` were of main interest in this paper, we adjusted them for multiple testing by Benjamini-Hochberg method.

For TCGA datasets, associations between molecular features and patient survival given a certain drug were estimated as:

```
coxph(Surv(Time, Status) ~ [Tumor_stage] + [IHC] + Feature + Drug + Feature * Drug)
```

using variables:

`Feature` – an original molecular variable or NEA score for the given gene;

`Drug` – binary variable: if the drug was administered to the given patient;

`[IHC]` – TCGA immunohistochemistry parameters (whichever available for the cohort).

Paper III uses both data described above for Paper I and additional data from public datasets and publications. In this paper we also created a medulloblastoma (MB) meta-cohort out of several sources, such as publications(98–101) and datasets available online (PBCA-DE and PEME-CA projects). We retrieved exome sequencing profiles as alongside with, when available, copy number alterations, gene expression and clinical data. We translated gene identifiers into gene symbols according to ENSEMBL annotations v.93 and then made sure all the gene symbols are found in the network and were up to date according to GeneCards(102) annotations.

For consistency with the publication datasets, we excluded non-functional types of mutations from PBCA-DE and PEME-CA sets, namely: intron variant, upstream gene variant, exon variant, 3_prime_UTR_variant, 5_prime_UTR_variant, intergenic region, downstream gene variant, missense variant, synonymous variant, and splice region variant. For a few patient IDs that were found in more than one dataset, their mutation profiles were merged (if different) using logical union operation.

Overall survival data was collected from the published datasets. Several patients with discrepant data were excluded. For 18 samples with different follow-up, we accepted the newest survival time values.

Data from all the obtained datasets were combined into one meta-cohort of 541 patients, covered with both clinical and exome sequencing data.

Paper IV uses GO and DO data and a number of various networks. Total 17 networks, presumably holding different topological properties, were tested, based on the following original networks: FC3, STRING, PWC, FC2, FC1, FClim. There were two methods for generating a new network:

1. Merging given network with PWC.

2. Edge filtering – subnetworks were created from networks by taking top N edges (sorted by confidence score).

Table 2 (corresponds to Table 1 in the original paper) demonstrates used networks and some of their characteristics. Number of components is measured as a number of disconnected components ("weak" mode in the respective function of igraph(103) package).

Table 2. Networks used in Paper IV. FC = FunCoup, STRING = STRING v. 10.5, PWC = Pathway Commons v9 (corresponds to Table 1 in Paper IV)

Network ID	Network name	Provenance	Number of edges	Number of nodes	Number of	Network diameter
FC3-1M, FC3	FunCoup v. 3 (104)	Most confident edges ranked by FunCoup confidence score	1,000,000	11,691	97	15
FC3-300K	Same as above		300,000	6,880	122	14
FC3-100K			100,000	4,160	104	12
FC3-30K			30,000	3,018	49	13

FClim	FunCoup-2014 (8)	Edges ranked with FunCoup confidence score > 3.5	911,327	14,586	121	10
FC2	FunCoup v. 2 (105)	Most confident edges ranked by FunCoup confidence score	911,327	12,638	377	13
FC1	FunCoup v. 1 (106)	Same as above	911,327	14,490	21	10
STRING	STRING v. 10.5 (107)	Most confident edges ranked by STRING score	1,000,000	17,977	4	9
PWC	Pathway Commons v. 9 (108)	All interacting gene and protein pairs downloadable from PathwayCommons database	755,608	18,550	3	6
FC3 & PWC		Merge of FC3 and PWC	1,739,208	20,503	4	6
FC2 & PWC		Merge of FC2 and PWC	1,679,650	22,187	149	9
FC1 & PWC		Merge of FC1 and PWC	1,682,108	22,842	11	8
STRING & PWC		Merge of STRING and PWC	1,599,902	21,098	4	6
FClim & PWC		Merge of FClim and PWC	1,679,192	22,097	35	8

3.3 Network enrichment analysis

The main focus of the study is network enrichment and, in particular, previously developed NEA method. All the papers included in this thesis focus primary on NEA's ability to discover novel genes for different purposes (Paper III), dedicated to practical applications of NEA and additional tools in everyday research process or clinical pipeline (Papers I and II) or studying fundamental properties of the developed algorithm, comparing it to other core methods and attempting to offer potential improvements (Paper IV). In the paper, not included in the scope of this thesis (see "Scientific papers not included in the thesis") NEA was used as an additional tool for estimating enrichment between the discovered biomarkers and well-known pathways affected in prostate cancer.

Network enrichment (as defined by NEA) between two gene sets of interest S_a and S_b is estimated by comparing the actual number of network edges $\mathcal{E}_{S_a \leftrightarrow S_b}$ that connect nodes of S_a with nodes of S_b in the real, biological network $G_B=(E,V)$ (defined by set of edges E and set of vertices V) with a number of connections expected by chance $\hat{\mathcal{E}}_{S_a \leftrightarrow S_b}$ in a random network $G_R=(E,V)$ where particular node degrees k of genes $\forall g_i \in \boldsymbol{S_a}$; $\forall g_j \in \boldsymbol{S_b}$; $g_i \not\equiv g_j$ are equal to those of the actual network (which implicitly assumes that the whole degree sequences of G_B and G_R are identical, too). In an earlier work (70), series of randomized instances of G_R were created using an algorithm of explicit edge permutation and used for estimating expected variance of ε. Later, it was demonstrated that $\hat{\mathcal{E}}_{i \leftrightarrow \mathbf{GS}}$ can be calculated analytically in a fast and unbiased manner:

$$\hat{\mathcal{E}}_{S_a \leftrightarrow S_b} = \left(\sum_{i=1}^{|S_a|} k_i * \sum_{j=1}^{|S_b|} k_j\right) \Big/ 2|E|$$

Then the difference between the actual and expected edge counts

$$\Delta\mathcal{E} = \mathcal{E}_{S_a \leftrightarrow S_b} - \hat{\mathcal{E}}_{S_a \leftrightarrow S_b}$$

is used to estimate significance of the relation $S_a \leftrightarrow S_b$ with a $\chi 2$ statistic:

$$\chi^2 = \frac{\Delta\varepsilon^2}{\hat{\varepsilon}_{S_a \leftrightarrow S_b}} + \frac{\Delta\varepsilon^2}{|E| - \hat{\varepsilon}_{S_a \leftrightarrow S_b}}$$

The χ2 value can be conveniently converted to p-values and then to Z-scores which could be used in downstream calculations in the same way as conventional, gene level variables.

In the simplest NEA case one of the sets is a single gene *i*:

$$\hat{\varepsilon}_{i \leftrightarrow \mathbf{S}} = (k_i * \sum_g k_g)/2|E|;\ \Delta\varepsilon_i = \varepsilon_{i \leftrightarrow \mathbf{S}} - \hat{\varepsilon}_{i \leftrightarrow \mathbf{S}};$$

$$\chi^2 = \frac{{\Delta\varepsilon_i}^2}{\hat{\varepsilon}_{i \leftrightarrow \mathbf{S}}} + \frac{{\Delta\varepsilon_i}^2}{|E| - \hat{\varepsilon}_{i \leftrightarrow \mathbf{S}}}$$

At the moment, NEA implementations work in 2 modes: direct and indirect. Direct mode takes into account only direct (first order) neighbours, while indirect works with paths of length 2 (second order neighbours) relative to AGS.

In Paper I NEA implementation used as a back-end for the interactive web-platform. Visualization tools offer user a way to obtain graphical representation of the performed analysis in addition to table form.

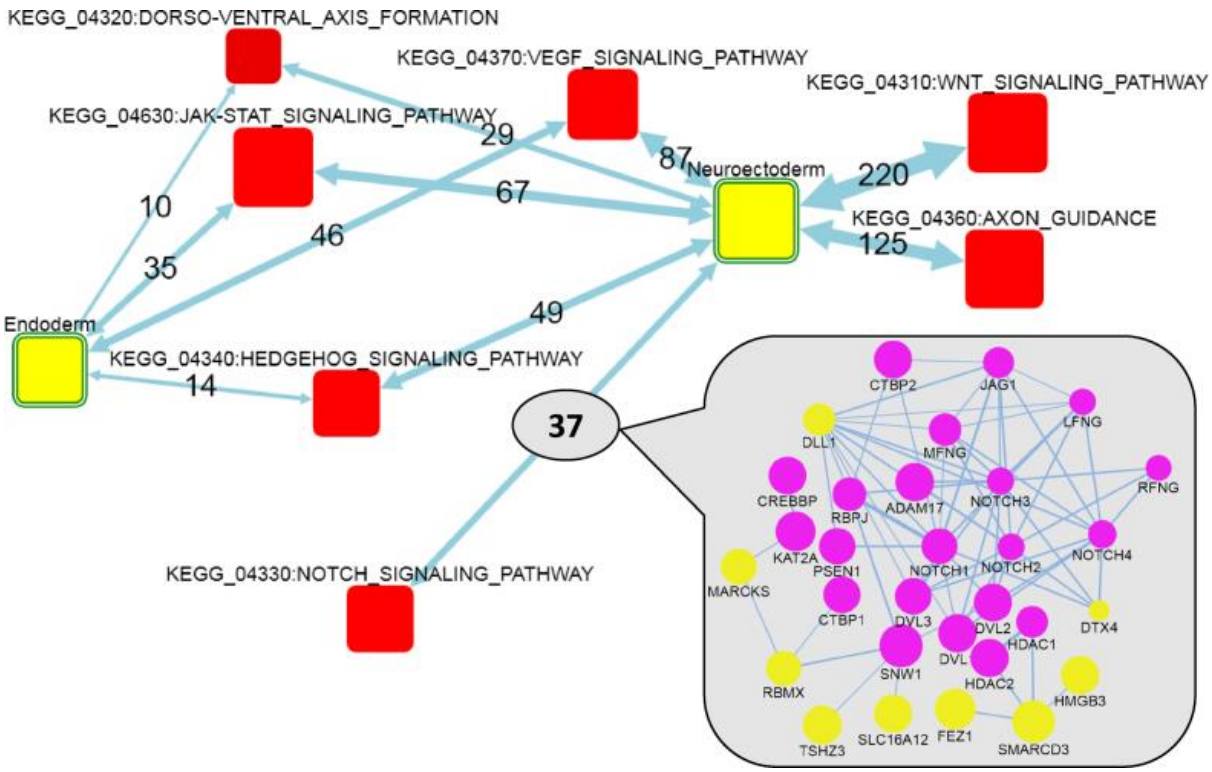


Figure 2. Example of NEA results visualization.

In Paper II pre-computed NEA results are available for user analysis and exploration, embedding into own models and visualizing them using tools developed in Paper I. NEA analysis was performed on 1655 selected pathways from MSigDB and FC3 network for cell lines (in case of CCLE) or individual patients (in case of TCGA). Two data types were considered separately: point mutations (MUT) and gene expression (GE), producing datasets NEA-MUT and NEA-GE respectively. For MUT, all mutations, disregarding their types, were taken as AGS for patient/cell line. For GE, each AGS was constructed as a list of top 100 genes with expression different from the respective cohort mean.

In Paper III NEA was used as one of the channels of evidence for discovering novel driver genes, named MutSet. It was combined with other evidence channel, named PathReg. PathReg used vectors `anchor.summary`, calculated for each pathway individually, for obtaining regression models (thus, this channel also relies on NEA). Specific NEA scores were calculated for every gene *i* present in the network (*N*=19035) versus every MGS in the given cancer cohort *c*. The `anchor.summary` values μ_{ic} were then obtained by summing up over all N_c available samples, regardless of mutation status of *i* in genome *j*:

$$\mu_{ic} = \sqrt{\log \frac{\sum_{j=1}^{j \le N_c} Z_{i \leftrightarrow MGS_j}}{N_c}};$$

Since the score $Z_{i \leftrightarrow MGS_j}$ is derived from the network patterns of mutated genes across the cohort and does not depend on the mutation profile of *i* itself, the μ_{ic} value would reflect a general propensity of *i* to interact with constellations of putative cancer genes. The transformations via $\chi 2 \rightarrow \mathrm{Z}$, log, and square root were imposed in order to render distributions closer to Gaussian.

The μ_{ic} profiles were rather scarce due to rare occurrence in mutated gene set (MGS) of true drivers that would interact with a given gene *i*. We thus further improved the gene specific values via modelling μ_{ic} with pathway NEA scores $Z_{i \leftrightarrow FGS}$. These were calculated for 320 FGS versus each of the *N* network genes and then used as a matrix of dependent variables **Φ** in `PathReg` model training. Sparse regression models were created using function `cv.glmnet` from R package `glmnet(109)`. The chosen package implements elastic net models for solving the problem:

$$\min_{\beta_0, \beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[\frac{(1-\alpha)\|\beta\|_{l2}^2}{2} + \alpha \|\beta\|_{l1} \right],$$

where α is a mixing parameter for balance between lasso and ridge regression (whereby α=0 and α=1 would lead to plain ridge and lasso regressions, respectively). In our case (α=1), `glmnet` solved just the lasso problem:

$$\min_{\beta_0, \beta} R_\lambda(\beta_0, \beta) = \min_{\beta_0, \beta} \left(\frac{1}{N} \sum_{i=1}^{N} (\mu_{ic} - \beta_0 - \beta^T \Phi) + \lambda \|\beta\|_{l1} \right)$$

Later, `MutSet` and `PathReg` evidence channels were combined. Both `MutSet` and `PathReg` produced output in the form of *q*-values, equivalent to false discovery rate which conveys the probability of a given driver prediction to be false. These values were integrated into the final value as a product $q(M\&P)=q_{\texttt{MutSet}} * q_{\texttt{PathReg}}$, which presented the

probability that neither channel have produced true predictions. Therefore, 1 – *q(M&*P) was the probability of either channel to be true and we convened to trust a driver prediction if *q(M&*P) < *c* (*c*=[0.01, 0.05]).

Paper IV, in addition to two different NEA implementations (R package(110) and Perl script with additional parameter, allowing it taking only direct neighbours or second order neighbours into consideration), uses two other well-established algorithms: PageRank with Priors (package `igraph`) and Random Walk with Restart (package `dnet(111)`). For each network we calculated topological features, such as number of nodes and edges, number of disconnected components, network density, connectivity etc.. For each combination of network and pathway, pathway-level variables were calculated (number of nodes for the pathway in the given network, number of modules, average distance between disconnected modules, closeness weight etc.). For each available node in any given network topological features were calculated: both global (degree, closeness centrality, betweenness centrality, eigenvector centrality etc.) and local (modular) closeness centrality for every possible combination of network and pathway). After performing network analysis with different methods and parameters, we attempted to establish correlations between topological characteristics and assigned scores. We also explored how damping factor and restart probability affect results due to different topological properties of the chosen network(77,78). Since, as was previously mentioned, cancer is a very heterogeneous disease and disease genes are considered to be low-degree genes, we paid special attention to poorly studied genes, which we identified by their degree in network and GO codes (for GO).

We used area under curve (AUC), partial area under curve (pAUC at FPR=0.15), F1 and Matthew's correlation coefficient (MCC) in order to estimate performance for different combinations of network, ontology, method and method parameter (when available). Performance was tested for each pathway. All pathways were filtered: groups of less than 3 genes and more than 300 genes were excluded from test (amount of genes were calculated for each network independently). Then each pathway had been split into two subsets: *train*, consisting two thirds of the pathway genes, which were used as seed nodes (AGS), and test, containing all the remaining pathway genes.

Results of network enrichment performed by different methods were analyzed for correlation with different topological metrics on different levels:

1. On network level – we compared performance for different methods and ontologies between various networks to determine if any of the tested network demonstrates the best performance for all the tested methods.
2. On pathway level – we used term-level features, described above, for creating linear regression models designed to explain significance of tested variables and their effect on performance. In other words, we attempted to find single pathway-level variables and their combinations, which could explain performance.

On gene level – results of network methods are matrices of size $||N|| \times ||P||$, where N is the set of nodes (genes) in the given network, P is the set of pathways in the given network under restrictions described previously. All results matrices were normalised with inverse normalisation procedure. For each pathway, correlation between vector of obtained values and vector of values for the chosen topological metric was calculated using implementation of Spearman method provided by standard function `cor` in R.

For creation of linear regression models, standard R function `lm` was used. Significance of model members was estimated using p-values produced by `lm` function. Sign of the member coefficient was used for assessing if it has positive or negative influence on performance. For each combination of network, ontology, method and parameter, three models were created, using different performance metrics as dependent variables: AUC, F1, MCC. In addition to regression models, we used Kendall τ for establishing correlations between performance metrics and individual variables.

4 Results and discussion

The thesis is dedicated to the study of network methods (NEA in particular) in order to understand current possibilities – can they be effectively used nowadays or are they an instrument of the past, which never achieved its full potential? – and their applicability to the emerging field of precision cancer medicine. All the presented papers focus on different aspects of this task.

In Paper I the main result is the creation of a web platform which allows researchers to perform network enrichment analysis on user-defined lists of genes (AGS) and visualise its results. It is applied to on one of the available networks or on a merge of several networks with pre-existing or user-uploaded FGS. User-friendly solutions provide researchers with the ability to incorporate the web platform into their research pipeline due to different possibilities to export data from the platform: as table data in plain format or as high-quality vector graphics. There are several alternatives to this system like MaxLink, cBioPortal and STRING, but, arguably, they do not achieve the same quality of service from the user point of view.

MaxLink utilizes its own original method, though related to NEA. It uses curated FunCoup network. Just like EviNet, it offers statistical analysis, though more limited, but it works only with one network and provides much less options in terms of visualization. Recent versions of MaxLink added some domain-specific functionality, namely "SARS-CoV-2 Search", which can be handy for researchers working on the specific topics. EviNet, on the other hand, despite being developed with cancer driver discovery in mind, attempts to be a universal tool by 1) offering a large compiled database of various FGS describing biological processes and diseases and 2) offering a possibility to download own data. The aforementioned advantages of EviNet allowed us to seamlessly integrate it into our next project, EviCore, for some network-related functionality. MaxLink offers only online functionality, while EviNet relies on an algorithm implemented earlier in an R package, which can be downloaded to user's local machines or research servers.

cBioPortal, while being an excellent source of cancer genome data, offers no network analysis and no statistics. However, it provides richer API compared to EviNet, and offers packages for both R and Python. Data exploration is probably the strongest part of cBioPortal, an issue (for EviNet) which we addressed in our next project, described in Paper II.

Last but not least, STRING has better network visualization – namely, it demonstrates different types of edges in multigraphs, such as "experimentally determined", "co-expression", "predicted by gene co-occurence etc.". The portal allows uploading of user data, but only after registration. It offers a variety of tools, but the entry threshold is higher compared to EviNet, since the site requires user to prepare data in specific format and offers a variety of narrow-tasked tools, abundance of which can confuse user. The latter feature makes STRING potentially more powerful tool for an experienced user, but EviNet offers a simple user interface coupled with interactive demos, thus minimizing time required to start conducting actual research with the chosen web platform.

Of course, there are more alternatives to the developed platform. We acknowledge potential weaknesses of EviNet, but we consider it a valuable instrument for researchers with minimum or no experience in bioinformatics and/or programming. This project should be considered together in complex with the other papers presented in this study, especially Paper II and Paper III.

Paper II demonstrates, among other project features, how to use data obtained from NEA for creating predictive models, e.g. predicting overall survival. Extending the results also to those in Paper III, we demonstrate that NEA results could be effectively combined with other omics or clinical variables in order to obtain more accurate models. The developed platform includes integration with the platform developed in Paper I and allows users to incorporate it into their own research pipelines: users can

visualize and explore data (including pre-calculated correlations and NEA results), create predictive models and download the desired results as vector graphics, table data, JSON or RData files, to embed illustrations into own web-based resources. For advanced users, we provide documented REST API, which they can use to create their own tools employing some functionality of the developed platform.

Just as the web platform developed in Paper I, EviCor definitely has some rivals in terms of functionality, such as CellMiner or OncoMine. However, according to our experience, the combination of functions provided by the developed platform is unique. One of the important features is data accessibility, unlike some alternatives, EviCor does not require registration and provides vast, documented and flexible API (which is an improvement compared to the previous project, EviNet). Just as EviNet, EviCor is intended to be as user-friendly as possible. The main advantages of EviCor are 1) inclusion of a unique data collection, containing both raw data from popular online databases as well as pre-processed data and statistical estimations of correlations between omics data and drug sensitivity and 2) ability to combine different data types for creating and exploring multivariate models. The latter function is unique and very useful in the complex field of oncology. For example, as epigenetics is an emerging yet understudied topic, provided operability allow researchers to create models incorporating both clinical variables and epigenetic profiles of selected genes for survival analysis.

Paper III describes the NEAdriver project, presenting proof that NEA could be used for discovering novel driver genes in pan-cancer cohorts. NEA proved to be a useful tool for cancer types possessing relatively few mutations but with a very large heterogeneity, such as medulloblastoma (MB), where common frequency-based methods are impossible to use. Not only did our approach make it possible to recover previously known driver genes, but we also managed to identify several pathways potentially important for tumorogenic processes and/or potential therapeutic targets. Moreover, we discovered that some of the significantly enriched genes in the MB cohort, namely HDAC1 and HDAC2, were never mutated in any samples. We discovered in the littaerature that these genes are proposed as therapeutic targets (112,113), signifying the

role of NEA as a tool for in silico analysis preceding in vitro and in vivo experiments. However, there are also some potential weaknesses with the NEA driver approach. One of them actually derives from advantages of the method: in cases with rare and understudied mutations, additional experiments are required. Results of such in silico experiments should always be taken with a grain of salt, since there is a risk that artefactual results can be produced by some combinations of topological features of the used network and the enrichment method, which will not reflect any actual biological truth. This issue is addressed in Paper IV. Another limitation of NEAdriver lies in area of "standalone" driver genes, which can trigger cancer development on their own. Thus they can be undetectable by all GBA methods, but can be discovered by frequency-based analysis (but only if sufficient amount of data is supplied) or function-based analysis of altered sequences. The aforementioned methods, however, have problems of their own. Frequency-based methods are not only dependent on the amount of collected samples in order to achieve statistical significance, but also prone to false-positive results, reporting often occurring passenger mutations as drivers due to methods' inability of functional assessment of alterations. On the other hand, methods based on functional assessment of alterations based on established biological knowledge are complex and, more importantly, undermine identification of lesser studied genes (this problem is partially addressed in Paper IV). Paper III is also connected with Paper I via NEA, being a proof-of-concept: researchers who want to employ NEA in their own analysis can either use standalone R package or the EviNet web platform.

In Paper IV we analysed various networks, topological properties of pathways belonging to different ontologies in these networks and compared NEA with other methods under different conditions. One of the main results is a proof that NEA is unbiased towards any topological characteristics of networks, pathways and individual nodes. We discovered that random walk methods can be more sensitive to certain centrality measures than to node degrees under certain circumstances and demonstrated that "local" topology (that is split of certain set of nodes into a number of modules in a certain networks) can drastically change the results. These results suggest that there is no "silver bullet" for adjustment of results produced by random walk methods. For example, normalization by node degrees does not solve the problem of centrality biases and parameter values

naturally reduces the bias towards some metrics while increasing score correlation with the others. According to our benchmarking on some of the networks (such as merges of FC3 and PWC or STRING and PWC) all the explored methods demonstrate better performance compared to smaller and sparser networks/THAN WHAT?/ (measured by different metrics), confirming the earlier suggested idea of correlation between network size and performance. For random walk methods, we attempted to find optimal parameters based on several performance metrics, and, in most cases, we obtained results quite different from parameters suggested by default by the explored methods. We measured difference in performance by AUC and, while in some cases difference in performance was neglectable, in some cases it was as big as 0.7. Analysis of created models indicated that the significance of certain pathway-level variables for random walk methods is considerably affected by parameter values. Of course, the provided study is not complete, some of the problems were not solved yet – such as the correlation between results produced by different methods and modular betweenness centrality. However, according to our knowledge, this is the most complete study of the influence of various topological properties of networks on the use of different network methods for biological purposes. Nevertheless, the results of Paper IV support and extend the results of Paper I (by highlighting strong and weak sides of the underlying algorithm of NEA) and Paper III (by demonstrating the ability of NEA to correctly retrieve rare and understudied genes, which was proved on certain GO codes and DO entries).

Overall, our study proves that NEA can effectively be used for various biomedical and clinical research tasks, which is specifically demonstrated in Papers III and IV. However, NEA has certain limitations, some of them are common to all GBA methods and cannot be fixed, while some of them has to be addressed in future research (see "Future perspectives"). But the method cannot be employed and offered as a viable technology without 1) comparison with other methods and 2) developing tools for the researchers.

In a narrow sense, we compare NEA only with other network methods, such as RWR and PPR. Despite the fact that under certain conditions aforementioned methods demonstrate better performance, but arguably the biggest problem is inability to pick the optimal parameter for the best performance, as highlighted in Paper IV. In a broader

sense, we have to compare NEA with methods based on completely different principles. Such comparison was done in Paper III as explained earlier (i.e. comparison with frequency-based methods). However, we do not claim that NEA is a "silver bullet". Instead, we seek a way to incorporate NEA into existing research pipelines and combine with other methods. As was mentioned in the beginning of this section, network analysis could be considered as an outdated tool by some, in this era of emerging AI. In Paper II we demonstrated how data generated by NEA can be combined with different machine learning techniques, such as regression models. It was mentioned before that pathway-level overview for cancer can offer a broader perspective and better understanding of cancer biology. Moreover, from machine learning point of view using NEA results means reduction of dimensionality: from tens of thousands genes we can shift to thousands of pathways or even just to several dozens pathways of interest.

Even the best method is worthless unless it is implemented and offered as a convenient tool. We addressed this issue in Papers I and II, attempting to offer best possible experience for users varying from researchers with no bioinformatics experience to bioinformaticians seeking ways to incorporate some parts of the developed platforms into their own projects. None of the developed portals is absolutely unique, but nonetheless they offer a number of advantages compared to the competitors. Alongside with statistical and mathematical proofs from Papers III and IV portals EviNet and EviCor offer solid options for practical research problems.

5 Conclusions

Network medicine is a relatively new branch of medicine, which itself is young, compared to mathematics. Personalized and precision treatment methods is a future of oncology. They will allow oncologists to create an effective treatment plan for each patient based on individual characteristics and with minimal side effects. Unlike radical mastectomy, which attempted to cure cancer at all costs, novel treatment methods will also take patient quality of life into consideration. Network medicine at the same time gives us more general understanding of cancer biology (by allowing exploration of cancer at pathway-level instead of at the single-gene level) and more detailed view (through discovery of novel drivers and therapeutic targets).

Some types of cancer have a high mutational burden, while others have high inter-patient heterogeneity and low mutational burden. In Paper III, not only do we demonstrate the ability of the developed method to discover novel drivers, but also the ability of network methods to surpass performance of other methods, such as frequency-based methods. Frequency-based methods are extremely ineffective for predictions of drivers in some types of cancer, such as medulloblastoma.

All presented papers prove that NEA can be effectively used (Papers III and IV) and provide user-friendly web tools for performing NEA (Papers I and II). The developed tools cover broad needs of the researcher-users: from exploring pre-downloaded and pre-processed data to statistical and network enrichment analysis and uploading results in widely used formats. All the data is publicly available, all the code is available on GitHub under open software licenses, APIs of the developed platforms are documented – thus all the performed research is verifiable, and the developed code is available for the future programmers to incorporate into their own research products.

6 Future perspectives

Despite the fact, that NEA has a number of advantages, it also has limitations. As was mentioned in Paper IV, at the moment NEA can take into account neighbourhoods of 1st and 2nd orders (relative to AGS). This limitation is artificial, future modifications of NEA will have to solve this issue. Several NEA modifications were introduced in Paper IV already, allowing to combine scores obtained by NEA runs with different parameters.

Another possible modification of NEA comes from PPR: different nodes belonging to AGS could have different weights in the beginning, thus simulating situation when researcher want to prioritize certain genes in AGS (e.g. by frequency of mutations in cohort) or probability of belonging to the established drivers.

In Paper III we performed an analysis on pan-cancer cohort, which included medulloblastoma patients. It is required to try to apply the developed algorithm to other types of cancer, especially rare and/or having low mutational burden.

Last but not least, in Paper IV we proved that NEA had no bias towards topological properties of individual nodes. However, additional study may be required, since we were able to cover only limited number of metrics. It is highly unlikely that we will discover a certain metric towards which NEA is biased, however novel metrics, which can be used to correct results (of NEA or other considered methods), can be proposed. In addition, there could be proposed an algorithm to combine NEA results with other methods, as it was done in Paper III. In addition to exploring properties of network methods, topological properties of networks and biological terms (such as signaling pathways) should be studied in more systematic way: there could potentially be a method to determine optimal parameter for certain method based on properties of the network and seed nodes in it.

7 Acknowledgements

First of all I want to thank my main supervisor Andrey Alekseenko for introducing me to the research field and supporting my studies through-out with patience. I also want to thank my supervisors, professor Ingemar Ernberg, for his continuous support and kind advice, and professor Erik Aurell, for offering thoughtful comments on role of mathematics in biology; professor Lars-Gunnar Larsson for organizing oncology seminars, which allowed the author to obtain the required level of understanding; Matti Nikkola, thanks to whom author had an opportunity to share his knowledge in AI and computer science with KI students.

The author would like to express his gratitude to Ingemar Ernberg group, who supported his research and gave constructive feedback during the entire period of performing this study, especially to Benedek Bozoky, with whom author performed a study on AI in cancer biomarker discovery and who introduced me to the clinical environment; to Vitaliy Grozman, for his explanations on needs of researchers at hospital side Special thanks to Kenneth W. Kinzler, Director at Ludwig Center at Johns Hopkins University, for permission to use illustration referred as Figure 1 in this paper.

8 References

1. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011 Mar 4;144(5):646–74.

2. Yeh AC, Ramaswamy S. Mechanisms of Cancer Cell Dormancy – Another Hallmark of Cancer? Cancer Res. 2015 Dec 1;75(23):5014–22.

3. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Yang THO, et al. The Immune Landscape of Cancer. Immunity [Internet]. 2018 Apr 5 [cited 2018 Apr 10];0(0). Available from: http://www.cell.com/immunity/abstract/S1074-7613(18)30121-3

4. Gonzalez H, Hagerling C, Werb Z. Roles of the immune system in cancer: from tumor initiation to metastatic progression. Genes Dev. 2018 Oct 1;32(19–20):1267–84.

5. Hanna TP, Nguyen P, Baetz T, Booth CM, Eisenhauer E. A Population-based Study of Survival Impact of New Targeted and Immune-based Therapies for Metastatic or Unresectable Melanoma. Clinical Oncology. 2018 Oct 1;30(10):609–17.

6. Chiappinelli KB, Zahnow CA, Ahuja N, Baylin SB. Combining Epigenetic and Immune Therapy to Combat Cancer. Cancer Res. 2016 Apr 1;76(7):1683–9.

7. Orlando D, Miele E, Angelis BD, Guercio M, Boffa I, Sinibaldi M, et al. Adoptive Immunotherapy Using PRAME-Specific T Cells in Medulloblastoma. Cancer Res. 2018 Jun 15;78(12):3337–49.

8. Merid SK, Goranskaya D, Alexeyenko A. Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis. BMC Bioinformatics. 2014 Sep 19;15:308.

9. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science. 2013;339(6127):1546–58.

10. Gröbner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, et al. The landscape of genomic alterations across childhood cancers. Nature. 2018 Mar;555(7696):321–7.

11. Banu MA, McKhann GM. Order in Chaos: Understanding Intratumoral Heterogeneity in Gliomas by Tracking Tumor Cell Fate. Neurosurgery. 2018 Jan 1;82(1):N4–6.

12. McGranahan N, Swanton C. Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution. Cancer Cell. 2015 Jan 12;27(1):15–26.

13. PCAWG Evolution & Heterogeneity Working Group, PCAWG Consortium, Gerstung M, Jolly C, Leshchiner I, Dentro SC, et al. The evolutionary history of 2,658 cancers. Nature. 2020 Feb;578(7793):122–8.

14. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014 Jan 5;505(7484):495–501.

15. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer genes. Nature. 2013 Jul 11;499(7457):214–8.

16. Grzywa TM, Paskal W, Włodarski PK. Intratumor and Intertumor Heterogeneity in Melanoma. Transl Oncol. 2017 Oct 24;10(6):956–75.

17. Pfister S, Remke M, Benner A, Mendrzyk F, Toedt G, Felsberg J, et al. Outcome prediction in pediatric medulloblastoma based on DNA copy-number aberrations of chromosomes 6q and 17q and the MYC and MYCN loci. J Clin Oncol. 2009 Apr 1;27(10):1627–36.

18. Smith JC, Sheltzer JM. Systematic identification of mutations and copy number alterations associated with cancer patient prognosis. Settleman J, editor. eLife. 2018 Dec 11;7:e39217.

19. Maciejowski J, Chatzipli A, Dananberg A, Chu K, Toufektchan E, Klimczak LJ, et al. APOBEC3-dependent kataegis and TREX1-driven chromothripsis during telomere crisis. Nat Genet. 2020 Sep;52(9):884–90.

20. Bolkestein M, Wong JKL, Thewes V, Körber V, Hlevnjak M, Elgaafary S, et al. Chromothripsis in Human Breast Cancer. Cancer Res. 2020 Nov 15;80(22):4918–31.

21. Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. Nat Rev Genet. 2006 Jan;7(1):21–33.

22. Cavalli FMG, Remke M, Rampasek L, Peacock J, Shih DJH, Luu B, et al. Intertumoral Heterogeneity within Medulloblastoma Subgroups. Cancer Cell. 2017 Jun;31(6):737-754.e6.

23. Wheeler DA, Wang L. From human genome to cancer genome: the first decade. Genome Res. 2013 Jul;23(7):1054–62.

24. Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nature Reviews Genetics. 2004 Feb;5(2):101.

25. Csermely P, Korcsmáros T, Kiss HJM, London G, Nussinov R. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. Pharmacol Ther. 2013 Jun;138(3):333–408.

26. Hu L, Huang T, Shi X, Lu WC, Cai YD, Chou KC. Predicting Functions of Proteins in Mouse Based on Weighted Protein-Protein Interaction Network and Protein Hybrid Properties. PLoS One [Internet]. 2011 Jan 19 [cited 2017 Aug 31];6(1). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3023709/

27. Zhang W, Chien J, Yong J, Kuang R. Network-based machine learning and graph theory algorithms for precision oncology. npj Precision Onc. 2017 Aug 8;1(1):1–15.

28. Schmitt T, Ogris C, Sonnhammer ELL. FunCoup 3.0: database of genome-wide functional coupling networks. Nucleic Acids Res. 2014 Jan 1;42(Database issue):D380–8.

29. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res. 2017 Jan 4;45(Database issue):D362–8.

30. Bader GD, Donaldson I, Wolting C, Ouellette BFF, Pawson T, Hogue CWV. BIND—The Biomolecular Interaction Network Database. Nucleic Acids Res. 2001 Jan 1;29(1):242–5.

31. Liu ZP, Wu C, Miao H, Wu H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database. 2015;2015:bav095.

32. Arita M. Scale-freeness and biological networks. J Biochem. 2005 Jul;138(1):1–4.

33. van Noort V, Snel B, Huynen MA. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO reports. 2004 Mar;5(3):280–4.

34. Ma HW, Zeng AP. The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics. 2003 Jul 22;19(11):1423–30.

35. Nacher JC, Akutsu T. Recent progress on the analysis of power-law features in complex cellular networks. Cell Biochem Biophys. 2007;49(1):37–47.

36. Clote P. Are RNA networks scale-free? J Math Biol. 2020 Apr;80(5):1291–321.

37. Lima-Mendez G, van Helden J. The powerful law of the power law and other myths in network biology. Mol Biosyst. 2009 Dec;5(12):1482–93.

38. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011 Jan;12(1):56–68.

39. Gursoy A, Keskin O, Nussinov R. Topological properties of protein interaction networks from a structural perspective. Biochemical Society Transactions. 2008 Dec 1;36(6):1398–403.

40. Chen L, Zhang YH, Huang T, Cai YD. Identifying novel protein phenotype annotations by hybridizing protein–protein interactions and protein sequence similarities. Mol Genet Genomics. 2016 Apr 1;291(2):913–34.

41. Hu L, Huang T, Liu XJ, Cai YD. Predicting Protein Phenotypes Based on Protein-Protein Interaction Network. PLoS One [Internet]. 2011 Mar 10 [cited 2017 Aug 31];6(3). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3053377/

42. Kim YA, Cho DY, Przytycka TM. Understanding Genotype-Phenotype Effects in Cancer via Network Approaches. Karchin R, editor. PLOS Computational Biology. 2016 Mar 10;12(3):e1004747.

43. Winter C, Kristiansen G, Kersting S, Roy J, Aust D, Knösel T, et al. Google Goes Cancer: Improving Outcome Prediction for Cancer Patients by Network-Based Ranking of Marker Genes. PLOS Computational Biology. 2012 May 17;8(5):e1002511.

44. Ghulam A, Lei X, Guo M, Bian C. Disease-Pathway Association Prediction Based on Random Walks With Restart and PageRank. IEEE Access. 2020;8:72021–38.

45. Huan T, Wu X, Bai Z, Chen JY. Seed-weighted random walk ranking for cancer biomarker prioritisation: a case study in leukaemia. Int J Data Min Bioinform. 2014;9(2):135–48.

46. Pathan M, Keerthikumar S, Ang CS, Gangoda L, Quek CYJ, Williamson NA, et al. FunRich: An open access standalone functional enrichment and interaction network analysis tool. Proteomics. 2015 Aug 1;15(15):2597–601.

47. Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, et al. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol. 2012;13(12):R124.

48. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, et al. A systems biology approach for pathway level analysis. Genome Res. 2007 Oct;17(10):1537–45.

49. Hornberg JJ, Bruggeman FJ, Westerhoff HV, Lankelma J. Cancer: A Systems Biology disease. Biosystems. 2006 Feb 1;83(2):81–90.

50. Eroles P, Bosch A, Pérez-Fidalgo JA, Lluch A. Molecular biology in breast cancer: intrinsic subtypes and signaling pathways. Cancer Treat Rev. 2012 Oct;38(6):698–707.

51. Murillo-Garzón V, Kypta R. WNT signalling in prostate cancer. Nat Rev Urol. 2017 Nov;14(11):683–96.

52. Brennan C, Momota H, Hambardzumyan D, Ozawa T, Tandon A, Pedraza A, et al. Glioblastoma subclasses can be defined by activity among signal transduction pathways and associated genomic alterations. PLoS ONE. 2009;4(11):e7752.

53. Yoshida GJ. Regulation of heterogeneous cancer-associated fibroblasts: the molecular pathology of activated signaling pathways. J Exp Clin Cancer Res. 2020 Jun 16;39(1):112.

54. Yuan M, Zhao Y, Arkenau HT, Lao T, Chu L, Xu Q. Signal pathways and precision therapy of small-cell lung cancer. Signal Transduct Target Ther. 2022 Jun 15;7(1):187.

55. Yoshimoto K, Mizoguchi M, Hata N, Murata H, Hatae R, Amano T, et al. Complex DNA repair pathways as possible therapeutic targets to overcome temozolomide resistance in glioblastoma. Front Oncol [Internet]. 2012 Dec 5 [cited 2021 Feb 13];2. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3514620/

56. Ellison DW, Dalton J, Kocak M, Nicholson SL, Fraga C, Neale G, et al. Medulloblastoma: clinicopathological correlates of SHH, WNT, and non-SHH/WNT molecular subgroups. Acta Neuropathol. 2011 Mar;121(3):381–96.

57. Creixell P, Reimand J, Haider S, Wu G, Shibata T, Vazquez M, et al. Pathway and Network Analysis of Cancer Genomes. Nat Methods. 2015 Jul;12(7):615–21.

58. Di J, Zheng B, Kong Q, Jiang Y, Liu S, Yang Y, et al. Prioritization of candidate cancer drugs based on a drug functional similarity network constructed by integrating pathway activities and drug activities. Mol Oncol. 2019 Oct;13(10):2259–77.

59. Franco M, Jeggari A, Peuget S, Böttger F, Selivanova G, Alexeyenko A. Prediction of response to anti-cancer drugs becomes robust via network integration of molecular data. Scientific Reports. 2019 Feb 20;9(1):2379.

60. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014 Jan;42(Database issue):D472-477.

61. Rodchenkov I, Babur O, Luna A, Aksoy BA, Wong JV, Fong D, et al. Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. Nucleic Acids Res. 2020 Jan 8;48(D1):D489–97.

62. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the Pathway Interaction Database. Nucleic Acids Res. 2009 Jan;37(Database issue):D674-679.

63. Köhler S, Bauer S, Horn D, Robinson PN. Walking the Interactome for Prioritization of Candidate Disease Genes. Am J Hum Genet. 2008 Apr 11;82(4):949–58.

64. Erten S, Bebek G, Ewing RM, Koyutürk M. DADA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization. BioData Min. 2011 Jun 24;4:19.

65. Le DH, Kwon YK. Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization. Computational Biology and Chemistry. 2013 Jun 1;44:1–8.

66. Jin W, Jung J, Kang U. Supervised and extended restart in random walks for ranking and link prediction in networks. PLoS One. 2019 Mar 20;14(3):e0213857.

67. Li Y, Li J. Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data. BMC Genomics. 2012 Dec 7;13(Suppl 7):S27.

68. Page L, Brin S, Motwani R, Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. 1999.

69. Komurov K, Dursun S, Erdin S, Ram PT. NetWalker: a contextual network analysis tool for functional genomics. BMC Genomics. 2012;13(1):282.

70. Alexeyenko A, Lee W, Pernemalm M, Guegan J, Dessen P, Lazar V, et al. Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. BMC Bioinformatics. 2012;13:226.

71. Huang JK, Carlin DE, Yu MK, Zhang W, Kreisberg JF, Tamayo P, et al. Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. cels [Internet]. 2018 Mar 28 [cited 2018 Apr 4];0(0). Available from: http://www.cell.com/cell-systems/abstract/S2405-4712(18)30095-4

72. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network Motifs: Simple Building Blocks of Complex Networks. Science. 2002 Oct 25;298(5594):824–7.

73. Schreiber F, Schwöbbermeyer H. Motifs in biological networks. In 2010.

74. Mavroforakis C, Mathioudakis M, Gionis A. Absorbing random-walk centrality: Theory and algorithms. arXiv:150902533 [cs] [Internet]. 2015 Sep 8 [cited 2017 Aug 14]; Available from: http://arxiv.org/abs/1509.02533

75. Borgatti SP, Everett MG. A Graph-theoretic perspective on centrality. Social Networks. 2006 Oct 1;28(4):466–84.

76. Perra N, Fortunato S. Spectral centrality measures in complex networks. Phys Rev E. 2008 Sep 5;78(3):036107.

77. Boldi P, Santini M, Vigna S. PageRank As a Function of the Damping Factor. In New York, NY, USA: ACM; 2005 [cited 2017 Aug 28]. p. 557–66. (WWW '05). Available from: http://doi.acm.org/10.1145/1060745.1060827

78. Bressan M, Peserico E. Choose the damping, choose the ranking? Journal of Discrete Algorithms. 2010 Jun 1;8(2):199–213.

79. Avrachenkov K, Litvak N, Pham KS. A Singular Perturbation Approach for Choosing the PageRank Damping Factor. Internet Mathematics. 2008 Jan 1;5(1–2):47–69.

80. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25–9.

81. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2020 Dec 8;49(D1):D325–34.

82. Schriml LM, Arze C, Nadendla S, Chang YWW, Mazaitis M, Felix V, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012 Jan 1;40(D1):D940–6.

83. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res. 2015 Jan 28;43(D1):D1071–8.

84. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19(1A):A68-77.

85. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019 May 8;1.

86. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. Cancer Discovery. 2015 Nov 1;5(11):1210–23.

87. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell. 2016 Jul;166(3):740–54.

88. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000 Jan 1;28(1):27–30.

89. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res. 2002 Jan 1;30(1):42–6.

90. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Research. 2023 Jan 6;51(D1):D587–92.

91. Nishimura D. BioCarta. Biotech Software & Internet Report. 2001 Jun;2(3):117–20.

92. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: Pathway Editing for the People. PLoS Biology. 2008 Jul 22;6(7):e184.

93. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Research. 2014 Jan;42(D1):D459–71.

94. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015 Dec 23;1(6):417–25.

95. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015 Jan 28;43(Database issue):D447–52.

96. Alexeyenko A, Sonnhammer ELL. Global networks of functional coupling in eukaryotes from comprehensive data integration. Genome Res. 2009 Jun;19(6):1107–16.

97. Alexeyenko A, Schmitt T, Tjärnberg A, Guala D, Frings O, Sonnhammer ELL. Comparative interactomics with Funcoup 2.0. Nucleic Acids Res. 2012 Jan;40(Database issue):D821–8.

98. Pugh TJ, Weeraratne SD, Archer TC, Pomeranz Krummel DA, Auclair D, Bochicchio J, et al. Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. Nature. 2012 Jul 22;488(7409):106–10.

99. Jones DTW, Jäger N, Kool M, Zichner T, Hutter B, Sultan M, et al. Dissecting the genomic complexity underlying medulloblastoma. Nature. 2012 Jul 25;488(7409):100–5.

100. Northcott PA, Buchhalter I, Morrissy AS, Hovestadt V, Weischenfeldt J, Ehrenberger T, et al. The whole-genome landscape of medulloblastoma subtypes. Nature. 2017 Jul 20;547(7663):311–7.

101. Robinson G, Parker M, Kranenburg TA, Lu C, Chen X, Ding L, et al. Novel mutations target distinct subgroups of medulloblastoma. Nature. 2012 Jun 20;488(7409):43–8.

102. Stelzer G, Dalah I, Stein TI, Satanower Y, Rosen N, Nativ N, et al. In-silico human genomics with GeneCards. Hum Genomics. 2011 Oct;5(6):709–17.

103. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Systems. 2006;1695.

104. Schmitt T, Ogris C, Sonnhammer ELL. FunCoup 3.0: database of genome-wide functional coupling networks. Nucleic Acids Res. 2014 Jan 1;42(Database issue):D380–8.

105. Alexeyenko A, Schmitt T, Tjarnberg A, Guala D, Frings O, Sonnhammer ELL. Comparative interactomics with Funcoup 2.0. Nucleic Acids Research. 2012 Jan 1;40(D1):D821–8.

106. Alexeyenko A, Sonnhammer ELL. Global networks of functional coupling in eukaryotes from comprehensive data integration. Genome Res. 2009 Jun;19(6):1107–16.

107. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015 Jan 28;43(D1):D447–52.

108. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011 Jan;39(Database issue):D685-690.

109. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010 Feb 2;33(1):1–22.

110. Jeggari A, Alexeyenko A. NEArender: an R package for functional interpretation of 'omics' data via network enrichment analysis. BMC Bioinformatics [Internet]. 2017 [cited 2019 Nov 6];18(118). Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1534-y

111. Fang H, Gough J. The `dnet' approach promotes emerging research on cancer patient survival. Genome Medicine. 2014 Aug 26;6(8):64.

112. Coni S, Mancuso AB, Magno LD, Sdruscia G, Rotili D, Mai A, et al. Selective inhibition of HDAC1 and HDAC2 counteracts medulloblastoma cell growth in mouse models through Gli acetylation. European Journal of Cancer. 2016 Jul 1;61:S143–4.

113. Pei Y, Liu KW, Wang J, Garancher A, Tao R, Esparza LA, et al. HDAC and PI3K Antagonists Cooperate to Inhibit Growth of MYC-driven Medulloblastoma. Cancer Cell. 2016 Mar 14;29(3):311–23.