DOI:10.15774/PPKE.ITK.2016.008

APPLICATION OF GRAPH MODELS IN BIOINFORMATICS

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

BALÁZS LIGETI

Roska Tamás Doctoral School of Sciences and Technology
Pázmány Péter Catholic University,
Faculty of Information Technology and Bionics

Supervisor:
Prof. Sándor Pongor

Budapest, 2016

DOI:10.15774/PPKE.ITK.2016.008

Abstract

Biomedical sciences use a variety of data sources on drug molecules, genes, proteins, complete genomes sequences, diseases and scientific publications, etc. This system can be best pictured as a giant data-network linked together by physical, functional, logical and similarity relationships. A new hypothesis or discovery can be considered as a new link that can be deduced from the existing connections. For instance, interactions of two pharmacons – if not already known – represent a testable novel hypothesis. Such implicit effects are especially important in complex diseases such as cancer. Currently huge amount of data is generated by experiments, such as whole genome sequencing of metagenomic data. Deriving new information, i.e. linking the experiments to microorganisms and supporting the new hypothesis with known data requires the proper analysis of a data-network.

The goal of the investigations carried out in this thesis is to predict novel drug combinations or novel biomarkers using the network of existing oncological and protein interaction databases and to interpret and analysis large metagenomic data using network principles.

I showed that the overlap of network neighborhoods is strongly correlated with the pairwise interaction strength of two pharmacons used in cancer therapy, and it is also well correlated with clinical data.

The strategy based on the hypothesis that novel, implicit links can be discovered between the network neighborhoods of data items lead to the discovery of novel biomarkers based on text analysis. In 2012 I prioritized ten potential biomarkers for ovarian cancers, two of which were in fact described as such in the subsequent years.

I showed that applying network principles and fast aligners in the evaluation of metagenomic whole genome sequencing experiments could improve the classification performance, and even sensitive detection of pathogens is possible.

The strategy seems to hold promises in several applications including prioritization of new drug combinations, discovering of novel biomarkers for experimental testing or sensitive detection of pathogens. Its use is naturally limited by the sparsity and the quality of experimental data; however, these aspects are expected to improve given the development of current databases.

DOI:10.15774/PPKE.ITK.2016.008

Acknowledgements

I would like to express my deepest gratitude to my advisor, Professor Sándor Pongor, for his valuable guidance and support, and that he made me possible to work on a topic of great interest to me. This thesis would have not been finished without his endless encouragement. I am also grateful to Professor Tamás Roska for his inspiration and that he made me believe I have that certain marshal's baton as well.

My sincere appreciation is extended to Professor Péter Szolgay for his unfailing support. This work would not exist without the support of the Faculty of Information Technology and Bionics, Pázmány Péter Catholic University.

I am also grateful to Balázs Győrffy and Gergely Lukács for their advices, comments and critics.

I also want to express my deepest appreciation to Roberto and Lőrinc: our work together provided the stimulating atmosphere (filled with a lot of fun) every scientist needs sometimes.

Here I have to mention my fellow PhD students and colleagues: Zsolt, Dóri, Endre, István, Zoltán, Norbi, Bence, Tamás, Miklós, Berci, Domokos, and many others. Having lunch together and discussing the big questions of life was always a refreshing moment of the day.

Last but not the least, I would like to thank my family: Noémi, for being my wife, my colleague, my secretary and my editor; my children, Emese, Alíz and Ágoston, for being the steady point in my life during these years.

A special thanks goes to Rufus: the one purring on my shoulders while writing this.

DOI:10.15774/PPKE.ITK.2016.008

Table of Contents

DOI:10.15774/PPKE.ITK.2016.008

DOI:10.15774/PPKE.ITK.2016.008

List of Abbreviations

AC	Drug regimen consisting of doxorubicin and cyclophosphamide
Bp	Base pair
BWT	Burrows-Wheeler transforms
CAF	Drug regimen consisting of 5-fluorouracil, doxorubicin and cyclophosphamide
CLI	Command line interface
CMF	Drug regimen consisting of 5-fluorouracil, methotrexate and cyclophosphamide
DBMS	Database management system
EC	Drug regimen consisting of epirubicin and cyclophosphamide
EGFR	Epidermal growth factor receptor
ER	Entity–relationship (model)
FAC	Drug regimen consisting of 5-fluorouracil, doxorubicin and cyclophosphamide
FEC	Drug regimen consisting of 5-fluorouracil, epirubicin and cyclophosphamide
GO	Gene ontology
HER2	Erb-b2 receptor tyrosine kinase 2
LCA	Lowest common ancestor
NCCN	National comprehensive cancer network
OC	Ovarian cancer
TAC	Drug regimen consisting of 5-fluorouracil, doxorubicin and cyclophosphamide
TOS	Target overlap score
WGS	Whole-genome shotgun
XML	Extensible markup language

DOI:10.15774/PPKE.ITK.2016.008

List of Tables

DOI:10.15774/PPKE.ITK.2016.008

List of Figures

DOI:10.15774/PPKE.ITK.2016.008

Ad maiorem Dei gloriam

DOI:10.15774/PPKE.ITK.2016.008

1. Introduction

The network view on biological data has profoundly influenced the ways we are looking at problems of diagnosis and therapy in life sciences today. In traditional paradigms, we used to look at data as isolated entities stored in organized databases. Today, we increasingly consider data as an interconnected network. There are many kinds of connections – for instance, drugs can be connected to diseases, to their protein targets, to genes producing the targets, or to drugs they can replace or antagonize. In a similar manner, proteins can be linked to other proteins they physically contact, to genes they regulate, to diseases they play a role in, etc. This is a very complex picture, because we have many types of entities and relationships that are defined in separate ontologies that in turn can be considered as networks of terms. The storage and manipulation of such a large body of data is clearly too demanding for current computers. In addition, such data networks are both incomplete and noisy. Namely, we have a seemingly large number of proteins, but the knowledge on proteins is rarely validated by experiment, and a large part of the annotations is just taken over from homologous proteins of various organisms. In addition, we cannot be sure whether two proteins are linked in all tissues and/or in all phases of the cell cycle. The solution of these problems is to break down the hypothetical data-network into specific - disease-specific, tissue-specific, pathogen-specific, etc. - manually curated parts that contain reliable information on a given problem. This tedious and labor-intensive solution is justified only in very important fields. Cancer-specific data networks are an example of this approach. In addition, two major information sources can help data-sparsity problems. On the one hand, various high-throughput experimental methods (two hybrid systems, DNA sequencing, Chip-seq, etc.) provide novel kinds of molecular interaction data that in principle can be easily added to the existing databases. However, high throughput data are most often laden with noise, which has to be handled. In such cases, hierarchical data networks (i.e. ontologies) may offer a good framework to balance between the reduction of noise and sensitivity to discover novel data links from the experiments.

On the other hand, literature databases that contain abstracts or full text of scientific papers provide a large body of new knowledge that can in principle be linked to molecular data. Again, the process is not trivial: scientific texts use natural language and concepts are often not analogous to the ones used in other texts or in molecular databases.

DOI:10.15774/PPKE.ITK.2016.008

Disease-specific databases and tools represent a current approach where the above problems are tackled by large communities of scientists. Cancer databases and tools are a typical example, since cancer is one of the most important complex diseases which is responsible for ~15% of all human deaths, and which has >100 more-or-less well-characterized types and >500 human genes associated with it [2]. Oncologists use a variety of traditional databases, but there are a number of data-collection efforts dedicated to the gathering of data on various cancer types. All this provides a solid knowledge base for designing integrated data-networks in which novel questions related to cancer therapy can be answered.

Here I am concerned with three types of questions that can be addressed via integrated data networks: i) finding drug combinations potentially useful for cancer therapy. I tackle this problem by using a simple network overlap measure applied to data networks; ii) finding novel gene-disease associations in ovarian cancer for generating a list of potential biomarkers. I approach this problem using a text mining approach applied to MEDLINE abstracts [3] as well as the STRING database [4]. iii) Finally, I present a practical application by testing a dedicated, data subnetwork in accelerating and improving the taxonomic identification. Here I take advantage of the fact that the taxonomic and even the functional subnetworks are hierarchical graphs, which allows a substantial speedup with respect to current algorithms. **Section 1** is an introduction to the problem of biological topics discussed in this work, and it covers the mathematical and computational background as well. **Section 2** presents the main database types used in this project. **Section 3** describes the principle of hypothesis generation via network overlap analysis, the identification of drug combinations via a network overlap measure and the prediction of pathogenic species. **Section 4** discusses conclusions and future trends while summarizing the scientific results.

1.1. Biological background

1.1.1. Chemotherapy

Chemotherapy is the most frequently used first-line treatment of cancer. Chemotherapeutic agents target all dividing cells in the body either by killing them (cytotoxic agents) or by blocking proliferation without cell elimination (cytostatic agents), regardless of their status as normal or neoplastic. Tumor cells proliferate rapidly, thus agents selectively damaging dividing cells exhibit a selective advantage. Victims of such a universal destruction are the fast-growing normal cells,

DOI:10.15774/PPKE.ITK.2016.008

accounting for the side effects of chemotherapy such as damaged hair follicles, irritated epithelium of the mouth and digestive tract, and suppression of myelopoietic precursors in the bone marrow.

Chemotherapeutic agents can be classified according to the mechanisms of their action. Drugs can destruct the structure of DNA, stop metabolic processes, and obstruct protein structures of the mitotic spindle. Cell cycle consists of four different phases: G1 (protein synthesis and cell growth), S (DNA replication), G2 (further protein synthesis and cell growth) and M (mitosis) – some agents are cell-cycle-phase-specific, while other agents require cell proliferation for action but are not linked to any given phases of the cell cycle [5]. Chemotherapeutic agents can be classified into five main categories.

1) Alkylating agents are not cell-cycle-phase specific, and their effects are dose-dependent, thus cell killing is a linear function of the applied dose of the medication. They form covalent bond with amino, sulfhydryl, phosphate and carboxyl groups to alkylate biologically active molecules and block the function of DNA, but also RNA and proteins [5]. The group consists of nitrogen mustards, platinum agents, nitrosoureas and cyclophosphamides. Nitrogen mustards are similar to mustard gas and are mainly effective in the hematopoietic system [6], while the lipid soluble nitrosoureas used to target brain tumors penetrate through the blood-brain barrier [5]. Carboplatin is a standard agent of care for ovarian cancer [7-9].

2) Antitumor antibiotics have been isolated from natural sources, such as plants, bacteria and fungi. Antibiotics intercalate between DNA base pairs, thus inhibit transcription and RNA synthesis. Their effectiveness is limited by dose-dependent cardiotoxicity as a main adverse effect [10]. Frequently used antibiotics are actinomycin-D, mitoxantron, and anthracyclines such as doxorubicin. Anthracyclines also inhibit topoisomerases I and II.

3) Antimetabolites are structurally similar analogues of naturally occurring molecules. They interfere with metabolic processes by either competing for key enzymes or substituting components of DNA during synthesis, thus block cell cycle in the S phase. Antimetabolites show a nonlinear dose-response, thus after a given concentration no further cells are eliminated. Methotrexate inhibits folate biosynthesis, ultimately leading to purine and pyrimidine depletion within the cell [11]. Nucleoside analog 5-fluorouracil and cytarabin interfere with pyrimidine synthesis, while mercaptopurin, azathioprine, pentostatin and thioguanin hamper purine production.

DOI:10.15774/PPKE.ITK.2016.008

4) Vinca alkaloids and taxanes consist of cell-cycle-phase specific antimicrotubule blocking chemotherapy agents. During the S phase vinca alkaloids bind to tubulin, prevent polymerization and eventually mitotic spindle formation. Taxanes on the other hand, such as paclitaxel and docetaxel, stabilize tubulin inhibiting depolymerization and cell division [12].

5) Topoisomerase inhibitors, such as camptothecin analogs (irinotecan) inhibit DNA elongation by blocking topoisomerase I in the S phase of the cell cycle [13].

Response to chemotherapy is classified as complete (tumor is untraceable), partial (50% shrinkage) or minimal (stable disease). When chemotherapy fails, tumor progression continues. Chemoresistance is a complex multifactorial phenomenon [14, 15]. Mechanisms of resistance include pharmacological factors such as inadequate drug concentrations due to low accessibility of the tumor. Cellular resistance factors include detoxifying or transport mechanisms reducing drug concentrations in the target cell, altered drug-target interactions including the ability of the cells to repair damaged DNA, tolerate stress and evade apoptotic death [16-20]. Inherited genetic variability also influences susceptibility to chemotherapeutic agents. Single nucleotide polymorphisms (SNPs) have also been linked to altered drug response [21]. The one-gene one-drug approach with relevance to cancer chemotherapy has been gradually replaced by studying genetic variation on entire biological or pharmacological pathways, such as the complex network underlying folate metabolism [22] or enzymes responsible for detoxification [23].

Combination chemotherapy blends cytotoxic drugs with different mechanisms of action. The goal is to eliminate a broader range of resistant cells in the heterogeneous population of cancerous cells, to prevent or slow the emergence of resistant clones, and to maximize the additive or synergistic effects of drugs on cell kill. Compelling evidence support combination treatments over sequential monotherapy [24]. Preferable combinations include drugs with different mechanisms of action, such as paclitaxel with cisplatin, and different pattern of resistance [5]. When applied sequentially, the order of combined agents influences responses. For example, carboplatin followed by docetaxel in advanced non-small-cell lung cancer patients suggested higher response rate when compared to reverse arrangements [25].

DOI:10.15774/PPKE.ITK.2016.008

1.1.2. Systemic therapy of recurrent or metastatic breast cancer

In 2012 alone over 1.7 million women were diagnosed with breast cancer being the most common cancer in women [26]. High numbers pose economic burden and affect the quality of life of an enormous population. The universal goal to increase treatment efficiency is not trivial, as breast cancer is a heterogeneous disease. Based on molecular features breast cancers are grouped into subtypes with distinct gene expression pattern comprising luminal A, luminal B, basal like and HER2 positive subtypes [27]. Each of these phenotypes requires different management. The picture is further complicated with cancer stage and menopausal status. Local treatment of primary breast cancer differs from the systemic treatment of advanced or metastatic disease. Preoperative, so-called "neo-adjuvant" treatments, such as anthracyclines or endocrine agents given preoperatively are expected to downstage the disease. Advanced incurable malignancies require a sturdier cytotoxic treatment compared to a less serious disease. The guidelines of the US National Comprehensive Cancer Network suggest a list of preferred single agents for recurrent or metastatic breast cancer (that is not HER2-positive): doxorubicin or pegylated liposomal doxorubicin, paclitaxel, capecitabine or gemcitabine, vinorelbin or eribulin. Other single agent chemotherapies include cyclophosphamide, carboplatin, docetaxel, albumin-bound paclitaxel, cisplatin, epirubicin, ixabepilone. Chemotherapy combinations are listed in **Table 1.1**

Table 1.1. Chemotherapy combinations for recurrent or metastatic breast cancer adapted from the Guidelines of the US National Comprehensive Cancer Network

Regimen[1]	Component 1[2]	Component 2[2]	Component 3[2]
CAF/FAC	cyclophosphamide	doxorubicin	fluorouracil
FEC	fluorouracil	epirubicin	cyclophosphamide
AC	doxorubicin	cyclophosphamide	
EC	epirubicin	cyclophosphamide	
CMF	cyclophosphamide	methotrexate	fluorouracil
NA[3]	docetaxel	capecitabine	
GT	gemcitabine	paclitaxel	
NA[3]	gemcitabine	carboplatin	
NA[3]	paclitaxel	bevacizumab	

[1]The abbreviation of the given chemotherapeutic regimen (i.e. CAF/FAC is a drug combination consisting of cyclophosphamide, doxorubicin, fluorouracil etc.) [2]Name of the component in the drug combination. [3]Not available.

DOI:10.15774/PPKE.ITK.2016.008

1.1.3. Targeted molecular therapy

Unfolding the molecular mechanisms underlying neoplastic transformation [28] opened a new, “personalized” era in clinical practice. Identification of driver mutations [29] allowed the rational design of molecular-targeting agents (MTAs). MTAs as single or combination therapies aim at aberrations that appear in a broad range of cancers and can be targeted in many tumor cells simultaneously. Patients are eligible to a therapy with MTAs only if their cancer bears a driver mutation targeted by the given agent. Therapies include monoclonal antibodies (mAbs), that deplete growth factor supply for the cells or prevent receptor dimerization, and small-molecule inhibitors, that block the initiation of intracellular signal transduction or possess catalytic activities [30].

The efficacy of monotherapies using molecularly targeted agents is often inferior compared to combination strategies. The reason for this is that relatively few malignancies depend on only one unique pathway to achieve the malignant transformation. For instance, targeting the hyperactive ABL1 kinase with small molecule tyrosine kinase inhibitors, such as imatinib and nilotinib produced superior clinical outcome in chronic myeloid leukemia [31, 32]. The complexity of signaling pathways and the heterogeneity of tumors called forth the combination of MTAs and cytotoxic agents. In this, agents are selected based on biological considerations to alter complementary pathways of signal transduction or to inhibit multiple target molecules within the same pathway [33]. In general, MTAs are considered to be less toxic than conventional chemotherapies [34], but when combined, the crosstalk between pathways may result in unpredictable toxicities [35].

1.1.4. HER2 positive breast cancer

Evolution of treatment choice in HER2-positive breast cancer illustrates the difficulties in targeting complex biological systems. About 20% of breast cancer patients overexpress Epidermal Growth Factor Receptor 2 (HER2), facing aggressive tumor growth and inferior prognosis [36]. The first successful targeted therapy approved by FDA in 1998 was an anti-HER2 monoclonal antibody, trastuzumab, combined with chemotherapy. The treatment dramatically changed the clinical outcome of the aggressive HER2-positive metastatic breast cancer [37, 38]. Trastuzumab monotherapy was effective in about 15-26% of patients [39], and combining trastuzumab with chemotherapy provided significantly better outcomes [40].

DOI:10.15774/PPKE.ITK.2016.008

HER2 (ERBB2/neu) belongs to the family of type I receptor tyrosine kinases (RTKs) including EGFR (ERBB1), HER3 (ERBB3), and HER4 (ERBB4). HER2 is overexpressed in tumor tissue but not in healthy cells, hence offers an ideal target for personalized therapy. Ligand binding of RTKs – except HER2 with no known ligand - induces receptor homodimerization or heterodimerization at the plasma membrane. Dimerization activates complex signal transduction involving the PI3K/Akt, Ras/MAPK, and JAK/STAT pathways, leading to cell transformation and cancer. Ligand and heterodimer compositions tightly regulate downstream signaling. With its permanently open conformation, HER2 is a favored dimerization partner of the other RTKs conferring lateral transmission to create a complex network of signaling pathways [41].

Redundant signaling cascades, as in the case of EGFR receptor family, facilitate bypassing the targeted node in the network [42]. Eventually, about 70% of patients develop resistance against trastuzumab. In addition, more superior patient stratification will be needed to improve initial clinical response. Despite constant evaluation of predictive biomarkers, the extent of HER2 expression remains the sole reliable trait for treatment decision [43]. Improved outcome can be obtained in case the inhibition involves other members of the EGFR receptor family. Preferred first line treatment includes simultaneous treatment with pertuzumab and trastuzumab assisted either by docetaxel or by paclitaxel [44, 45]. Pertuzumab targets the second extracellular domain of HER2 and prevents its dimerization with HER3.

Following the most current NCCN guidelines, in case the preferred first line treatment cannot be implemented, the subsequent regimes should include the antibody-drug conjugate trastuzumab emtansine (T-DM1), consisting of trastuzumab covalently linked to a microtubule inhibitor [46]. Trastuzumab is also suggested to be utilized in combination either with paclitaxel and carboplatin, or with one of the following: docetaxel, vinorelbine, or capecitabine. Lapatinib, a small-molecule tyrosine kinase inhibitor blocks EGFR and prevents its dimerization with HER2. After trastuzumab failure, addition of lapatinib to chemotherapy improved post-progression free survival rates in metastatic breast cancer patients [47]. Trastuzumab combined with lapatinib offers a chemotherapy-free alternative to trastuzumab exposed HER2-positive breast cancer. Trastuzumab exposed HER2-positive breast cancer may also be treated with the combination of lapatinib and capecitabine, or trastuzumab and capecitabine. Trastuzumab can be combined with other single agents as long as anthracyclines are avoided due to increased cardiac cytotoxicity.

DOI:10.15774/PPKE.ITK.2016.008

1.1.5. Ovarian cancer

Ovarian cancer is the fifth most common cancer in women worldwide and the deadliest gynecologic malignancy, as less than 30% of patients with advanced disease reach 5-year survival [48]. It is characterized with extreme heterogeneity, as a considerable proportion of tumors does not originate from the ovaries [49, 50]. Their common features are their shared location and dissemination to the pelvic organs. Histological subtypes display cellular and molecular diversity and distinct pathogenesis. Type I tumors progress from benign precursor lesions and consist of low-grade serous, low-grade endometrioid, clear cell, mucinous and Brenner carcinomas with a distinct genetic profile and a low malignant potential. Type II tumors including high-grade serous, high-grade endometrioid, undifferentiated and mixed-mesodermal tumors are highly aggressive, genetically unstable, lack precursor lesions, frequently harbor p53 mutations (<90%), and approximately 20% of them carry BRCA1/2 mutations [51, 52].

About 80-85% of women are diagnosed with serous carcinoma, followed by endometrioid (10%), with clear cell and mucinous cancers being the least common subtypes. Clear cell and mucinous tumors appear in earlier stages more frequently than serous cancers, providing a generally good prognosis. Type I patients fare better after surgery and usually do not require chemotherapy [53]. However, the lack of symptoms at an early stage frequently results in a late diagnosis when the tumor has already reached an advanced state. The extent of surgical tumor-mass reduction is an important prognostic factor of survival. Complete resection of advanced tumors improves both progression-free and overall survival compared to suboptimal surgical outcomes [54]. Surgery complemented with adjuvant or neo-adjuvant carboplatin-paclitaxel treatment has been the standard of care in the past 15 years [55, 56]. In case of paclitaxel intolerance pegylated liposomal doxorubicin (PLD)-carboplatin or docetaxel-carboplatin treatments provide an alternative solution [57, 58]. Despite the good initial responses about 70% of patients relapse within the first three years. The relatively poor survival data compared to other types of solid tumors necessitated a more refined methodology. Ovarian histotypes are treated now as distinct diseases with different mutational profiles and treatment requirements, influencing early detection, clinical trial design and the identification of new drugable targets [59, 60].

Contrary to the good results associated with early stage mucinous cancers, women with advanced mucinous tumors do worse compared to other histological types of advanced disease related to the high frequency of platinum-resistance [61]. Mucinous tumors represent a distinct

DOI:10.15774/PPKE.ITK.2016.008

spectrum of ovarian cancers ranging from benign to invasive with an individual molecular profile featuring frequent KRAS but infrequent p53 and BRCA mutations [62, 63]. Oxaliplatin combined with 5-fluorouracil represents a promising alternative treatment specific to mucinous tumors, validated in vitro and on xenografts [64].

Ovarian clear cell cancers (OCCC) in advanced stages are also particularly malignant, and refractory to platinum-based chemotherapy [65]. Clean cell cancers are characterized by high frequency mutations in the PIK3CA catalytic subunit of the PI3K gene [66] and mutations in the chromatin remodeling ARID1A gene [67]. The gene expression profiles resembling renal clean cell cancers, such as MET overexpression and overactivation of IL6-STAT3-HIF signaling pathway, suggest that antiangiogenic treatment used on renal clean cell tumors, such as the multi-kinase inhibitor sunitinib, may be applicable to OCCC [68].

The high-grade serous ovarian cancer (HGSOC) is of particular interest, as it accounts for 70-80% of ovarian cancer fatalities. Large portion of HGSOCs originate from outside of the ovaries, from the distal part of fallopian tubes [50]. Despite its sensitivity to platinum derived medications and other DNA damaging agents, patient survival has not been improved for years, as therapies targeting specific tumor biomarkers were lacking. Transcriptional profiling separated mesenchymal, immune, differentiated and proliferative subtypes associated with different prognosis, although the distinction has not yet been translated to clinical decisions [69]. Sequencing HGSOC revealed a frequent driver p53 missense or nonsense mutation indicating its role in tumor initiation [70], and the inactivation of tumor suppressor genes RB1, NF1, RAD51B, and PTEN [71]. CCNE1 encoding cyclin E1 for cell cycle progression is amplified in a large proportion of HGSOC that lacks defects in the HR pathways, likely representing an early event in tumor progression [51]. Drugs targeting cells deficient in DNA repair, such as poly (ADP) ribose polymerase (PARP)-inhibitors, selectively kill tumor cells with dysfunctional BRCA1/2. Olaparib, the first PARP-inhibitor have been approved for use as a maintenance therapy in Europe and for advanced recurrent disease in the USA, to the great advantage of BRCA-related ovarian cancer patients, particularly with a platinum-sensitive disease [72].

1.1.6. Future perspectives

Precision medicine targeting specific mutations has its limitations and the transcriptional targets of key driver genes are still elusive [73]. The initial enthusiasm seems to dampen as long-

DOI:10.15774/PPKE.ITK.2016.008

term survival data have emerged with limited success. For example, initially well responding patients with cutaneous melanomas treated with the BRAF inhibitor vemurafenib relapsed shortly after treatment [74]. ALK-positive lung cancer patients treated with crizotinib showed a 65% response rate – but the median duration of response was only 8 months [75]. In HER2-positive breast cancer, the majority of patients develop resistance within the first year of trastuzumab treatment [76-78]. Current guidelines suggest the simultaneous combination of molecularly targeted and immune checkpoint therapy [79]. The concept behind this approach is that T cells of the adaptive immune system show a remarkable ability to match the diversity and adaptability of tumors. Immune therapy can unleash T cells specific to many antigens present in the tumor by targeting a single immune checkpoint. In spite of promising ongoing studies, current results suggest durable tumor [80] inactivation only in a fraction of patients.

1.2. Metagenomics concepts

Metagenomics is the genomic study of microbial communities, i.e. microbes coexisting in the same space and time. The term *metagenome* was coined to designate the genome content of the whole community. The study of these communities is as old as microbiology itself, since at the beginning it was not yet possible to separate microbial samples into taxonomic units, and later studies relied mostly on morphological features. Without going into details of this long and impressive development, one can concentrate on the current trends of metagenomic analysis today as summarized by a recent review of Escobar-Zepeda et al. [81]. First, sequencing technologies allow us to investigate the species as well as functional (gene) variety present in microbial samples. Species composition is easier to determine. We can analyze the 16S RNA genes of bacterial species according to the classical concept of Carl Woese. In this case, we can use amplicon sequencing of a short segment of bacterial genomes and compare it with 16S RNA databases. In this analysis, first step is assembling the whole sequence of the relatively short 16S RNAs present in the sample. This analysis gives an estimate of the known bacterial species present in a bacterial community and estimates unknown taxonomic units. 16S RNA analysis carries out a search in a collection of short sequences with sensitive alignment techniques, like BLAST [82]. This category is considered as the standard method, although it has its limitations. For instance, there is an essential need for PCR amplification, which comes with extra overhead and experimental bias. Word-based techniques and artificial intelligence may offer a solution by making it possible to build databases of clade-specific recognizers this way facilitating the usage of rapid string matching techniques [83].

DOI:10.15774/PPKE.ITK.2016.008

Another approach is to use whole genome shotgun sequencing, i.e. a full-scale sequencing of the entire sample. Complete assembly of the reads into full genomes is not possible – there are simply too many species in the sample. Partial assembly of genome fragment is possible with current, high coverage sequencing and in-depth data processing, but the analysis is never complete. Instead, the individual sequencing reads are simply mapped onto the genome of known bacterial species, a technique known as binning. This analysis often gives more detailed species composition as 16S RNA sequencing, but it tends to be limited to microbial species with known genomes. There are two main categories of computational approaches addressing the above problem. The first one is the group of marker-based methods, which aims to overcome the difficulties by reducing the search space, namely by using small, specified datasets. MetaPhlAn applies a small database of clade-specific sequence markers built from the genome sequences of the known taxa and then searching this with general-purpose aligners. Within large microbial communities, this kind of search can be exceptionally fast and accurate for identifying taxa and their approximate proportion. However, because of the fact that many strains often share identical markers, this approach many times lacks lower (e.g. strain-level) identification. This is a problem in the case of pathogenic strains of bacteria like *E. coli.*

Taxon assignment [84], which is one of the most critical points for all approaches, is mainly achieved by different kind of lowest common ancestor searches in a taxonomic hierarchy. This means that the reads are assigned to one taxon (e.g. an *E. coli* strain), if possible, or to more than one taxon (e.g. 100% identity with an *E. coli* strain and an *E. fergusoni* strain), but in this case the lowest common taxonomic ancestor (the genus *Escherichia*) is reported. Numerous programs use this approach, like MEGAN [83, 85], Mothur [86] and SOrt-ITEMS [87].

The above-discussed tools represent a high variability in the computational approaches meaning that there is a demand for further improvements. For instance, current tools being developed for general purposes means that there is a need for specified tools. Furthermore, qualitative and quantitative answers cannot always be clearly separated: e.g., *E. coli* reads in an output may express the presence of it in the sample; however, the abundance of the species can not be measured by the number of identified reads. At this moment, only MetaPhlAn is able to supply reliable quantitative results for species abundance [88].

The third direction of metagenome analysis is directed towards the functional analysis of genes present in metagenomics samples. The principle is based on whole genome sequencing, but

DOI:10.15774/PPKE.ITK.2016.008

the read to genome hits are evaluated in terms of gene groups, rather than in terms of strains and taxa. The analysis of gene function is based on the classification schemes such as COGSs (mentioned below).

1.3. Networks in biology

Biological databases, including the ones related to cancer therapy and metagenomics, contain annotated data items cross-referenced to each other. In the mathematical sense, such an entity can be pictured as a subgraph or subnetwork, in which some of the edges (cross references) point to other entities or subgraphs defined in other databases. For instance, a drug in the drug interaction database can be linked to another drug item within the same database, as well as to a disease defined in a medical ontology [89, 90], a protein defined in Uniprot, etc. In principle, there is no problem to represent all such subgraphs in one large network that we term here a data network. The advantage of such a network is that it allows a large variety of queries to be answered within the same system. In practice, the construction of such a large network is prohibitively difficult. First it would be far too large, second it would contain a large number of heterogeneous and partly conflicting data types [91]. The current solution is to build partial networks that allow one to answer a few questions related to a given project.

From the practical point of view, cancer data networks consist of, on the one hand, dedicated cancer-related sequence databases, and on the other hand, molecular and molecular interaction databases that include drug and drug interaction databases. The former ones are collected by focused next generation sequencing projects carried out by an often large number of research groups (**Table 1.2**). Such projects contain data on cancer mutations, and are often divided into type-specific datasets or comprehensive datasets. Another subgroup of these databases are data resources that are made available via WWW interfaces and include dedicated search facilities.

Molecular and molecular interaction databases used to build cancer data network consist of those datasets that help one to describe and interpret cancer-related sequence information. These databases can be roughly categorized as 1) general-purpose sequence databases, 2) drug-related databases, 3) molecular interaction databases and 4) literature databases.

A wide range of experimental methods used to study molecular interactions fall into two broad categories: i) Traditional methods of molecular biology focus on functionally proven

DOI:10.15774/PPKE.ITK.2016.008

interactions and try to gather fine details by studying the interacting partners with methods like x-ray crystallography [92, 93], nuclear magnetic resonance [94, 95], often in conjunction with structural bioinformatics and/or conventional biochemical methods. Interaction data of a selected protein can be collected with methods such as affinity chromatography or co-immunoprecipitation [80, 96, 97]. These are typically "small-scale" (focusing only on very few molecules) and traditional biochemical methods. ii) Large-scale or system-level approaches can be used to collect a large number of interaction data in one experiment. One of the best known methods for detecting protein-protein interactions is the yeast two-hybrid system [98]. The underlying idea is that the expression of the reporter genes depends on two separate components, a binding domain (BD) and an activation domain (AD). If the two domains are indirectly connected via a protein-protein interaction, where one of the interaction partner is fused with BD and the other fused to the AD, then one can detect the reporter gene. This approach makes it possible to detect a large number of interactions by screening a certain protein against a DNA library representing all possible proteins the organism can have. Another system-level technique, proteomics, can be used to study post-translational modifications or protein-protein interactions via affinity purification coupled with mass spectrometry (AP-MS). This approach can also be useful for detecting strong connection between proteins, thus exploring protein complexes [99]. High-throughput methods are productive but there are several drawbacks and biases – among others, the number of erroneous interaction assignments can exceed 10 percent.

In addition to experimental methods, the body of databases available in other fields is also a source of information. While experiments provide data on the biological entities themselves, the databases provide information on a wide variety of concepts. In this way we broaden the scope of molecular interaction data to "data networks" that allow us to link biological data to the results of further scientific fields. For instance, a drug database such as Drugbank [100] provides information on chemical structures and their biological targets (proteins and genes) and/or the diseases. A database of scientific publications, on the other hand, provides information on a large class of descriptions (scientific abstracts) that are linked to each other by common keywords, authors, statements etc. Further examples for special data network are the ontologies. These are special, hierarchical knowledge representations; for example, the Anatomical Therapeutic Chemical Classification (ATC) System classifies drugs into groups at five levels in a hierarchical way. Thus the classification system can be seen as a simple ontology, more specifically a forest (disjoint union of trees). The roots of the individual trees are the first level characters/classes of the ATC system

DOI:10.15774/PPKE.ITK.2016.008

and the leaves are the full ATC codes (7 characters). There are 14 main groups at this level such as code A (Alimentary tract and metabolism), code B (Blood and blood forming organs),

General-purpose databases such as Uniprot [101], Ensemble [102] or RefSeq [103, 104], GenBank [105] hold high quality and reliable information about proteins and genes (focusing on the amino acid or nucleotide sequence, protein names or descriptions, and citation information). Usually they provide data mining tools and APIs as well.

Drugbank database [100] is one of the most comprehensive and freely available, complex data source about drugs. Currently, it holds information about 2200 FDA approved and more than 6000 experimental drugs. It also provides detailed information about the food-drug and drug-drug interaction information. The information was manually curated from web resources and published papers and has been continuously developed [106, 107]. It also provides data about drug mechanism of action and drug labels and ADMET (drug metabolism, absorption, distribution, metabolism, excretion and toxicity) profile, thus the drug card of Drugbank could be a rich source of text mining.

TTD database [108] is tailored to peptide molecules and its target information. It also includes information about diseases and drug combinations, however the last one is only available as excel tables, but not in a structured format, such as XML. Both Drugbank and TTD contains manually curated data.

STICH [109-111] is an automatically created, integrated database. It was created by using similar concepts as those of the STRING network. The database focuses on small molecules and their relations to other small molecules and proteins. Similarly to the STRING database there are various types of associations between the molecular entities. It mainly contains protein-chemical and chemical-chemical links based on text mining and other complex predictions extended with chemical structure description strings.

The Drug Combination Database [112, 113] focuses on agents combined together to achieve some therapeutically advantage over single agent drugs. Drug regimens are typically used in treating cancer and other complex diseases. The database is partly based on the FDA orange book [114], clinical trials (https://clinicaltrials.gov/), and publications. It also holds information about the individual drug components, such as ATC codes, target and cross references.

DOI:10.15774/PPKE.ITK.2016.008

Furthermore, it also provides annotations for drug combinations, such as possible mechanism of actions, interaction type, suggested doses, etc.

Drug side effects and drug interactions are often not covered in standard public databases. These kinds of data are available, for instance, in the SIDER database [115, 116], where the side effects are extracted from the drug labels (using controlled vocabulary such as UMLS [90]). A well-maintained collection of drug side effects is provided by the Tatonetti Lab [117].

Experimental results of protein-protein interaction measurements are deposited in various primary databases such as the Database of Interacting Proteins (DIP) [118], Biomolecular Interaction Network Database (BIND) [119], Molecular Interactions Database (MINT) [120-123], Biological General Repository for Interaction Datasets (BioGRID), Human Protein Reference Database (HPRD), IntAct Molecular Interaction Database [124].

The DIP database contains large number of manually curated and reviewed interactions from numerous species [118, 125]. It also provides some services and visualization tools for the available data [126], and a cytoscape plugin (MiSink) [127]. Different evidences for the interactions were integrated and considered manually.

Human Protein Reference Database (HPRD) [128] contains various types of data about proteins such as post-translational modification, known or predicted disease associations, cellular localization, tissue expression, mainly from publications. The data have also been reviewed by scientific experts. The database contains information about 30047 proteins and 41327 interactions among them.

Another important protein-protein interaction database is IntAct [124, 129-131], developed and maintained by the European Bioinformatics Institute (EBI), updated on regular basis. The interactions were partly curated from literature (14074 publication) in collaboration with the Swiss-Prot team, or the data were submitted directly. They also use controlled vocabularies [132] (PSI-MI [133, 134], gene ontology [135] and NCBI taxonomy terms [136]) for annotating the interactions and the proteins. The database contains information about the interacting domains as well.

The information about the interactions is dispersed among different databases. Sometimes, these databases were curated/reviewed redundantly, so it is a natural need to make a standard data

DOI:10.15774/PPKE.ITK.2016.008

representation and data integration. The MiNTAct [137], Imex [138], Mentha [139] consortial databases integrate the molecular interaction data collected from 11 databases.

STRING (Search Tool for the Retrieval of Interacting Genes) is one of the largest integrated protein interaction databases, which covers 66.9 Mio predicted and known interactions between proteins of 1100 organisms. The majority of the interactions (44.1 Mio) are predictions. The links between the proteins are some kind of associations (among them several indirect ones) - not only physical interactions. The evidence types for the associations are neighborhood, gene fusion, co-occurrence, co-expression, experiments, databases, text mining, and homology. Each type of association has a confidence score, which is a probabilistic measure of the reliability of the link. The several types of links and their confidences can be combined into one association with one confidence score.

Transcription factor databases contain sequence motifs and genomic locations collected from genomic data using bioinformatics methods. In the network representation of the database the nodes are DNA motifs linked to genomic locations. A typical example of transcription factor databases is Transfac, first published by Edgar Wingender's group in 1994 [140]. The database is manually and continuously updated. The current release contains 7915 sites assigned to 6133 transcription factors. Further examples of this database are given in **Table 1.2.**

DOI:10.15774/PPKE.ITK.2016.008

Table 1.2. Cancer-related databases and resources

Database	Description	URL	Refs.
Comprehensive databases and resources			
TCGA	The Cancer Genome Atlas	http://cancergenome.nih.gov/	[141]
CGP	Cancer Genome Project	http://www.sanger.ac.uk/research/projects/cancergenome/	[142]
CPTAC	Clinical Proteomic Tumor Analysis Consortium	http://proteomics.cancer.gov/programs/cptacnetwork	[143, 144]
ICGC	International Cancer Genome Consortium	https://www.icgc.org/	[145]
Data mining resources			
COSMICMart	BioMart tool for COSMIC	https://cancer.sanger.ac.uk/cosmic/login	[146]
IntOGen BioMart	BioMart tool for IntOGen	http://biomart.intogen.org/	[147]
UCSC Cancer Genomics Browser	A visualization and analysis tool specialized to cancer data	https://genome-cancer.ucsc.edu/	[148, 149]
ICPS	An Integrative Cancer Profiler System	http://server.bioicps.org/	[143]
NCG 4.0	Network of Cancer Genes	http://ncg.kcl.ac.uk/	[150, 151]
CGWB	The Cancer Genome WorkBench	http://cgap.nci.nih.gov/cgap.html	[152]
CancerMA	A web-based tool for analyzing microarray data	http://www.cancerma.org.uk/information.html	[153]
ICPS	An Integrative Cancer Profiler System	http://server.bioicps.org/	[143]
Databases of genetic variations in cancer			
COSMIC	Catalogue of Somatic Mutations in Cancer	http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/	[154]
CaSNP	Cancer SNP data on CNAs	http://cistrome.dfci.harvard.edu/CaSNP/	[155]
DriverDB	Cancer driver genes and mutation database	http://driverdb.ym.edu.tw/DriverDB/intranet/init.do	[156]
IntOGen	Integrative Oncogenomics	http://www.intogen.org/	[157]
MoKCa	Mutations, Oncogenes, Knowledge & Cancer	http://strubiol.icr.ac.uk/extra/mokca/	[158]
CGAP	Cancer Genome Anatomy Project	http://cgap.nci.nih.gov/	[159]
Mitelman Database	Database of chromosome aberrations and gene fusions in cancer	http://cgap.nci.nih.gov/Chromosomes/Mitelman	[160]
CGC	The Cancer Gene Census	http://cancer.sanger.ac.uk/cancergenome/projects/census/	[161]

DOI:10.15774/PPKE.ITK.2016.008

Table 1.2. (Continued)

Databases of epigenetic, proteomic and transcriptome variations in cancer			
CanProVar	Human Cancer Proteome Variation Database	http://bioinfo.vanderbilt.edu/canprovar/	[162]
MethyCancer	A database of human DNA methylation and cancer	http://methycancer.psych.ac.cn/	[163]
CellLineNavigator	Expression profiles of cancer cell lines	http://www.medicalgenomics.org/celllinenavigator/	[164]
ITTACA	Integrated Tumor Transcriptome Array and Clinical data Analysis	http://bioinfo.curie.fr/ittaca	[147]
PubMeth	Cancer methylation database based on text-mining of PubMed	http://matrix.ugent.be/pubmeth/	[165]
OncomiRDB	A database of experimentally verified oncomiRs	http://bioinfo.au.tsinghua.edu.cn/member/jgu/oncomirdb/	[166]
Cancer-specific clinical and drug resources			
CancerDR	Cancer Drug Resistance Database	http://crdd.osdd.net/raghava/cancerdr/	[167]
HPtaa	The Human Potential Tumor Associated Antigen database	http://www.bioinfo.org.cn/hptaa/	[168]
CancerResource	A resource of cancer-relevant compound and protein interactions	http://bioinf-data.charite.de/cancerresource/	[169]
CanGEM	Cancer Genome Mine	http://www.cangem.org/	[170]
DTP	Anti-cancer agent database	http://dtp.nci.nih.gov/docs/cancer/searches/standard_mechanism.html	[171, 172]
ITTACA	Integrated Tumor Transcriptome Array and Clinical data Analysis	http://bioinfo.curie.fr/ittaca	[147]
Cancer-type-specific resources and databases			
RCDB	RCDB	http://www.juit.ac.in/attachments/jsr/rcdb/homenew.html	[173]
curatedOvarianData	Clinically annotated data for the ovarian cancer transcriptome	http://bcb.dfci.harvard.edu/ovariancancer/	[174]
PED	Pancreatic Expression Database	http://www.pancreasexpression.org/	[175]
HLungDB	Human Lung Cancer Database	http://www.megabionet.org/bio/hlung/	[176]

DOI:10.15774/PPKE.ITK.2016.008

Special types of molecular interactions are metabolic and signal transduction molecular interactions. One of the oldest pathway databases is KEGG [177]. However, the current version holds information related to pathways such as genome, diseases and related drugs. It provides a global map for each pathway.

Reactome [178], similarly to KEGG, is a comprehensive, manually curated, high quality pathway database with support of enrichment analysis and data visualization.

The Human Metabolome Database [179], however, concentrates on small molecule metabolites, and it is a rich source of biomarker discovery. It also provides enzymatic, biochemical, and clinical data.

The signaling and metabolic pathways are often handled as separate entities, however, crosstalks and regulatory coupling exist between the pathways [180]. The Signalink [181] and NDEx databases [182] not only offer manually curated and reviewed pathway information, but provide more context for pathway analysis such as transcriptional and post-transcriptional regulators.

Scientific literature databases contain data collected from scientific journals using increasingly automated electronic submission links. Medline/Pubmed [183] is perhaps the best known representative of public scientific literature databases, it collects scientific abstracts from the publishers and provides them with a unified system of keywords (mesh terms, reference [184]). In the network representation of the database, the nodes are scientific abstracts; the edges correspond to shared keywords, citation links (X cites Y), etc. The Medline database was first published in 1971 and it gained a very wide acceptance as it became available via the PubMed search facility in 1997. For machine learning purposes, the database is downloaded, and word combinations are identified via natural language processing techniques in order to create new index tables. Further examples of this database are given in **Table 1.3.**

DOI:10.15774/PPKE.ITK.2016.008

Table 1.3. Representative examples of molecular and molecular interaction databases relevant to cancer therapy

	Contents	URL	Ref.
1. General-purpose databases			
Uniprot	Comprehensive database of protein sequences	http://www.uniprot.org/	[101]
RefSeq	Genome sequences	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/	[185]
GenBank	Comprehensive database of genetic sequences	http://www.ncbi.nlm.nih.gov/genbank/	[105]
Ensemble	Comprehensive database of sequences with data mining tools	http://www.ensembl.org/index.html	[102]
2. Drug-related databases			
DrugBank	Drug data and drug-drug interactions	http://www.drugbank.ca/	[100]
Therapeutic Target Database (TTD)	Therapeutic Target Database	http://bidd.nus.edu.sg/group/cjttd/TTD_HOME.asp	[108]
STITCH	Drug molecular interactions	http://stitch.embl.de/	[109]
DCDB	Drug Combination Database	http://www.cls.zju.edu.cn/dcdb/	[186]
Offsides, TwoSides	Drug adverse effects and Drug-Drug Interactions	http://tatonettilab.org/resources/tatonetti-stm.html	[117]
Drugs.com	FDA approved drugs linked to diseases and target proteins/genes	www.drugs.com	
SIDER	Drug adverse effects	http://sideeffects.embl.de/	[115]
3. Protein /protein interaction databases			
DIP	Experimentally and manually validated molecular interactions	http://dip.doe-mbi.ucla.edu/dip/Main.cgi	[118]
HPRD (Human Protein Reference Database)	Experimentally and manually validated molecular interactions	http://www.hprd.org/	[128]
Intact	Manually curated molecular interaction	http://www.ebi.ac.uk/intact/	[124]
MIntAct	Manually curated integrated database	http://www.ebi.ac.uk/intact	[137]
STRING	Protein/protein interactions as well as connections derived from other databases	http://string-db.org/	[187]

DOI:10.15774/PPKE.ITK.2016.008

Table 1.3. (Continued)

4. Transcription factor databases			
TRANSFAC	Transcription factors and binding sites	http://www.gene-regulation.com/pub/databases.html	[140]
JASPAR	Transcription factor binding profile database	http://jaspar.genereg.net/	[188]
DBD	Transcription factor prediction database	http://www.transcriptionfactor.org/	[189]
5. Metabolic pathways			
KEGG	Kyoto Encyclopedia of Genes and Genomes	http://www.genome.jp/kegg/	[177]
Reactome	Curated pathway database	http://www.reactome.org/	[178]
MetaCyc	Metabolic pathway database	http://metacyc.org/	[190]
HMDB	Human Metabolome Database	http://www.hmdb.ca/	[179]
6. Signal transduction databases			
NetPath	Manually curated signal transduction pathways	http://www.netpath.org/	[191]
Signalink	Manually curated signal transduction pathways	http://signalink.org/	[181]
NDEx	Integrated network database	http://www.ndexbio.org	[182]
7. Mutation databases			
COSMIC	Somatic mutations found in human cancer	http://cancer.sanger.ac.uk/cosmic	[192]
OMIM	Disease gene and mutation database of humans	http://www.omim.org/	[193]
8. Literature databases			
PubMed/Medline	PubMed/Medline	http://www.ncbi.nlm.nih.gov/pubmed	[194]
EMBASE	Biomedical and pharmacological bibliographic database	http://store.elsevier.com/embase	
Scopus	Bibliographic database of peer-reviewed literature	http://www.scopus.com/	

DOI:10.15774/PPKE.ITK.2016.008

1.4. From databases to data networks

One way to picture data network construction is to take a database of cancer genes or proteins, and then cross-reference it to general-purpose sequence databases, drug-related databases, etc. that will form a network among various types of entities allowing the cross querying of diverse biological databases in a unified manner. In practice, the construction of such a large network is prohibitively difficult, partly because of the incompatibility of ontologies, partly because of the sheer size of the network [91]. The design of these data networks is largely facilitated by database frameworks capable to handle an arbitrary set of biological entities and relationships [195]. Physical or structural connections rely on the well-known fact that molecules practically never function alone but rather in association with other molecules such as ligands, lipids, amino acids, proteins and nucleic acids. On the one hand, there are structural associations between the elements that can be "strong" such as covalent bonding and tight associations in the cytoskeleton (microfilaments are polymers of G-actin proteins), or "transient" such as in the case of receptor-ligand associations. Understanding the nature and the type of these relationships is crucial for interpreting complex biological phenomena such as disease mechanisms. As an example, the active forms of proteins are most often complexes assembled from various types of other proteins or other types of molecules such as RNA, DNA or small molecules. On the other hand, there are functional associations, such as between members of signaling pathways, transcriptional or metabolic networks. Functional associations may not even involve structural interactions, for instance distant members of a metabolic pathway are functionally related. The common motif in these widely different scenarios are the links between molecules that can involve various structural and functional aspects. For instance, a transient interaction act in catalyzing sequential steps within a metabolic network, or in a signaling pathway such as modifying the protein by adding phosphate group, etc.

Biological databases, including the above examples, contain annotated data items cross-referenced to each other. In the mathematical sense, such an entity can be pictured as a subgraph or subnetwork, in which some of the edges (cross-references) point to other entities or subgraphs defined in other databases. For instance, a drug in the drug interaction database can be linked to another drug item within the same database, as well as to a disease defined in a medical ontology, a protein defined in Uniprot, etc. In principle, there is no problem to represent all such subgraphs in one large network, which we term here a data network – but such a network would be prohibitively complicated for practical uses. One of the solutions is warehousing, wherein

DOI:10.15774/PPKE.ITK.2016.008

databases are stored as parallel items within the same computer, and integrated concepts and new data types take care of appropriate matching of underlying entities and attributes, including the resolution of conflicts. The result of such a common representation can be best pictured as a network of data, where the original data items (say drugs, target proteins, diseases, mutations) are represented in a common large network. For instance, a network combining drug targets and protein-protein interactions will contain links (network paths) between proteins that are targeted by the same (or similar) drugs, or drugs acting on proteins that are in physical contacts. In such a data network all data items (say drugs, target proteins, diseases, and mutations) are connected via a variety of different links, which makes processing complicated and time-consuming. As a practical workaround, one can construct a dedicated database tailored to a specific task, and that can be queried with simpler tools [195].

From the practical point of view, it is useful to distinguish comprehensive resources that aim to cover, for instance, all known genes and proteins and one selected type of interaction (say, regulatory connections). On the other hand, specialized resources concentrate on a selected species (Homo sapiens), or on a selected tissue type, or on a selected mechanism (signal transduction, or protein kinases). A few representative examples of databases are listed in **Table 1.2.**

1.5. Graph – definitions and notations

From the logical point of view, all interaction networks and data networks are graphs in which nodes are entities, such as molecules, diseases, i.e. biological, physical, as well as conceptual objects, while the edges or links between nodes are relationships, such as molecular interactions, drug-disease connections, drug compatibilities, etc.

A graph or network can be defined by a set of vertices and a set of edges. Two vertices are connected if they are linked to each other. For example, let the nodes be the cities and the edges be the roads. In this structure, there is an edge between two vertices if two cities are connected directly by a road. The graphs can be grouped by their different properties, such as weighted or unweighted edges, or by degree distribution, etc. [196].

DOI:10.15774/PPKE.ITK.2016.008

The network is an ordered set of vertices and edges, $G=(V,E)$, where V is the set of vertices and E denotes the set of edges. The two sets define the graph. In this paper the nodes are denoted by their indices (i.e. k, v_k or x_k).

There are several simple properties and classification of graphs.

The degree of a node v_k is the number of edges being incident to the node and it is denoted by $\deg(v_k)$.

Indegree is the number of incoming edges; outdegree is the number of edges that leave the vertex (outgoing links). If we consider an undirected weighted graph, then the degree of a node will be the sum of the weight of incident edges.

If a number (a weight) is associated to the edges, we talk about weighted graph. The weight might mean cost or the strength of the chemical association between two molecules, lengths, etc. In our example, the weight can be the length of the road between two settlements.

Let x, y be two nodes of a graph $x,y \in V$, and $e(x,y)$ the edge between node x, y, $e(x,y) \in E$. The graph is undirected if $e(x,y)$ has no orientation. For example, if there is a road from city A to city B, it means that we can travel from A to B and B to A as well, there is no distinction. On the other hand, if $e(x,y)$ does have orientation, then the graph is directed. If $e(x,y) \in E$ it does not necessarily implicate that $e(x,y) \in E$ is true.

Path of length l is an ordered collection of l vertices $(v_1, v_2, \ldots, v_l \in V)$ and $l-1$ edges $(e(v_1,v_2), e(v_2,v_3), \ldots, e(v_{l-1},v_l) \in E)$. The distance of two nodes is the minimal number of edges traversed on a path between v_i and v_j. If there does not exist a path between the two elements, the distance is infinite. The shortest path between v_i and v_j is the minimum number of traversed edges: $\min\left\{l(v_i,v_i) \mid i,j=1,2,\ldots|V|\right\}$.

A graph is connected if and only if there exists a path between any two distinct vertices in the graph. If the graph is not connected, then it consists of smaller independent components.

DOI:10.15774/PPKE.ITK.2016.008

A directed graph is strongly connected if exists a path between any pairs of nodes considering the orientation of the edges.

For practical uses, a graph is often represented as an adjacency matrix. It describes which nodes are connected and which are not. The adjacency matrix A is an $N \times N$ matrix, where $N = |V|$:

$$a_{ij} = \begin{cases} 1 & \text{vertex } v_i \text{ and are } v_j \text{ adjacent} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

If the graph is weighted, then the entries of matrix A represent the weights of the edges. If the graph is undirected, then the matrix becomes symmetric.

The Laplacian matrix of a graph is often called admittance matrix and it is often used. The Laplacian matrix is an $N \times N$ matrix:

$$L_{ij} = \begin{cases} \deg(v_i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ vertex } v_i \text{ and are } v_j \text{ adjacent} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The Laplacian matrix can be formulated by the matrices:

$$L = D - A \tag{3}$$

Where D is a diagonal matrix and A is the adjacency matrix of the graph.

$$D = diag(d_1, d_2, \cdots, d_N), \tag{4}$$

where $d_i = \sum_j^N A_{ij}$. If the graph is directed then the d_i elements are the outdegrees of the node.

DOI:10.15774/PPKE.ITK.2016.008

1.6. Network analysis techniques

A great large number of mathematical and algorithmical approaches exists for gaining insight knowledge of a network. Each of them give us an information from different perspectives. The global descriptors usually give us information about the network itself, while the different cluster analysis techniques reveal the substructures inherently existing in a network. A large number of algorithms has been developed for exploit the complex relationship existing between nodes or set of nodes in a network. This work is concerned with the concept of network neighborhood that cat can be defined as a subnetwork or subgraph around a selected node. Defining a subnetwork in a data-network can be carried out either by i) static or ii) dynamic methods.

i) Static methods use the data network "as is", and simply omit those data that do not fulfill some criteria. For instance, we omit those data types and connection types that do not belong to the subnetwork. In this way, we can define tissue-specific networks, or we can define the neighborhood of a gene as nodes and edges that are less than n steps away within the network, using paths that contain only a given set of edges. For instance, a neighborhood of a potentially affected drug can be defined as a set of genes that are in the same metabolic or signaling pathway as the known drug target.

ii) Another, probabilistic way is to define a subnetwork as an effect that propagates from a central node, such as a drug target. This is a dynamic approach since the nodes of the network get weighted in an iterative fashion during propagation, and at the end, one can select those nodes that have weights exceeding some threshold value. We are concerned with two kinds of propagation algorithms used in several fields of computer science, PageRank [197-199] and diffusion [200-203].

1.6.1. Random walk based algorithm: PageRank

This algorithm is a special case of random walk on data network: a walker starts at a certain data node, and then randomly selects the next node from its neighbor, then moves there, and so on. In the case of PageRank, the walker not only selects a neighboring node randomly, but also, it can move to any other nodes with a certain probability ("restart probability"). If the walker is only allowed to move to a specific set of nodes or to the neighboring nodes, then this is the PageRank with prior algorithm [198, 199, 204]. If there is prior knowledge available about which nodes are

DOI:10.15774/PPKE.ITK.2016.008

more relevant, then one can use this information to bias the original PageRank scores. For example, known drug targets, known diseases can be used as prior knowledge; in that case, the walker initiated from these specific data nodes, and in every iteration it goes back, restarts with a certain probability.

More formally, first, we define the prior probabilities as pr, and then we use this information in the random walk in the following way:

$$P(v)^{i+1} = (1-\beta)\left(\sum_{u=1}^{d_{in}(v)} p(u,v)P(v)^{i}\right) + \beta pr(v) \tag{5}$$

$P(v)^{i}$ denotes the personalized PageRank score at iteration step i, $u,v \in V$ and $p(u,v)$ is the probability of traveling from node u to node v. β is the "restart probability" ($0 \le \beta \le 1$). β is the probability that we restart the random walk, meaning that we go to the starting nodes according to the prior probability distribution, thus biasing the results towards to the initial conditions. If $pr(v) = 1/N \quad \forall v \in V$, then $P(v)^{i}$ is the original PageRank score at iteration step i.

It is possible to reformulate the PageRank iteration in matrix form:

$$P^{i+1} = (1-\beta)M^{T}P^{i} + \beta pr, \tag{6}$$

where M is the probability transition matrix that can be generated from the adjacency matrix of G. Let D denote the diagonal matrix defined in **equation 4**, then simply:

$$M = DA \tag{7}$$

Note that M is not necessary symmetric even if A is.

Other well-known algorithms based on random walks include k-step Markov [204], HITS [205], HITS with Prior [204].

DOI:10.15774/PPKE.ITK.2016.008

1.6.2. Random walk and kernel methods

Random walk based methods are also used to quantify the similarities between nodes, however it is not a trivial question how we can measure the relatedness (or similarity) between the network nodes. There are several ways to deal with this: for example, by counting all possible paths between any pair of nodes, or by considering a diffusion process on the graph, which resembles to the random walk on graph. The connection between diffusion and random walk has been well established.

The kernel methods offer a very good framework for quantifying the similarities [202]. They also offer a wide range of powerful algorithms for all kind of problems of pattern analysis, for example for clustering, classifications, rankings, principal component analysis, correlations, and regression problems. Furthermore, kernel-based algorithms can be used on general types of data, such as sequences, text documents, sets of points, vectors, images, graphs, etc. by leveraging the typical properties of such type of data. Kernel methods use the following (theoretic) working flow: The first step is mapping the data to high or infinite dimension space, called feature space. In the feature space, each coordinate of the transformed input vector is a feature of the data. In the second step, we apply our algorithm in this new space in order to find relations between the data. The mapping function is not necessarily linear; it is possible to use wide scale of nonlinear functions. Kernel methods compute the inner product between all pairs of data vectors from the feature space. This approach proves to be very efficient, because it is computationally cheaper than finding the exact feature vectors for all data points, and plenty of algorithms can operate using only the scalar product. One of the most famous examples is Support Vector Machines (SVM), but there are kernel based algorithms for Fisher's linear discriminant analysis (LDA), principal components analysis (PCA), canonical correlation analysis, ridge regression, and for spectral clustering [202].

A *kernel* is a function $k : X \times X \to \mathbb{R}$ that for all $x, z \in X$ satisfies:

$$k(x,z) = \langle \Phi(x), \Phi(z) \rangle, \tag{8}$$

where X is the input vector space, and Φ is the mapping function from X to a feature space F.

$$\Phi : x \to \Phi(x) \in F \tag{9}$$

DOI:10.15774/PPKE.ITK.2016.008

There are typical kernel and mapping functions, one of them is the following: let $X \subseteq \mathbb{R}^n$

$$k(x,z) = \langle x,z \rangle^2 \tag{10}$$

This *k* kernel function corresponds to the following mapping function:

$$\Phi : x \to \Phi(x) = (x_i x_j)_{i,j=1}^{n} \in \mathbb{R}^{n^2} \tag{11}$$

If the kernel function and some input vectors are given, then one forms a matrix that is called kernel matrix or Gram matrix. The elements come from the evaluated kernel function on all pairs of data points and the matrix has the following properties:

I. **Symmetry:** The kernel matrices are always symmetric:

$$k(x,z) = \langle \Phi(z), \Phi(x) \rangle = \langle \Phi(x), \Phi(z) \rangle = k(z,x) \tag{12}$$

II. **Gram matrix:** Given a set of vectors $S = \{x_1, \ldots, x_n\}$ the **Gram matrix G** is an $n \times n$ where $g_{ij} = \langle x_i, x_j \rangle$ If we evaluate the *k* kernel function on the input data with the corresponding mapping function Φ, we get a **Gram matrix**:

$$g_{ij} = \langle \Phi(x_i), \Phi(x_j) \rangle = k(x_i, x_j) \tag{13}$$

III. **Kernel matrices are positive semidefinite matrices.** A symmetric matrix is positive semi-definite if its eigenvalues are all non-negative. This holds if and only if

$$v^T A v \geq 0 \tag{14}$$

1.6.3. Kernels on graph

Kernels can be applied on several types of structured data, such as graphs, strings, trees, etc. These kernels allow us to study the graph structure and they can reveal important features, such

DOI:10.15774/PPKE.ITK.2016.008

as the similarities between nodes or clusters. If node x is linked to node y, and node y is linked to z, it is reasonable to say that x is similar to z, although there may not exist any direct link between them. Diffusion kernels quantify the similarities between two nodes by considering all possible paths between them. It is a reasonable consideration that longer paths have a lower contribution to the similarity than the shorter ones. Depending on the type of the discount factor applied on the path length one can define various different graph kernel types [202]. The path counting is indirectly carried out by powering the adjacency matrix.

Kondor and Lafferty introduced the exponential diffusion kernel [206] as (K_{ed}):

$$K_{ed} = \sum_{i=1}^{\infty} \frac{\alpha^k A^k}{k!} = \exp(\alpha A), \tag{15}$$

where the discount rate is exponential. Parameter α regulates the decay of the longer path.

Another slower discount rate leads to another type of kernel, called von Neumann diffusion kernel [201], defined by the following formula:

$$K_{ND} = \sum_{i=1}^{\infty} \alpha^k A^k = (I - A)^{-1}, \tag{16}$$

which only exists if the $\alpha \leq \|A\|_2^{-1}$ Where $\|A\|_2$ is the spectral norm of A.

It is easy to see that these kernels are well defined since they are positive semi definite and symmetric because of the construction process. The exponential diffusion kernel and the von Neumann diffusion kernel have the same eigenvector as the adjacency matrix, the only difference is that how they reweight the eigenvalues of A.

The exponential function is always positive, therefore the rescaled eigenvalues are positive as well, thus the exponential diffusion kernel is positive definite. In the case of von Neumann diffusion kernel, the rescaled eigenvalues are $(1-\alpha\lambda)^{-1}$, thus this kernel is also positive semi definite.

DOI:10.15774/PPKE.ITK.2016.008

T. Ito showed that kernels give us a unified framework for the importance and relatedness [200]. At first sight, one can say that the importance measure is different from the similarity measure since importance is defined on a node while similarity is defined between nodes. Let $G(V,E)$ be a weighted (all weight is positive) connected undirected graph and A the corresponding adjacency matrix and v an importance score vector of G. For example, v could be the dominant eigenvector of A, and it is well known that the entries of v are the scores of how important a node is [197, 200, 205]. Let us consider the vv^T matrix. The i th row (or column) gives a ranking, which is identical to the one defined by v.

Let λ be the dominant eigenvalue of A, if it has multiplicity one, then:

$$\lim_{n\to\infty}(\frac{1}{\lambda}A)^n = vv^T \tag{17}$$

It is also proved that:

$$\lim_{\alpha\to\lambda^{-1}}\left(\sum_{k=1}^{\infty}\alpha^k A^k\right)^n = vv^T \tag{18}$$

It has a consequence in the case of von Neumann kernel, if we choose α as $\frac{1}{\lambda}$, then the i th row (or column) of K_{ND} gives the same ranking as the HITS ranking. The parameter α can be interpreted as bias towards the ranking based on importance. If we choose a small α, then the importance of the node is not really dominant. However, if $\alpha \geq 0$ then the importance has an effect on the similarity defined between the nodes, thus von Neumann kernel offers a framework to study the similarities and the importance together.

The Laplacian exponential diffusion kernel K_{led} is almost the same as the K_{ed}. The difference is that instead of adjacency matrix we use minus Laplacian matrix in the formula. This can be interpreted as a heat diffusion on the graph. Diffusion is a physical metaphor used to model transport phenomena on networks. In our case, we assign an imaginary quantity, such as “energy” or “drug action” to one node of the network – for instance the gene targeted by the drug – and then use an iterative process to compute how this quantity diffuses along the network. Let x_i be the

DOI:10.15774/PPKE.ITK.2016.008

quantity of the energy on node. It diffuses to the neighboring nodes with rate a_{ij}, so we can write that the energy of node i is increased by $\sum_{j=1}^{N} a_{ji} x_j \delta t$ between a small time interval δt. The energy loss of the node is: $\sum_{i=1}^{N} a_{ij} x_i \delta t$. It leads to the following differential equation for $x(t)$:

$$\frac{dx(t)}{dt} = -Lx(t) \tag{19}$$

The solution of this differential equation with respect to the initial condition $x(0)$ is:

$$x(t) = e^{-Lt} x(0), \tag{20}$$

In a similar way to PageRank with prior, it is possible to incorporate prior knowledge about the data network, i.e. relevant drugs to a disease by regularizing the Laplacian matrix [200]. The regularization could be interpreted as alteration of diffusion process by i.) controlling (increasing or decreasing) the energy loss of a node, ii.) altering (increase or decrease) the input energy flow on certain edges, iii,) both of the above. All of the above alterations can be described with different regularization parameters, more formally the regularized Laplacian matrix defined as:

$$L_{\mu,\gamma} = QD - WAW \tag{21}$$

Where the W and Q matrices are defined as follows:

$$w_{ij} = \begin{cases} \gamma & \text{if } i = j \text{ and } \mathrm{x}_i(0) \neq 0 \\ 1 & \text{if } i = j \text{ and } \mathrm{x}_i(0) = 0 \\ 0 & \text{if } i \neq j \end{cases} \tag{22}$$

$$q_{ij} = \begin{cases} \mu & \text{if } i = j \text{ and } \mathrm{x}_0^i \neq 0 \\ 1 & \text{if } i = j \text{ and } \mathrm{x}_0^i = 0 \\ 0 & \text{if } i \neq j \end{cases} \tag{23}$$

Where t, μ and γ are the regularization parameters of the algorithm.

DOI:10.15774/PPKE.ITK.2016.008

The evaluation of an extremely large system of ordinary differential equations could be a challenging task; however, using sparse linear algebra and leveraging the sparseness of a typical data network, the solution could be computed in reasonable time. Instead of focusing on computing the matrix exponential, one could focus on the approximation of the matrix-vector product gaining a significant speed up. The expression $e^{-Lt}x(0)$ could be approximated using iterative methods such as Arnoldi algorithm [207-209].

There are more graph kernels and variations such as commute time kernel [210], that is the Moore-Penrose pseudo inverse of the Laplacian matrix. F. Fouss [210] also showed in his paper that the average commute times and the average first passage times of the random walk can be computed using the kernel.

1.6.4. Methods for graph kernel computation

Computing a kernel matrix of an extremely large graph is a computationally expensive task (the time complexity is $O(n^3)$, where n is the number of nodes in the graph). The above presented graph kernels can be classified into two groups, one is kernels based on von Neumann series, and the others are kernels based on matrix exponential. It is well known that computing the inverse or matrix exponential of a general, large matrix is a computationally challenging problem. The size of a human protein-protein interaction network is approximately 20000 x 20000 (this is considered to be large). Since the computational time is considerable, finding the optimal kernel parameters for an application has also become challenging. To overcome this limitation, approximation algorithm can be used, such as Arnoldi algorithm [209], or Cholesky decomposition [211]. Further advantage of that algorithms is that they exploit the graph sparsity, which is an inherent property of the biological networks [212].

It is a frequent case in practice, that not the whole kernel matrix is needed, but the product of the kernel with a vector, and in that case, Arnoldi algorithm can be extremely useful for computing that product [207, 209, 213].

The computation of von Neumann series is also possible using series approximation, however the approximation speed does greatly depend on the condition number of the adjacency or the Laplacian matrix, which is typically very high in biological networks like STRING [4], thus the convergence is slow. The other approaches are based on the matrix decomposition, like

DOI:10.15774/PPKE.ITK.2016.008

Cholesky decomposition technique [211]. Although, this decomposition technique approach is "direct", thus it does not leverage the sparsity property directly, it is still possible to reduce computation time by transforming the adjacency matrix by using the reverse Cuthill–McKee algorithm [214].

1.6.4.1. Krylov space methods

The evaluation of matrix functions, such as matrix exponential or inverse, is often challenging in practice. In a number of such cases, Krylov subspace methods are easily applicable, especially in the time evolution of a complex system, but we can use them in several other fields, such as solving linear equations, finding the eigenvalues of a matrix, and evaluation of square root or trigonometric functions [205].

Let matrix be large and sparse. Then Krylov space methods are useful, especially if we want to evaluate the following matrix vector product:

$$f(A)b, \quad \text{where } A \in \mathbb{C}^{N\times N}, b \in \mathbb{C}^{N}$$
$$f : \mathbb{C} \subset D \to \mathbb{C} \text{ is a function for which } f(A) \text{ is defined} \tag{24}$$

In most applications, only the product $f(A)b$, and not $f(A)$ is needed. For example, when we solve the linear equation $Ax = b$, the solution is $A^{-1}b$, thus the solution is the result of the multiplication of the inverse matrix with b. Another example of solving a system of ordinary differential equations given: $\frac{dx(t)}{dt} = Ax(t)$, and $x(0)$ as the initial condition. Then the solution is $x(t) = e^{At}x(0)$. In that case, the solution is also a vector, given by a matrix vector product.

Krylov space algorithms are iterative methods, and they give an approximation of the solution with an acceptable error within a reasonable computational time. They have some further advantages over other iterative methods, namely the matrix A is required only for computing matrix vector products. It is a very advantageous property while working with large sparse matrices. Usually, the computation cost of a matrix vector product is $O(N \times N)$, but it is easy to construct

DOI:10.15774/PPKE.ITK.2016.008

data structures and procedures, for which the cost of evaluating of Ab is only $O(\rho N)$, where ρ is the number of nonzero elements per row. Another important benefit is that the approximation of $f(A)b$ for a smooth function f (f is continuously differentiable, such as the exponential function) converges super linearly [207].

The family of Krylov space methods has many members. The Generalized minimal residual method (GMRES) [215], Conjugate gradient method (CGN) [216], Biconjugate gradient method (BiCG) [217] solve linear equation (you can decide whether your matrix is hermitian or not). Based on Arnoldi and Lánczos algorithm, it is possible to find the extreme eigenvalues of A as well.

1.6.4.2. Arnoldi algorithm

Most of the Krylov space methods are based on the outcome of the Arnoldi algorithm. The main idea is to project matrix A to a smaller space, called Krylov subspace, and looking for a solution in this smaller space. Arnoldi algorithm computes the projection of A, and builds an orthogonal basis on the Krylov subspace, and this projection can be seen as the compressed version of A[209].

Krylov subspace is a vector space spanned by the vectors of Krylov sequence, K_s which is a set of vectors generated by A and b:

$$K_s = \left\{b, Ab, A^2b, \ldots, A^{m-1}b\right\} \tag{22}$$

The m th Krylov subspace of $A \in \mathbb{C}^{N\times N}$ and $b \neq 0, b \in \mathbb{C}^N$ is:

$$K_m = span\left\{b, Ab, A^2b, \ldots, A^{m-1}b\right\} = \left\{q(A)b : q \in P_{m-1}\right\} \tag{23}$$

P_m are the polynomials with degree m. If A is sparse, we can generate K_s very easily since in every iteration we just multiply A with the result vector of the previous iterations.

The approximation of the matrix function $f(A)b$ with the help of Krylov space is done by looking for a polynom [209]:

$$f(A)b \approx p_{m-1}(A)b \ , \tag{24}$$

DOI:10.15774/PPKE.ITK.2016.008

where $A \in \mathbb{R}^{N \times N}$, v is a nonzero vector and p_{m-1} is a polynomial with degree m-1. The approximation is also the element of the Krylov subspace (K_m). The goal is to find an element in K_m that approximates $f(A)b$.

The Arnoldi approximation of $f(A)b$ is defined by [209]:

$$f_m = \beta Q_m f(H_m) e_1, \quad \text{where } \beta = \|b\| \tag{25}$$

The rationality behind this approximation is that H_m represents the compression of A into $K_m(A,b)$ with respect to the basis Q_m and $b = \beta Q_m e_1$, where e_m denotes the m th unit coordinate vector. The matrix values of Q_m and, H_m as calculated by a modified Gramm-Schmidt process [209] that generates a m orthonormal vectors in K_m.

1.6.4.3. Approximating the matrix exponential

The exponential of a matrix plays an important role in many applications such as in the analysis of networks. Although **formula 25** computes a matrix vector product, it is possible to get the full $\exp(\alpha A)$ column by column. One gets back the i th column of $e^{\alpha A}$ by choosing b as e_i, where $e_i \in \mathbb{R}^{N \times N}$ and e_i is the unit coordinate vector. The approximation formula for i th column of matrix exponential of A is:

$$e^{\alpha} e_i = Q_m e^{\alpha H_m} u_1 \ , \tag{26}$$

where u_1 also denotes the unit coordinate vector, but the dimension of u_1 is m ($u_1 \in \mathbb{R}^m$). First, we compute the first column, then the second column, and so on.

1.6.5. Ranking in networks

In bioinformatics, many problems are related to rankings. A typical example in bioinformatics is protein annotation, where an unknown protein is queried against a large and comprehensive protein database with a proper aligner such as BLAST. The output is a ranked list of proteins in the database, where the top hits are the most similar proteins to the query. These top lists might be the basis of the next step in the research. But not only the sequences are usually subjected to ranking, but all sorts of biological entities, such as drugs, drug combinations, pathways,

DOI:10.15774/PPKE.ITK.2016.008

mutations, peptides, nodes in network, etc. The questions is naturally arising: how to define ranking on networks.

Ranking of nodes can be the representative of importance within a network, or in other cases, it can indicate how close the node is to our query representing our basic knowledge on the subject. That type of knowledge can vary depending on the field of application, for example in the case of drug repositioning, the most important knowledge is about the known applications of that drug, or in the case of the disease candidate gene prioritization, the input is the known genes related to the disease.

We assume that graph G is connected, and the adjacency matrix of the graph $A(G)$ has dimension $N \times N$. The prior knoweldge is represented as vector denoted by p_0 with dimension $N \times 1$. and also have a vector with dimension N denoted by p_0.

1.6.5.1. Ranking by using PageRank with priors

This algorithm is very similar to the k-step Markov in a sense that this is also an iterative algorithm and we get probability distribution in every iteration step what we can use for ranking or prioritization. But this also converges to a steady state distribution, which is not equal to the original PageRank vector, since it is biased towards an initial distribution, which is our prior knowledge about the system. Using prior knowledge as the initial distribution of the random walk (p_0 and p_0 also normalized to 1: $\sum_{i=1}^{N} p_0(i) = 1$), a probability distribution is generated in every step (denoted by $p^{(k)}$), which can be seen as a ranking vector. The i th entry of $p^{(k)}$ will be the score of node i. This will be the probability of being at this node after k step with respect to the initial condition and the restart parameter β as described in **Section 1.6.1**. The higher probability means higher rank. This is similar to the relative importance of a node. In the case of PageRank, the pr in **equation 6** will be replaced by p_0 .

1.6.5.2. Ranking based on kernels

The entries of the kernels can be considered as the similarity between a pair of nodes. Here one can simply reweight the prior knowledge vector with the proper kernel values. In the case of the exponential diffusion kernels (exponential diffusion kernel, Laplacian exponential diffusion kernel, regularized Laplacian exponential diffusion kernel), we choose the diffusion and the

DOI:10.15774/PPKE.ITK.2016.008

regularization parameters as we want (of course they have to be positive number). On the other hand, when we work with von Neumann diffusion kernel, Laplacian diffusion kernel, or with the regularized Laplacian kernel, we have to choose the diffusion parameter (α) carefully, otherwise the convergence of the series is not guaranteed. The $0 \le \alpha \le \frac{1}{\lambda_{\max}}$, where $\lambda_{\max}$ is the spectral norm of matrix A (or L, depending on which kernel we investigate). Let K be a kernel matrix, and the query vector be p_0 , then the ranking p is simply defined as :

$$p = Kp_0 \tag{27}$$

There are other ways of using kernels. Since the entries are inner products, it is easy to derive distance measure from the kernel. Another question is whether we have to normalize the kernel or not. The normalization here means that the norm of the vectors in the feature space is one.

1.6.5.3. Measuring ranking performance

A common question in bioinformatics is that how good a ranking is [1, 218]. In such cases, the traditional ROC (receiver operating characteristic) analysis [1] could be useful, since the ranking can be seen as a classification task. The various indices derived from the ROC curve, such as AUC (Area Under ROC Curve), precision, recall are particularly useful for characterizing the performance.

The original use of ROC analysis is related to binary classification tasks, where the goal is to tag an element either as positive (+) or as negative (-). The classification algorithm is generally trained on elements with known class (training set). Usually, a testing phase follows the training process in order to assess the performance of the classification algorithm with respect to a given parameter set. The test set is distinct from the training set, and its label is also known (+ or -).

The algorithm classifies each element in the element test set as either positive or negative. If the element is positive and is classified as positive, then this is a true positive (TP) hit, if it is classified as negative then this is a false negative hit (FN); on the other hand, if an object is negative and mapped as positive it is called false positive (FP), if it is classified as negative, then this is a true negative hit (TN). The classification result is then summarised in a so-called contingency table that contains the counts of TP, FP, TN, FNs. Then various indices could be derived from the contingency table, such as accuracy, sensitivity, F-measure, etc. For ROC analysis, the false

DOI:10.15774/PPKE.ITK.2016.008

positive rate (FPR) calculated as: $FP/FP+TN$ and the true positive rate (TPR) calculated as: $TP/FN+TP$ are particularly important.

In many cases, the classification is not binary, i.e. the classifier assigns a score instead of a label to each object representing the "certainty" of the membership to a class. In that case, the result is a ranking defined on the score. The classifier is considered good if the positive examples are ranked in the top of the list, while the negatives are ranked in the bottom of the list. The classification is poor if the positive elements are uniformly distributed in the ranked list. Following this guideline. one can transform non-binary classification into a binary classification by applying a threshold value, where the objects having a larger score are considered as positives and the other objects as negative. After the thresholding step, one can build a contingency table and derive the corresponding TPR, FPR values. In the final step, the ROC curve is created from the TPR and FPR values by applying various threshold values on the ranking. Sonego et al. used the below figure (**Figure 1.1**) to illustrate the process.

After the construction of the ROC curve, the AUC value could be calculated. This is 1 if the ranking is perfect (all positive elements are ranked at the top), and 0.5 if the classification is random.

1.6.6. Inference in ontologies

Ontologies, often represented as directed acyclic graphs (DAG) or trees, are special networks capable of expressing hierarchical biological knowledge. The practical graph mining of that special structures often require other approaches than simple random walk or diffusion due to the hierarchy. Typical examples for hierarchical structures are gene ontology (GO) [135], disease ontology (DO) [89], clusters of ortholog groups (COG), taxonomy, etc.

DOI:10.15774/PPKE.ITK.2016.008

Values	Class
1	+
0.98	+
0.84	+
0.71	+
0.69	–
0.69	+
0.64	+
0.62	+
0.56	–
0.5	+
0.39	–
0.37	–
0.36	+
0.32	+
0.32	–
0.29	–
0.28	–
0.25	–
0.11	–
0.00	–
AUC	0.87

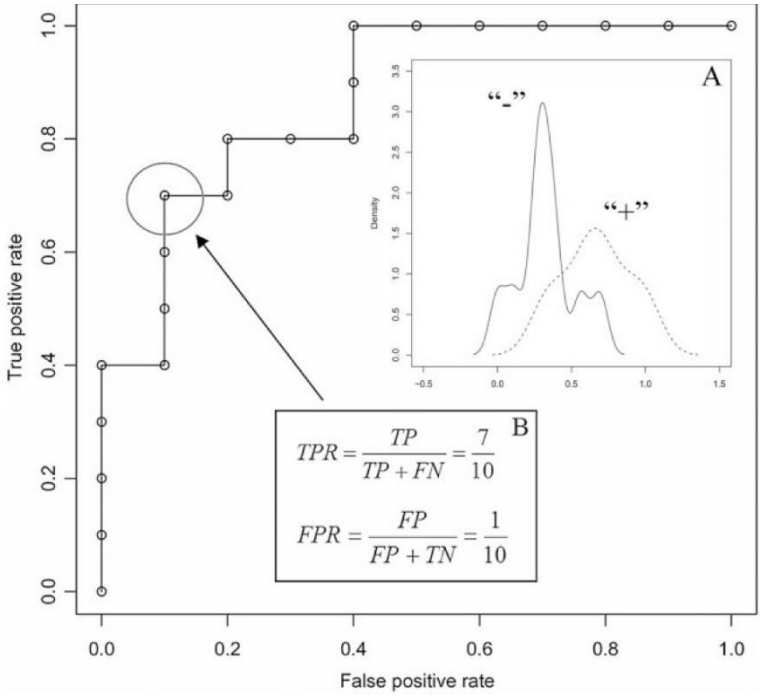

Figure 1.1. Constructing a ROC curve from ranked data (taken from Sonego et al. [1])

The TP, TN, FP and FN values are determined compared to a moving threshold; an example is shown by an arrow in the ranked list (left). Above the threshold, + data items are TP, - data items are FP. Therefore, a threshold of 0.6 produces the point FPR = 0.1, TPR = 0.7, as shown in inset B. The plot is produced by moving the threshold through the entire range. Data were randomly generated based on the distributions shown in inset A [1].

The similarity between two or more nodes in the hierarchy is usually characterized through their lowest common ancestor. The similarity value is calculated based on a property of one of the common ancestors, such as information content (IC), or the shared path, or distance, or the combination of the above. The IC quantifies the specificity of a node at a given level. The IC of a node (x) is the negative log likelihood of the probability of the node $p(x)$ being used. These can be estimated from the annotation frequencies. Resnik measure [219] is one of the most commonly used semantic similarity measures [220, 221], that is the most informative common ancestor (MICA) of two nodes:

$$sim\left(x_i, x_j\right) = \mathrm{IC}\left(MICA\left(x_i, x_j\right)\right) \tag{27}$$

In many cases the data nodes (ontology terms) are assigned to various entities, such as proteins and drugs; usually, the cardinality of such a relationship is many to many. If the goal is to

DOI:10.15774/PPKE.ITK.2016.008

quantify the similarity between the entities and not between the terms, then two types of approach could be used, one that discards the structure of ontology, the other is aggregating all similarity between the terms annotated to the entities. The former reduces the ontology knowledge into a presence absence vector (g_i), where the i th entry is 1 if the i th term is annotated to the entity, 0 otherwise. In the final step, the similarity between the entities is computed with the help of a standard vector similarity or distance measure, such as Jaccard, cosine or Hamming distance. The cosine similarities between two drugs (or any other type of entities) can be computed as:

$$S_{GO}\left(D_i, D_j\right) = 1 - \frac{g_i^T g_j}{\left\|g_i\right\|\left\|g_j\right\|} \tag{28}$$

The score is 1.00 if the two entities have the completely same term annotation and 0 if they do not share any term.

DOI:10.15774/PPKE.ITK.2016.008

2. Databases and methods

Many tools, pipelines, IDEs, libraries, technologies and programming languages have been utilized throughout the projects as well as a large body of datasets. During the development I used a desktop computer with Intel(R) Core(TM) i5 CPU (3.33GHz) and 16GB RAM as well as a DELL PowerEdge R720 server equipped with 2 Intel(R) Xeon(R) CPU E5-2640 (2.50GHz) processors and 32 GB RAM.

2.1. Databases

The protein-protein interaction data were taken from the STRING database [222] (http://string.embl.de/, retrieved on 28th august of 2012). I also used an archived version of STRING database, release version 6.3 (in use from December 12, 2005 to January 15, 2007) for hypothesis testing. The drug related data (drug targets, synonyms, aliases, ATC codes) were taken from the Drugbank [106] via the JBioWH [195] (https://code.google.com/p/jbiowh/, retrieved on 12th September of 2012), STITCH [110] (http://stitch.embl.de/, retrieved on 4th September of 2012) and TTD [223] (http://bidd.nus.edu.sg/group/TTD/ttd.asp, retrieved on 23th July of 2012) databases. The drug interaction data were taken from http://drugs.com/ (retrieved on 11th November of 2013). The drug combination data were taken from the DCDB [186] (http://www.cls.zju.edu.cn/dcdb/, 4th March of 2012), and TTD [223] (http://bidd.nus.edu.sg/group/TTD/ttd.asp, retrieved on 23th July of 2012) databases.

From the STRING database, the human protein-protein associations and their combined confidence scores were used. From the STITCH database only those drug-protein associations were considered which had i) experimental evidence or ii) database evidence with at least 0.800 confidence, and the overall confidence was at least 0.900. Molecules such as Na^{+}, Ca^{2+}, ATP, etc., that had more than 45 targets were excluded from the dataset. All filtering algorithms were implemented in MATLAB R2014a.

Published clinical trial data on trastuzumab were collected from the ClinicalTrials database (www.clinicaltrials.gov) using the word ‘trastuzumab’ in pairwise combination with all the 43 chemotherapeutic agents approved for breast cancer (amsacrine, azacitidine, bleomycin, cabazitaxel, capecitabine, carboplatin, carmustine, chlorambucil, cladribine, cyclophosphamide,

DOI:10.15774/PPKE.ITK.2016.008

cytarabine, dacarbazine, daunorubicin, daunorubicin <liposomal>, docetaxel, doxorubicin, epirubicin, estramustine, etoposide, fludarabine, fluorouracil, gemcitabine. idarubicin, ifosfamid, irinotecan, ixabepilone, lomustine, mercaptopurine, methotrexate, mitomycin-c, mitoxantrone, nelarabine, oxaliplatin, paclitaxel, pemetrexed, pentostatin, temozolomide, teniposide, thioguanine, topotecan, vinblastine, vincristine, vinorelbine) on the 1st of January 2013. ClinicalTrials.gov is developed by the U.S National Institute of Health and contains summary information about clinical studies conducted all over the world. Only 18 agents were studied in combination with trastuzumab in 81 trials. The findings were narrowed down to trials in which the effect of the combined therapy was studied (n=43). For trials in which trastuzumab was studied in combination with more than one agent, these duplicates were included only once. Only the data recorded according to Response Evaluation Criteria In Solid Tumors Criteria (RECIST) [224] were used. Overall clinical response (rate) (OR) was calculated from percentage of patients with complete response (CR) and partial response (PR) (OR = CR + PR) [224]. The Confirmed Clinical Benefit (CCB) was calculated from CR, PR, and stable disease (CCB = CR + PR + SD) [224]. Finally, the median progression free survival (PFS) and the median overall survival (OS) data were added in months.

2.2. Data preprocessing

The data networks have been stored in traditional relational databases. In my projects I used MySQL and Oracle based systems. For designing the entity-relationship (ER) diagrams I partly used the Oracle SQL Developer program (http://www.oracle.com/technetwork/developer-tools/sql-developer/) and the MySQL Workbench Tool (http://www.mysql.com/products/workbench/).

The DCDB was integrated with Drugbank, TTD, STRING, STITCH and JBioWH data, and the necessary constraints and indices were built. The various types of ambiguities have been handled manually. The programs accessed the databases through JDBC (Java Database Connectivity).

The STRING, DrugBank and STITCH databases were preprocessed (filtered to human related proteins and chemicals as described above) in Java and later in Python using dedicated parsers written for that purpose. All the protein ids were mapped to UniProtKB accession numbers, using the mapping files provided by the UniProt (idmapping tool).

DOI:10.15774/PPKE.ITK.2016.008

The programs also transform the data into necessary format to make the data import into the database easier.

Although the Oracle DBMS supports XML and has advanced retrieval and indexing utilities, I implemented a dedicated parser for Drugbank in Python, because it is more simple to integrate the information via dedicated data structures. For that purpose I used *xml.etree.ElementTree* module (https://docs.python.org/3/library/xml.etree.elementtree.html).

The data was loaded into the database either with the Oracle SQL Developer migration tool or with the SQL*Loader utility.

The document sets in our experiments were acquired from the MEDLINE database through its PubMed system [194] using the Entrez Programming Utilities [225] via Biopython [226]. Each document set consisted of citations that comprised of abstracts obtained from PubMed by executing boolean queries. The target sets of texts were restricted to abstracts of articles, because unlike the majority of full texts, they are freely available online in XML format.

The sequence databases used in our experiments were created from indexed blast database files. The raw indexed files were retrieved from the ftp site of the NCBI along with the NCBI taxonomy and the corresponding GI (genbank identifier) - ncbi taxon id mapping. In my experiments I used the NT database as the main sequence source. The raw sequence database was stored in fasta format. During the second step of database preprocessing the nt database was split into ~4Gb pieces and the header of each sequence was replaced by the GI identifier and the taxon id of the organism. Then the database was indexed by the Bowtie2-build program. Since the original NT database contains all types of sequences from various organisms, not only microorganisms, it is reasonable to create subsets (i.e. only bacteria, only bacteria and virus etc.). Such precalculated databases are available at https://code.google.com/p/taxoner/wiki/07_Databases. The preprocessing steps were implemented via UNIX shell scripts in Python and in C.

2.3. Methods: programs and environments

The network neighborhood analysis was carried out in the Matlab programming environment. The data networks have been stored in dedicated MATLAB data structures using object-oriented programming features. The raw data loaded from the database via JDBC and SQL queries. The

DOI:10.15774/PPKE.ITK.2016.008

queries describe the filtering process, i.e. which type of interactions should be included and which not. In my experiment the weight of the edges between the proteins were the confidence values provided by the STRING database. The weight of drug-protein and drug-drug connections has been universally 1. In my application all types of drug – protein associations were considered to be a link (i. e. proteins/genes targeted by drug, enzymes, carriers). The network object furthermore holds information about itself and can compute various network properties: i.e. finding the connected components, calculates the degree distribution, regularized Laplacian matrix, corresponding probability transition matrix, (see **equation 7** for details). It also contains the proper id mappings and annotations. The network itself is represented as a simple sparse matrix. It is necessary to filter the data network to the largest connected component, otherwise the numeral stability cannot be guaranteed. In the case of large, undirected networks (number of nodes > 15000), the largest connected component was calculated via heuristics that use the small world property. Basically, I started a random walk from the node with the largest degree (likely to be in the largest connected component) and I continued the iteration until all nodes had been found in the component. It is implemented as a simple matrix-vector multiplication, where the vector is updated in every iteration, in which non-zero entry implies that the node is indeed in the component. On the other cases the *graphconncomp()* function is used. Generally, as much data as possible was pre-calculated and stored in matlab binary format. The data network, kernel matrices, random distributions, different drug-drug interaction scores have been stored and they were loaded on demand. It is reasonably, since many times the same data is required and it is computationally much cheaper to store and retrieve the data then compute on demand. The execution time of such calculations heavily depends on the size of the data network. In our case (human related database) it requires 6-24 hours on an average desktop computer. The Gene Ontology was retrieved via the Bioinformatics Toolbox™ of MATLAB.

```
./taxoner -dbPath path/to/database/fasta/ -taxpath
/path/to/nodes/nodes.dmp -seq
/path/to/fastq/illumina.fastq -p 6 -o Results/ -dbNames
bacteria;archaea
```

The Taxoner uses other type of network approaches. The algorithm was implemented in ANSI C, since C is suitable for building computationally demanding programs. The program is only running in GNU/LINUX environment. In order to harness the advantages of many core architectures the POSIX Thread and MPICH third party libraries were used for support parallelism. The first step in the pipeline is aligning the reads against the reference genomes using the *bowtie2* program with the provided or default parameters. Since the database may consist of several pieces, such as in the

DOI:10.15774/PPKE.ITK.2016.008

case of nt, each read has to be mapped to the sequences in each sub-database. After all mappings are calculated the taxonomic binning is started. For each hit in a SAM file the LCA is calculated and reported. In the end the various temporary output files are merged with help of the UNIX *sort* command and the final LCA, alignment information of the best hit is reported. Optionally, it is possible to generate a MEGAN compatible report.

For usage in large-scale pipelines, Taxoner can be run from the command line. An example is:

```
./taxoner -dbPath path/to/database/fasta/ -taxpath
/path/to/nodes/nodes.dmp -seq /path/to/fastq/illumina.fastq
-p 6 -o Results/
```

where the `-dbPath` tells Taxoner where to find the sub-databases (and Bowtie2 indexes), the -taxpath is the path to the NCBI nodes.dmp file, the `-seq` is the input fastq reads, `-p` specifies the number of threads and -o is the output folder for the results. As an alternative, where the user wants to align reads to only a part of the database (say all the bacteria and archaea), an extra parameter `-dbNames` can be added, with a semicolon separated list of prefixes for each database subset:

```
./taxoner -dbPath path/to/database/fasta/ -taxpath
/path/to/nodes/nodes.dmp -seq
/path/to/fastq/illumina.fastq -p 6 -o Results/ -dbNames
bacteria;archaea
```

The running time was measured in a UNIX environment using shell scripts. Each measurement was repeated at least 3 times, then the average value was reported. The NCBI taxonomy was used as reference taxonomy. In order to compare the results of MetaPhlAn with Taxoner's, MetaPhlAn's taxonomy was mapped to the NCBI taxonomy, thus each marker and clade received a unique NCBI taxon id. The ambiguities were resolved manually (as many as 22), however. these clades were not related to my datasets. MEGAN was set to use the same taxonomy as Taxoner. The Taxonomy tree module was implemented in Python in order to assess the necessary statistics (e.g. counting hits in different branches and taxonomy levels, calculating F-measures, false negative rates etc.).

DOI:10.15774/PPKE.ITK.2016.008

3. Results and discussion

3.1. Discovering novel drug combinations

In the past few decades, the number of novel marketed drugs has fallen much below the expectations despite the growing resources invested in this area [227-229]. Many biological pathways have rich regulatory loops, which can be utilized to compensate various perturbations. In cancer therapy, drugs acting on the HER2 (erb-b2 receptor tyrosine kinase 2) and EGFR (epidermal growth factor receptor) pathways have shown this type of drug evasion effects. Multitarget drugs or drug combinations have been proposed as a general strategy to circumvent this phenomenon [230, 231]; one of the reasons is that combinations often have less toxicity and higher therapeutic success [232]. The number of approved drug combinations is on the increase, even though most of them were established by experience and intuition [233, 234].

About one-fourth of breast cancer patients express HER2 (human epidermal growth factor receptor-2), a transmembrane receptor tyrosine kinase of the epidermal growth factor receptor (EGFR) family. In HER2 positive patients, administration of trastuzumab, an anti-HER2 therapy, improved the progression free survival (PFS) and the overall survival (OS) [76]. It also enhanced survival as adjuvant therapy combined with chemotherapy [235] or as monotherapy after chemotherapy [236]. Since 2006, trastuzumab is also approved for use in adjuvant settings in HER2 positive early breast cancer. Anti-HER2 therapy is highly successful: although high HER2 expression was previously associated with worse survival, today HER2 positive patients have better prognosis as compared to women with HER2 negative disease [37].

According to current NCCN guidelines (www.nccn.org), trastuzumab is given in combination with adjuvant chemotherapy only. Preferred regimes for chemotherapy with trastuzumab include Adriamycin, Cyclophosphamide, Paclitaxel, Docetaxel and Carboplatin. Numerous other agents are also included in protocols used for breast cancer patients including Methotrexate, Epirubicin, Fluorouracil and protocols containing combinations of these (FAC, CAF, CMF, EC, FEC, TAC, etc.). Thus, the combination of various agents into multi-agent protocols represents the backbone of the state of the art in systemic treatment for HER2 positive breast cancer. However, finding the most efficient combinations of these is not an easy task given the complexity of the underlying biological system.

DOI:10.15774/PPKE.ITK.2016.008

Several experimental methods, even high-throughput methods [237], have been developed for measuring the efficiency of drug combinations, such as Bliss independence or Loewe additivity [238-240]. Wong et al. used a stochastic search algorithm [241], while Calzoari and associates employed sequential decoding algorithms for finding the best combinations [242]. Yang et al. used differential equations to find a perturbation pattern that can revert the system from disease state to a normal state [243]. Jin and associates employed a Petri net based model to microarray data in order to predict synergism of drug pairs [244]. A common feature of these computational methods is that they require a large number of experiments or deep knowledge of the kinetic parameters of the pathways even if the search space is small.

Other studies used various combinations of data mining methods to integrate pharmacological and network data [245, 246]. Li and coworkers used the concept of network centrality and disease similarity to prioritize drug combinations [247]. Wu and associates used the microarray profile of the individual drugs for the predictions [245], and others applied the concept of synthetic lethality, and the available gene interaction data [248, 249]. Despite the countless attempts, there are still many challenges and open practical questions. In particular, finding suitable data representations and similarity measures is not a trivial problem because of the heterogeneity of information sources. Currently there are published data on a large number of drug combinations (six hundred in the DCDB and TTD databases as of March 2013), that refer to a variety of diseases and therapeutic targets. It is an open question whether or not the correlations and tendencies extracted from such heterogeneous datasets can be successfully applied to a specific problem, such as that of trastuzumab.

Here I present a novel principle that is based on the assumption that perturbations generated by the pharmacological agents propagate through an interaction network to other targets that constitute what we call a propagation neighborhood. Overlaps of multiple propagation neighborhoods can then cause unexpected synergies at target genes that are not in the immediate vicinity of the original targets of the individual agents. I introduce a novel Target Overlap Score (TOS) that is based on the overlap of the propagation neighborhoods of the target proteins. I show that TOS is correlated with the known efficiency of beneficial and deleterious effects of drug combinations reported in the DCDB, TTD and Drugs.com databases. I also show that there is a correlation between TOS and the outcome of recent clinical trials where trastuzumab was used in combination with anthracycline- and taxane-based systemic chemotherapy in HER2-receptor positive breast cancer.

DOI:10.15774/PPKE.ITK.2016.008

3.1.1. TOS: a network-based Target Overlap Score

Drug molecules reach their therapeutic effects by acting on specific targets in the organism and activating or inhibiting the functions of their targets. Drug effects naturally do not end here, since drug targets are members of large interaction networks through which the perturbation can propagate. For instance, by inhibiting the action of a single molecule such as BRAF (B-Raf proto-oncogene, serine/threonine kinase), the entire RAF/MEK/ERK (Raf-1 proto-oncogene, serine/threonine kinase, mitogen-activated protein kinase 1, mitogen-activated protein kinase 1) pathway will be tuned down, and as a consequence, collateral pathways including PI3K (phosphatidylinositol-4,5-bisphosphate 3-kinase) and RALA (v-ral simian leukemia viral oncogene homolog A (ras related)) will also be affected. In other words, a drug acting on a single target will concomitantly perturb a group of linked targets that is termed here as network neighborhood **(Figure 3.1).**

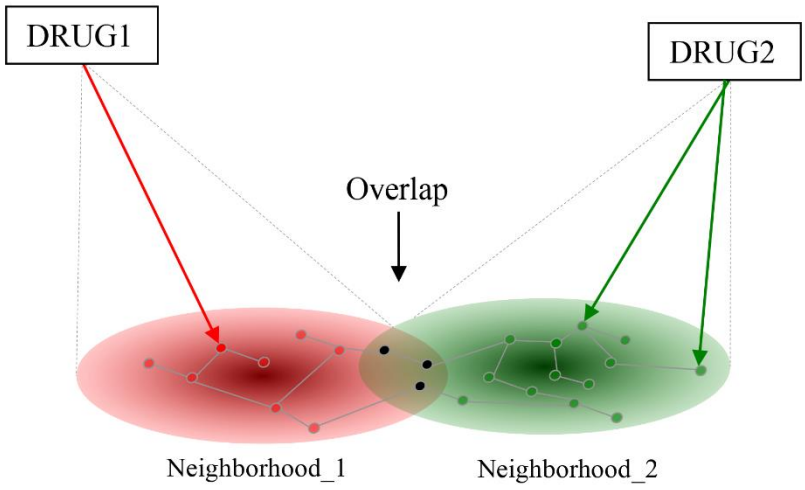

Figure 3.1. The network interaction hypothesis
The effect of two drugs ($Drug_1$, $Drug_2$) reaches their imminent targets first (arrows) and the effect will then propagate to their network neighborhoods (subnetworks) indicated in red and green, respectively. Targets in the overlap are affected by both drugs, and we suppose that drugs affecting a number of common targets will influence the effects of each other. The overlap is quantified as the proportion of jointly affected targets within all affected targets (in set theory terms: intercept divided by union).

We hypothesize that two (or more) drugs can have an unexpected combined effect if their perturbation neighborhoods overlap. In order to capture this property, I define a Target Overlap Score (TOS) for two drugs as the number of jointly affected targets divided by the number of all affected targets. This simple definition has a few plausible consequences: i) TOS has a value between zero and 1.0, higher values indicating stronger joint effects. ii) As a mathematical

DOI:10.15774/PPKE.ITK.2016.008

consequence, a drug will give TOS = 1.0 with itself. Note that even though a combination of two identical drugs does not occur in the clinical practice, it can cause a statistical bias in the comparisons so they have to be removed from the datasets used in the statistical comparison. iii) The concept of TOS can be generalized to more than two interacting drugs. Naturally, we have to decide in advance if, at one extreme, we want to consider genes perturbed by more than one agent in a drug combination only, or, at the other extreme, we consider just those genes that are perturbed by all of them. Here I used the former definition iv) The concept of TOS does not include any supposition about the beneficial or detrimental nature of combined drug effect. This is an important point since "drug interaction" in pharmacology denote negative, detrimental effect while the term "drug combinations" usually refer to beneficial, i.e. therapeutically useful combined effects. In principle, TOS can be correlated with both as I in fact show it in the next chapter. v) Finally, the definition of TOS is different from several other concepts related to traditional measures of drug interactions (antagonism, agonism, etc.) that mostly refer to effects of drugs on the activity of one target, such as a receptor. In contrast, TOS depends on the number of targets, and does not at present consider the magnitude (nor the positive or negative nature) of the effect.

From the practical point of view, a subnetwork is a graph that can be described in terms of its nodes, links or its substructures (known subgraphs). For instance, a node-based description can be the number of nodes, shared by the two subnetworks (network neighborhoods). Since network neighborhoods can be of very different sizes, it is safer to normalize this value to the total number of affected nodes, which takes us to the well-known formula of the Jacquard or Tanimoto coefficient, which I use as a target overlap score (TOS). The TOS of two subnetworks net_1 and net_2 is:

$$TOS(net_1, net_2) = \frac{\left|V_{net_1} \cap V_{net_2}\right|}{\left|V_{net_1} \cup V_{net_2}\right|} \tag{29}$$

This score can also be used for connections instead of nodes, and can be transformed into a weighted form. Namely, some algorithms assign weights to the nodes (edges). In such cases we can represent intersection and union by the sum of weights calculated for the participating nodes (edges) which then leads to the weighted variant of the overlap score. A probabilistic weighting is especially important, as it is generally applicable. In such a case, the weight assigned to a node can be the significance of the node (edge) derived from Monte-Carlo simulations; in this case, the overlap coefficient will have a statistical interpretation.

DOI:10.15774/PPKE.ITK.2016.008

Here I defined network neighborhood as the set of genes that are significantly perturbed by a drug. This was determined by Monte Carlo simulation, by repeating the diffusion process 10,000 times and determining the nodes (genes) whose activity changed at a chosen level significance (typically $p < 0.05$). As a numerical measure for drug-drug interaction, I define the Target Overlap Score (TOS) as the Jaccard coefficient (similarity measure between sets) calculated between the neighborhoods significantly affected by a pair of drugs.

Although several different propagation methods could be used for calculating the network neighborhood, I preferred the diffusion to PageRank, because it proved to be more robust in terms of changes in parameters, however, in some cases PageRank performed better. Furthermore, I tried to use as less parameters as possible to avoid the overfitting and to reduce the computation time needed for the optimization of the parameters.

TOS is 1.00 for a pair of drugs affecting the same targets and 0.00 for agents that do not significantly affect any target in common.

S_{DCM} , the perturbation of the i th drug D_i can be expressed as a vector:

$$S_{DCM}(D_i) = e^{-L_{\mu,\gamma}t} x(0) \ , \tag{30}$$

where the j th entry of $x(0)$ nonzero if it is targeted by the drug:

$$x_j(0) = \begin{cases} 1 & \text{if protein } j \text{ is drug target} \\ 0 & \text{otherwise} \end{cases} \tag{31}$$

The j th element of $S_{DCM}(\mathrm{D}_i)$ measures the disruption effect of D_i on protein j . I used the parameters $\mu = 0.1$ and $\alpha = 0.005$ throughout this study. Then the network neighborhood or subnetwork of drug D_i , net_{D_i} consists of the significantly perturbed network elements:

$$net_{D_i} = \left\{ v_i \mid v_i \in V, p_{v_i} < 0.05 \right\} \tag{32}$$

Then the target overlap score is calculated as in **equation 29.**

DOI:10.15774/PPKE.ITK.2016.008

Furthermore, this measure can be generalized to handle multiple drug combinations. For this purpose, one can determine the number of nodes that were significantly perturbed by at least two drugs, divided by the size of the affected subnetwork.

$$TOS(net_1, net_2, \cdots net_M) = \frac{\left| \bigcup_{i,j=1\cdots M; i \neq j} V_{net_i} \cap V_{net_j} \right|}{\left| \bigcup_{i=1\cdots M; i \neq j} V_{net_i} \right|} \tag{33}$$

This coefficient is zero if the neighborhoods do not overlap and 1.0 if they are identical. Identity is in fact a problem since a drug's overlap with itself is not meaningful. To avoid this problem, the drugs participating in the analysis should be pre-screened and drugs with identical targets should be excluded from the analysis, either a priori, or by omitting their combination from the results.

3.1.2. TOS is correlated with the strength of both beneficial and deleterious drug combinations

For the evaluation, I chose a simple ranking test, i.e. I compared the TOS value calculated for known drug pairs with the TOS of randomly chosen drug pairs, and calculated an AUC value for the ranking using ROC analysis [1] as described in **Section 1.6.5**.

It is noted that strong interactions are expected to give AUC values close to 1.0, while AUC values for randomly selected pairs are expected to be around ~ 0.5. In the present study, I used the STRING/STITCH interaction network, and the first question I asked was whether or not the evaluation system fulfils these fundamental criteria; for this purpose, I used the database of FDA-approved drugs [106] and generated all possible binary combinations. Trivial interactions (drugs acting on the same target and drug pairs with identical or nearly identical chemical structures) as well as drug pairs known to have positive or negative effects were omitted from the analysis which left 733542 pairs. This evaluation gave an AUC value of 0.48 (**Figure 3.2**, left) which is very close to the random value of 0.5 This finding thus shows that, given the TOS algorithm applied to the STRING/STICTH network, the randomly chosen FDA-approved drug pairs indeed behave as random. I have to mention that the randomly selected drug pairs may have contained cases in which the interaction has not been discovered yet. A related question is that of drugs having identical targets. These should by definition give a TOS value of 1.00, and I found 271 such drug pairs. Also,

DOI:10.15774/PPKE.ITK.2016.008

drugs having close to identical chemical structures are likely to affect similar targets. I found 179 such drug pairs but only eight of these were common with the previous subset. The comparison shows that both subsets give high TOS values, which will statistically bias the comparison if included either in the positive or in the negative dataset of non-interacting drugs. Therefore, for the statistical evaluation described below, I left out these drug pairs from both datasets.

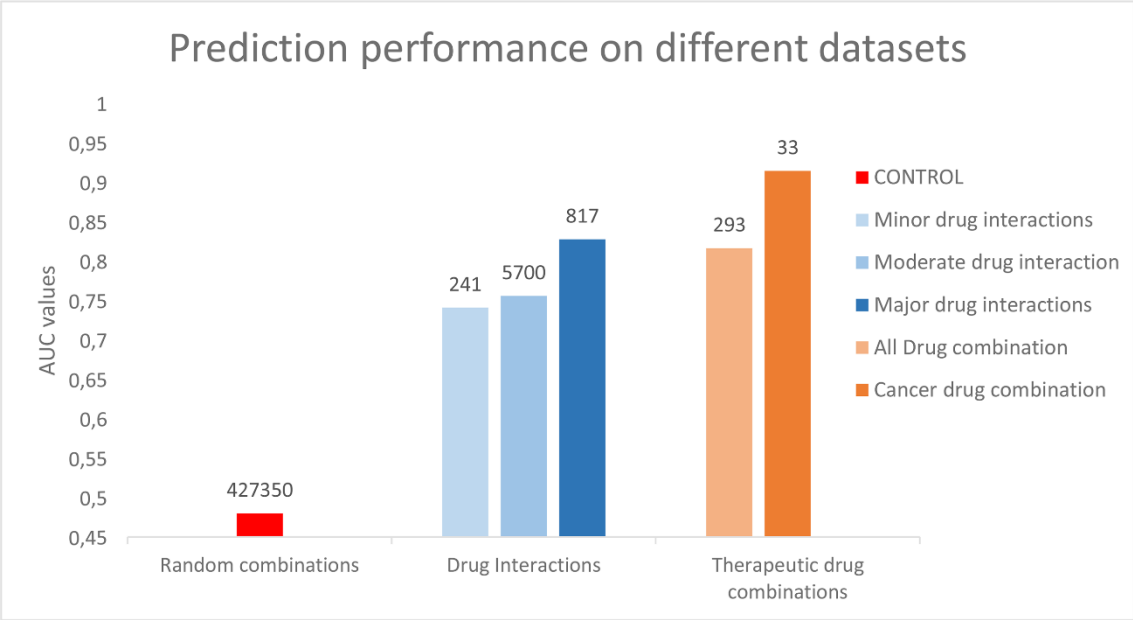

Figure 3.2. Prediction performance on known drug interactions and combinations
The prediction performance was measured on several different training sets (for details **see Table 3.1)**, cancer-related drug-drug interactions and drug combinations. The prediction method is based on a simple measure, the Target Overlap Score (TOS). The prediction procedure was repeated 100 times with different negative sets, and then the average value was reported. The standard deviation of AUC values (not shown) are between 0.0001 and 0.006 for the different datasets.

Next, I wanted to test whether or not TOS can help to identify the drug pairs that are empirically known to have a beneficial or detrimental effect. In pharmacology, two drugs are called "interacting" if their joint administration has a detrimental effect [250]. Drug pairs listed at http://drugs.com are classified into three groups according to the severity of the negative effects, such as major, moderate and minor. There are 10323 strongly, 92958 moderately and 17193 weakly interacting drugs in the database; however, in the selection I considered only cancer-related drug pairs, i.e. those, in which one of the agents was or was proposed to be used in treating cancer, which resulted in 817 strongly, 5700 moderately and 241 weakly interacting drugs from the database, denoted as sets A, B and C, respectively (**Table 3.1**). The results show that the interacting drug pairs show remarkably higher AUC values than the randomly selected drug pairs; moreover, these

DOI:10.15774/PPKE.ITK.2016.008

values qualitatively follow the strength of the interaction (**Figure 3.2**). Namely, strongly interacting drug pairs show substantially higher AUC values than the moderately interacting ones, etc.

Table 3.1. Datasets

	Dataset	Original size	Size after filtering[1]	Data source
Detrimental drug interactions[2]				
Severe	A	1053	817	Drugs.com
Moderate	B	6857	5700	Drugs.com
Minor	C	273	241	Drugs.com
Beneficial drug interactions[3]				
Total from DC[4]	D	429	293	DCDB, TTD
Cancer-related[5]	E	55	33	DCDB, TTD

[1] I filtered the available drug pairs by leaving out the drug combinations where the components have exactly the same targets, or the components were structurally similar, as described above. Drugs with no available targets were also discarded; [2]Taken from Drugs.com (November 11, 2013) as described below; [3]Taken from the Drug Combination Database (March 8, 2012) and the Therapeutic Target Database (July 23, 2012); [4]All approved drug combination were included; [5]All approved drug combination that are used in cancer treatment.

I also tested drug pairs that are known to have a beneficial effect when administered together. In pharmacology, the term "drug combinations" refers to drugs that are administered together because they have an empirically known beneficial therapeutic effect. Such therapeutically useful drug combinations are included in the Drug Combination Database (DCDB) [186] as well as in the Therapeutic Target Database TTD [223], along with the specific mechanism of their interaction. Using the same selection criteria, I selected 293 combinations (dataset D, **Table 3.1**). The results in **Figure 3.2,** right, show that therapeutic drug combinations yield AUC values substantially different from the random combinations.

Next, I carried out the same comparisons for all drugs. In this case, the datasets were naturally larger. The results show the same general tendencies as seen in the case of all drug combinations (data not shown). Namely, i) the known interactions are substantially different from the combinations of non-interacting drugs; ii) the AUC values of minor, moderate and strong, detrimental interactions follow the correct order, i.e. the stronger the interactions the higher the AUC values; and iii) the values of beneficial, therapeutic combinations are also substantially different from the average and the AUC value of 0.91 in cancer-related combinations can be considered especially convincing iv) in **Figure 3.2**, the beneficial interactions show higher AUC values. We have no ready explanation for this phenomenon; however, we speculate that one of the reasons could be that therapeutic combinations are usually optimized via careful clinical studies.

DOI:10.15774/PPKE.ITK.2016.008

3.1.3. TOS vs. GO and ATC codes

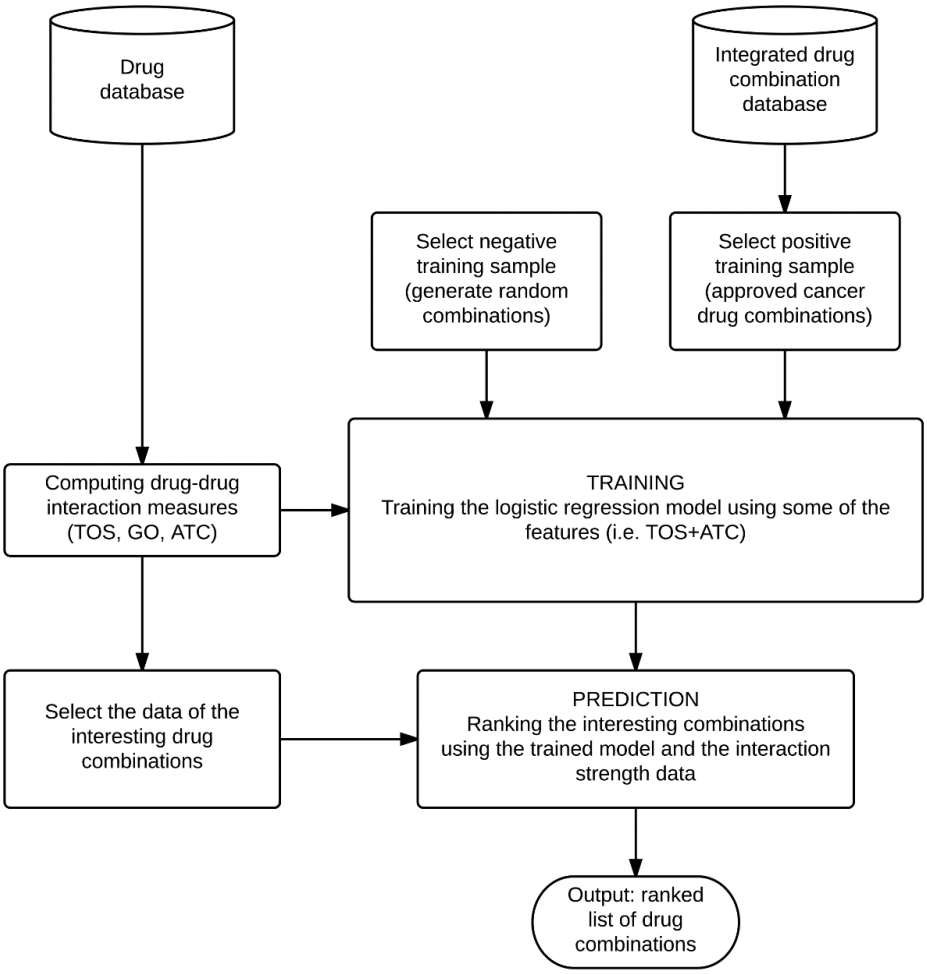

Figure 3.3. Flow chart of the training and the prediction procedure
The input is a list of candidate combinations (i.e. combinations selected for clinical trials) and the set of known combinations (i.e. previously approved cancer combinations). The first step is to compute the Target Overlap Score (TOS) and the drug interaction measures (GO, ATC) for all possible drug combinations. The database consists of the random generated drugs and of the components of the candidate and the known combinations. After the selection of the training sample (both the positive - known cancer combinations - and the negative one - random combinations) a logistic regression was trained using the previously computed TOS and similarity values. In the next step, the trained model is used for ranking a set of candidate combinations. The output is the ranked list of the drug combinations.

Since TOS is conceptually different from other measures used to characterize drug interactions, one might expect that additional parameters successfully used in other studies can increase its ranking power. The most obvious way of boosting the performance of a classifier is to include more and more relevant knowledge on the drugs. Earlier studies suggest that the integration of disease similarity [247] or therapeutic information such as ATC code based similarity [221, 246,

DOI:10.15774/PPKE.ITK.2016.008

251] as well as target similarity, such as GO annotations, could be useful as well [246, 249]. In order to test these possibilities, TOS was combined with GO or ATC based similarity metrics using logistic regression [252], a standard method in machine learning studies. The computational steps are summarized in **Figure 3.4**. I investigated various types of similarity measures defined in hierarchical networks (see **Section1.6.6**. for more details,) and I found that in the case of GO the simple cosine similarity, while in the case of ATC classification system the Resnik similarity produces the best results. Pairwise and ternary combinations of the three interaction measures (TOS, GO, ATC) were calculated by the logistic regression model [252]. Briefly, for a series $m_1, m_2, \ldots, m_n$ measures to be combined, the logistic regression model will calculate a combined measure M as

$$M\left(m_1, m_2, \ldots, m_n\right) = \frac{1}{1 + e^{\beta_0 + \sum \beta_i m_i}}, \tag{34}$$

where the β_i regression coefficients are estimated by linear regression.

The results in **Figure 3.4** show that inclusion of new parameters did not substantially change the picture.

The fact that the ranking power of TOS was not substantially improved when other parameters were added shows that TOS in itself captures a property that is well correlated with the empirically known interaction strength of various drug combinations.

3.1.4. TOS shows correlation with the outcome of clinical trials

In a clinical trial (also called „interventional study"), patients receive specific interventions according to a well-defined protocol [341]. In our case, trial data were collected from http://clinicaltrials.gov and consisted of studies in which combinations included trastuzumab either as an interaction partner or as a basis for comparison, and only those clinical scores were used that were collected according to RECIST [221]. The list of drugs tested in clinical trials included bevacizumab, capecitabine, carboplatin, cyclophosphamide, docetaxel, doxorubicin, epirubicin, fluorouracil, gemcitabine, ixabepilone, lapatinib, oxaliplatin, paclitaxel, pertuzumab, sunitinib.

DOI:10.15774/PPKE.ITK.2016.008

First, I analyze the statistical dependence between the clinical outcomes and the TOS values calculated for the drug regimens used for the treatment. Several regimens included more than two agents, such as trastuzumab and three additional drugs, A, B and C. Spearman's rank correlation coefficient was used for quantifying the statistical dependence between the TOS score and the clinical outcome measures. The TOS scores of the drug combinations are in **Table 3.2**.

.

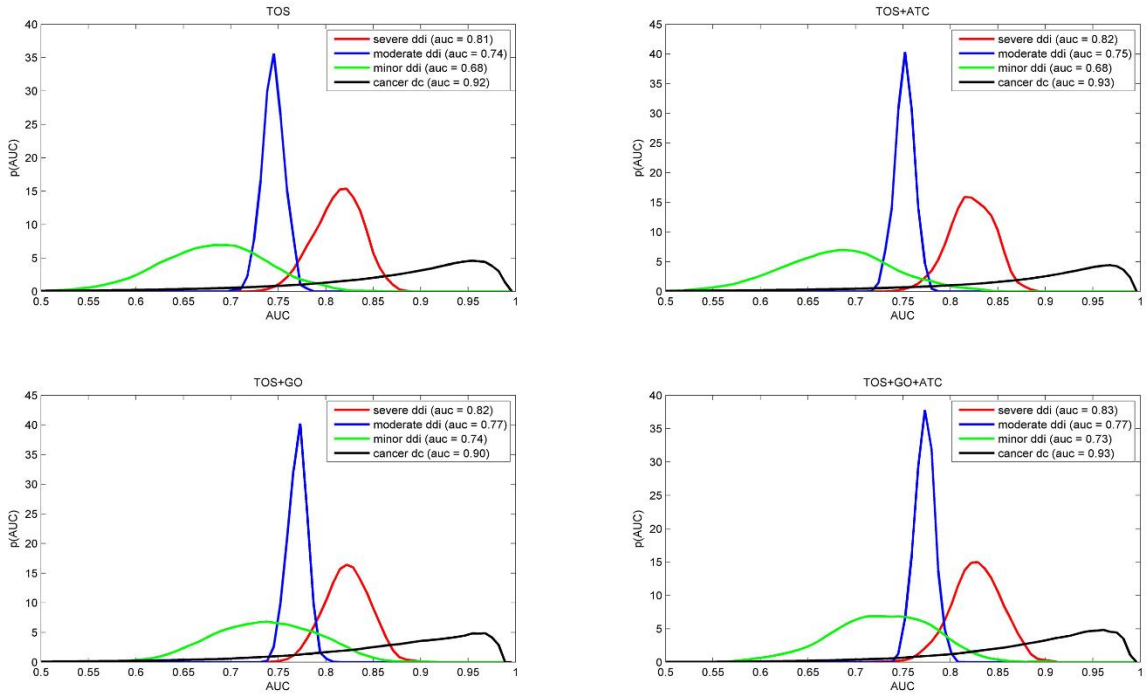

Figure 3.4. Performance of combined predictors on different training sets
The short titles TOS, TOS+ATC, TOS+GO or TOS+GO+ATC refer to the combination used. The curves represent the AUC value distribution (as a probability density function) obtained via a kernel density estimation (KDE) approach. The data were obtained by a 5-fold cross-validation procedure. Note that the distributions are quite similar to the TOS values (top left) which indicates that TOS effectively captures the drug combination phenomenon.

The TOS score shows substantial correlation with the overall response (OR) (r=0.64; p=0.0028), **Figure 3.5**. Furthermore, the overall survival rate (OSR) and Confirmed Clinical Benefit (CCB) correlate well with TOS r=0.87; p=0.017 and r=0.84; p=0.0021).

DOI:10.15774/PPKE.ITK.2016.008

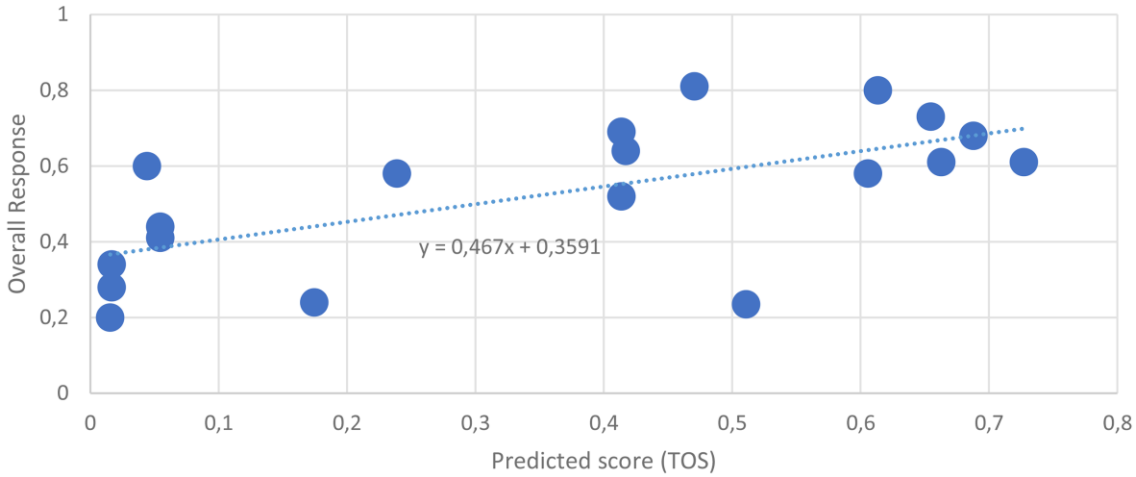

Figure 3.5 Scatter plot of prediction scores and Overall Response
The predicted scores are on the x axes, the clinical outcome, Overall Response (for the definition of outcome measures see the RECIST [35]) are on the y axes. Each data point corresponds to a multicomponent combination.

Table 3.2. TOS scores of binary and multicomponent combinations

Trastuzumab in binary combinations[1]	TOS score[2]	Trastuzumab in multiple combinations[1]	TOS score[2]
tra+doc	0.4138	tra+lap+5fu+cyc+epi+pac	0.7272
tra+gem	0.4067	tra+5fu+cyc+epi+pac	0.6629
tra+flu	0.3888	tra+doc+sun	0.6548
tra+pac	0.3732	tra+per+doc	0.6133
tra+dox	0.2509	tra+dox+doc	0.6058
tra+epi	0.1105	tra+cap+doc	0.5108
tra+cap	0.0845	tra+bev+doc	0.4707
tra+cyc	0.0806	tra+gem+car	0.4170
tra+ixa	0.0441	tra+ixa+car	0.0544
tra+oxa	0.0155		
tra+car	0.0110		

[1]All combinations presented here were under clinical investigation as of 1st of January 2013. Components in the combinations were lapatinib (lap), fluorouracil (5fu), cyclophosphamid (cyc), epirubicin (epi), paclitaxel (pac), pertuzumab (per), docetaxel (doc), carboplatin (car), doxorubicin (dox), gemcitabine (gem), carboplatin (car), ixabepilone (ixa), oxaliplatin (oxa); [2]The scores were computed using the generalized TOS. The larger the score the stronger the interaction.

In conclusion, the data suggest that there is a significant correlation between the TOS scores and the outcome of clinical trials.

DOI:10.15774/PPKE.ITK.2016.008

3.1.5. Discussion

The TOS score is based on the intuitive expectation that drugs perturbing overlapping neighborhoods within a gene network will combine their effects either in the positive or in the negative sense, and that the strength of the combined effect is proportional to the ratio of jointly affected targets within all affected targets. The TOS measure will detect the overlap, but a high TOS will not tell if the variations are caused by positive or by negative synergies. Our tests showed that TOS is in a consistently good correlation with both, and that this correlation could not be substantially strengthened by including GO and ATC terms.

A special advantage of TOS is the ability to rank potential drug combinations, so in addition to the best combination it can also show how the other potential combinations perform in a relative comparison. The examples in **Table 3.2** list cases where trastuzumab was combined with a cytotoxic drug (binary combination), or was part of a larger regimen consisting of more drugs which were studied in clinical trials. The results illustrate our message for trastuzumab: the highest ranking scores were achieved by combinations containing docetaxel. In addition, the most potent single agent to be administered with trastuzumab was also docetaxel. A few agents reached low scores (cyclophosphamide, oxaliplatin, carboplatin, ixabepilone) when applied together with trastuzumab. Nevertheless, interestingly, some of these, like cyclophosphamide and epirubicin, have achieved much higher scores when applied in complex regimens, which underlies the complex nature of the therapeutic response. In this context, it is worth to note that we apparently cannot yet estimate whether TOS can help to predict progression free survival, which is one the most important measure of clinical outcomes.

Even though the correlation of TOS with drug combination data is promising, its eventual use in predictive settings has important limitations. First, TOS relies on the protein-protein (or gene-gene) interaction data available in the databases. Though such data are accumulating at a growing pace, interactions missing from the current datasets may lead to erroneous predictions. An important property of the TOS score is that it can not by itself differentiate between positive and negative effects. Therefore, a high TOS value can mean either a positive, synergistic effect or a negative, deleterious drug interaction effect. As more information becomes available on the direction, strength and type (such as inhibition, activation, binding, etc.) of the interactions between the drug targets, some of the above limitations will be gradually eliminated. I also have to mention that completely or partly identical targets will by definition lead to high TOS values. While the

DOI:10.15774/PPKE.ITK.2016.008

former are trivial, the second may be worth to evaluate. The distinction is not built into the TOS score itself, but these cases can be identified by straightforward computations.

In addition, there is a conceptual difference between TOS and many of the other concepts of drug interactions. Namely, TOS does not limit drug-drug interactions or perturbations to identical drug targets or affected pathways. Instead, TOS captures a multitarget effect, and I think this is why the direction of the combined effect (i.e. beneficial vs. deleterious). can not be easily captured by the measure. Here a note on "beneficial" vs. "detrimental" effects is perhaps in place. Namely, many of the current, therapeutically useful drugs, including immunosuppressants or anti-cancer drugs are effective because they are toxic to a restricted population of cells. In the context of a single biochemical network, such effects would be considered as deleterious, even though in the therapeutic sense they are beneficial to the entire organism. I think it is the task of experimental studies to decide whether a combination with outstanding TOS score is therapeutically useful.

3.2. Prediction of cancer biomarkers by integrating text and data networks

Predicting disease biomarkers consists in suggesting genes potentially associated with a disease. Traditionally, gene-disease associations are based on experimental data that have been validated by careful clinical studies. With the emergence of high-throughput techniques, it is possible to experimentally compare the behavior of all human genes in healthy and diseased states. However, the evaluation of such lists is not simple [253]. Computational methods of "gene prioritization" were developed for this purpose [254]. Most of the methods combine the new experimental data with a background database containing information on co-occurrence, functional annotations, protein-protein interactions, pathways, and gene expression. Briefly, we can view new experimental data as numerical scores assigned to genes, and the background database as a network of genes in which the links are defined by one of the methods mentioned above. In the process of gene prioritization, the experimental scores are updated using the gene network data and the genes are re-ranked based on the new scores. Updating of scores can be based on graph distance (shortest path), on a propagation algorithm such as the popular PageRank [255] or on diffusion methods [203], for example. The resulting methods differ in the kind of score updating methodology, the background database used, and most importantly, the size of the data they can handle. Relatively few methods can select genes from entire genomes or accept input data on all genes. For instance,

DOI:10.15774/PPKE.ITK.2016.008

it is customary to restrict the scope of candidate genes to a small region of the chromosome using methods of linkage analysis or to use known disease genes as a training set. One of our goals is to use approaches analogous to the methods of gene prioritization in order to further increase the sensitivity of hypothesis generation.

Text mining has been successfully applied to finding various gene-disease associations [256], such as suggesting disease marker genes from MEDLINE records and ranking (prioritizing) genes based on biomedical literature [257]. Reviews of the earlier work are found in [258] and [259]. More recently, Hristovski and associates combined DNA microarray data and semantic relations extracted from MEDLINE, for generating novel hypotheses [260]. Frijters and colleagues also presented an application of their literature mining method in an open-ended retrieval of hidden relations for hypotheses in terms of gene-disease, drug-disease and drug-biological process associations [261].

3.2.1. Theory

In the framework of knowledge discovery, a biomarker is a hypothesis generated from the evaluation of scientific literature. In the ideal case, hypothesis generation leads to a list of hypotheses that can be ranked according to various criteria. Generally speaking, a hypothesis is a previously unknown, indirect connection between term A (disease) and term C (cause). Knowledge discovery posits that such a hypothesis is validated if both the cause and the disease are related to the same set of intermediate concepts, which is the basis of the well-known ABC model of Swanson. Hypothesis generation is a different task as it seeks to identify novel hypotheses rather than confirming one. The RaJoLink model of Petrič et al [262] approaches this problem by looking at “rare terms”, i.e. concepts sporadically appearing in the literature that may be linked to common, hitherto unknown causes. In an ideal case, the causes emerging in this manner can be ranked by importance [3]. The same principle can be used to select potential biomarkers. In this case, the starting phenomenon is a disease (i.e. the same as before) but the terms we are looking at are the names or symbols of genes or pathways. The approach presented in this section relies on the supposition that genes (pathways) sporadically mentioned in the scientific literature my point to a set of genes (pathways) that is the cause of the disease, so mutations in these genes (pathways) may then be used as biomarkers for the disease. The goal of this section is to show how a common data network of scientific abstracts and protein-protein interactions can be used to automate this process. Namely, in the data network described in the previous sections diseases are linked to genes (as well

DOI:10.15774/PPKE.ITK.2016.008

as drugs). In this section, a new type of link, "co-occurrence" was added, denoting that a disease and a gene are mentioned in the same scientific publication. On such an enhanced network, the problem of hypothesis generation can be defined as a neighborhood-overlap problem as shown in **Figure 3.6**. In this section, ovarian cancer was used as the test case.

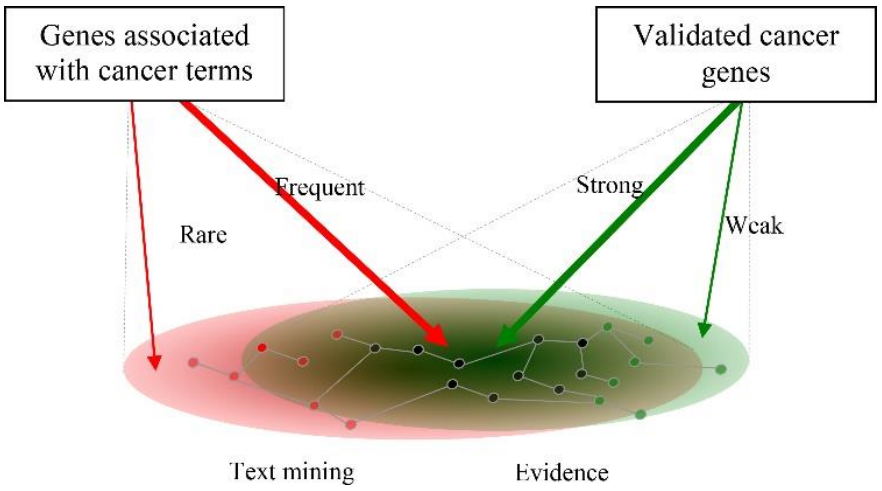

Figure 3.6. The principle of biomarker prediction using terms rarely associated with cancer and a set of validated genes

A "hypothesis" is a gene worth to be experimentally tested. Such a gene (network node) is expected to rarely - but appreciably – associate with cancer terms in scientific papers (light red area), but has no, or no strong evidence for cancer involvement (light green area). In the picture, such genes are located within the intersection of the light-shaded areas, and the ones of interest are identified by ranking them according to a suitable criterion. The principle of biomarker prediction using terms rarely associated with cancer and a set of validated genes.

3.2.2. Constructing a data-network with molecular and literature-based links

The benchmark datasets were designed to test whether or not a method could efficiently predict that a gene plays a certain role, which was then experimentally confirmed later. For this test, a corpus of abstracts is needed being published before a certain biological role was confirmed. I chose ovarian cancer as the model disease and used a recently published list of 37 ovarian cancer biomarker (OC biomarkers) genes [263] (**Table 3.3**) as test cases. I then wanted to determine if the relationship between these genes and ovarian cancer could have been predicted on the basis of literature published beforehand. In order to have a sufficient number of genes in the analysis, I selected the year 2007 as a separating line. A total of 10 OC biomarkers have been proposed after this date.

DOI:10.15774/PPKE.ITK.2016.008

OC biomarker abstracts were selected using the search phrase*: (biomarker OR biomarkers OR marker OR markers) AND (“cancer of ovary" OR "ovary cancer" OR "cancer of the ovary" OR "ovarian cancer" OR "malignant neoplasm of ovary" OR "malignant ovarian neoplasm" OR "malignant tumor of ovary" OR "malignant tumor of the ovary" OR "malignant neoplasm of the ovary" OR "malignant ovarian tumor" OR "malignant tumor of ovary" OR "ovarian malignancy" OR "ovarian carcinoma")*. This search resulted in 4,878 abstracts published before the year 2007. This set was defined as the OC biomarker test corpus. Separately, 26,979 abstracts about the known OC biomarker genes [263] (**Table 3.3**) published up until May 14th 2012 were obtained and these formed the OC biomarker prediction corpus. HGNC gene symbols, names and their synonyms were used (downloaded on December 23rd 2011). Such HGNC nomenclature was then applied to the terms that were automatically extracted from collections of MEDLINE abstracts.

Table 3.3. List of ovarian cancer biomarker genes published before May 2012

	Symbol	Gene		Symbol	Gene		Symbol	Gene
1	CA125	CA 125 [264-267]	14	P16	p16 [268, 269]	26	BIRC5	Survivin [270]
2	KRT19	Cytokeratin 19 [271, 272]	15	CDKN1A	p21 [273-275]	27	TERT	hTERT [276]
3	KLK6	Kallikrein 6 [277]	16	CDKN1B	p27 [278-281]	28	EGFR	ERBB1 [282, 283]
4	KLK10	Kallikrein 10 [284]	17	RB1	pRB [285, 286]	29	ERBB2	ERBB2 [287]
5	IL6	Interleukin-6 [288]	18	E2F1	E2F1 [289]	30	MET	c-Met [290]
6	IL7	Interleukin-7 [291]	19	E2F2	E2F2 [292]	31	MMP2	MMP-2 [293]
7	IFNG	γ-interferon [294]	20	E2F4	E2F4 [292]	32	MMP9	MMP-9 [295]
8	FAS	sFas [296, 297]	21	TP53	p53 [298, 299]	33	MMP14	MT1-MMP [300]
9	VEGFR	VEGFR [301]	22	TP73	p73 [302]	34	SERPINB5	Maspin [303]
10	CCND1	Cyclin D1 [273, 304]	23	BAX	Bax [305, 306]	35	BRCA1	BRCA1 [307]
11	CCND3	Cyclin D3 [308]	24	BCL2L1	Bcl-xl [309]	36	ERCC1	ERCC1 [310]
12	CCNE	Cyclin E [311-314]	25	BIRC2	cIAP [315]	37	WFDC2	Epididymis protein 4 [316-318]
13	P15	p15 [319]						

3.2.3. Principle of evaluation

From the mathematical point of view, genes selected by text mining analysis can be viewed either as an unranked set of gene names or as a ranked list wherein genes are characterized by their names as well as by a numerical score. I used two kinds of methods for re-ranking the genes selected by the enhanced RaJoLink rare-term algorithm described here: a) standard gene prioritization methods available via gene prioritization web servers (ToppGene and Endeavour) [320, 321] and

DOI:10.15774/PPKE.ITK.2016.008

b) propagation-based methods that were implemented on the STRING database [322], as briefly described in **Section 1.6.1** and **1.6.3.**

More specifically, the PageRank iteration was initiated from the known disease-associated genes, biomarkers, thus the vector pr is defined as:

$$pr_i = \begin{cases} \frac{1}{M} & \text{if protein i is a known OC biomarker} \\ 0 & \text{otherwise} \end{cases}, \quad (35)$$

where M is the number of validated biomarkers.

Similarly, the diffusion process was initiated from known biomarker genes:

$$x_i(0) = \begin{cases} 1 & \text{if protein i is a known OC biomarker} \\ 0 & \text{otherwise} \end{cases} \quad (36)$$

3.2.4. Testing the methods on the rediscovery of known OC biomarker genes

The goal is to establish whether the genes that have been proposed as OC biomarkers after 2007 could have been predicted on the basis of prior literature evidence and knowledge. Genes suggested as biomarkers are those that co-occurred with the term “marker” or “biomarker” in MEDLINE abstracts. MEDLINE abstracts, MeSH and HUGO terms published before 2007 and standard propagation algorithms (PageRank or diffusion kernel methods [203, 255]) were used for re-ranking the results, using the network of the STRING database, release version 6.3 (in use from December 12, 2005 to January 15, 2007). In the re-ranking step I could not use the gene-prioritization servers as the current servers contain information entered after 2007.

Out of the 37 ovarian cancer genes listed in **Table 3.3**, 27 are mentioned together with “marker” or “biomarker” in MEDLINE articles published before 2007. The remaining 10 genes (the target genes) are: BCL2L1, CCND3, E2F1, E2F2, E2F4, ERCC1, IL7, MET, MMP9, WFDC2. Six genes were identified with the enhanced RaJoLink method. For five of these six genes, the ranks could be substantially improved by propagation/re-ranking (**Table 3.4**).

DOI:10.15774/PPKE.ITK.2016.008

Table 3.4. List rediscovery of genes suggested as OC biomarkers

Gene symbol	Gene	Year when first mentioned as ovarian cancer prognostic marker	Rank			
			Original RaJoLink	New RaJoLink	New RaJoLink + PageRank	New RaJoLink + Personal Diffusion
BCL2L1	Bcl-xl	2007	NA	337	5	10
CCND3	Cyclin D3	2007	NA	165	43	31
E2F1	E2F1	2008	NA	NA	NA	NA
E2F2	E2F2	2007	69	140	36	3
E2F4	E2F4	2007	39	16	NA	NA
ERCC1	ERCC1	2007	NA	NA	NA	NA
IL7	Interleukin 7	2007	54	297	82	80
MET	c-Met	2007	NA	NA	NA	NA
MMP9	MMP-9	2007	44	86	22	49
WFDC2	Epididymis protein 4	2009	NA	NA	NA	NA

3.2.5. Prediction of new OC biomarkers

I wanted to establish if any putative gene biomarkers might exist for ovarian cancer on the basis of currently available published knowledge. To achieve this, an experiment similar to the previous one was completed where i) the data input was the ovarian cancer prediction corpus which includes abstracts about the known OC biomarker genes [263] (**Table 3.3**) published up until May 14, 2012 and current versions of STRING, MeSH, HUGO nomenclature data, and ii) the propagation step was carried out with the standard propagation algorithms (PageRank or diffusion kernel methods [203, 255]), and also with the gene prioritization servers ToppGene and Endeavour [320, 321]. It was apparent that a number of well-known cancer-related genes appear in the top of these lists.

For a better overview, I compared the top of the lists and picked 10 genes that ranked highly in most of the rankings (**Table 3.5**). These include RUNX2, SOCS3, BCL6, PAX6, DAPK1, SMARCB1, RAF1, E2F6, P18INK4C (CDKN2C), and PAX5. These are all cancer-related genes that have not previously been proposed as OC biomarkers and have not been mentioned in literature

DOI:10.15774/PPKE.ITK.2016.008

sources together with ovarian cancer. These may represent genetic markers upon which hypotheses can be formulated in relation to ovarian cancer.

Table 3.5. Predicted OC biomarker genes

Genes predicted as ovarian cancer biomarkers		Validation after May, 2012
Gene symbol	**Gene**	
RUNX2	Runt-related transcription factor 2	[323, 324]
SOCS3	Suppressor of cytokine signaling 3	NA
BCL6	B-cell lymphoma 6 protein	[325, 326]
PAX6	Paired box protein Pax-6	NA
DAPK1	Death-associated protein kinase 1	NA
SMARCB1	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily B member 1	NA
RAF1	RAF proto-oncogene serine/threonine protein kinase	NA
E2F6	Transcription factor E2F6	NA
P18INK4C (CDKN2C)	Cyclin-dependent kinase 4 inhibitor C	NA
PAX5	Paired box protein Pax-5	NA

3.2.6. Discussion

Here I applied an open knowledge discovery method that generates hypotheses that are not known in advance. The methodology suggests potential disease-gene associations based on text databases (in our case MEDLINE), MeSH terms, and the HUGO nomenclature. It was found that unequivocal matching to gene symbols is a very important factor, and that re-ranking text-based predictions either by standard propagation algorithms (PageRank, Diffusion Ranking) applied to the STRING network, or by gene-prioritization servers available on the web can improve the efficiency of text-mining searches.

The searches revealed a number of genes that were previously not associated with progression and prognosis of ovarian cancer in MEDLINE abstracts. The RUNX2 transcription factor is a putative tumor suppressor gene localized at chromosome 1p36, a region showing frequent loss of heterozygosity events in colon, gastric, breast and ovarian cancers [327]. RUNX2 has also been associated with prostate [328], lung [329], breast cancer [330], osteosarcoma [331], thyroid tumors [332]. Several other studies demonstrated a link between RUNX2 and the hormonal system in prostate [333] and breast cancer [334]. All these data suggest a possible contribution of RUNX2 to proliferation via enhancing the growth factor effects of sexual hormones. The potential

DOI:10.15774/PPKE.ITK.2016.008

of the gene in this association is supported by the prognostic power of hormone receptors in ovarian cancer [335]. In 2012 it was also confirmed that RUNX2 is associated with advanced tumor progression in epithelial OC [324]. In addition, the inhibition of RUNX2 lead to significant decrease of cell proliferation [336].

BCL6 (B-cell CLL/lymphoma 6) is another transcription factor found to be frequently mutated in diffuse large-cell lymphoma. The gene was related not only to lymphomas [337] and leukemias [338], but also to progression to breast [339], gastric [340] and lung cancer [341]. Interestingly, both BCL6 [342] and RUNX2 [343] are influenced by prolactin secretion. Wang et al. showed that BCL6 is a negative prognostic factor in ovarian cancer [325] and the inhibition of BCL6 along with NACC1 [326] reduced the invasion capabilities of cancer cells.

The tumor suppressor DAPK1 (death-associated protein kinase 1) is one of the key regulators of the extrinsic apoptotic pathway [344]. Genetic variation of DAPK1 is associated with survival in breast cancer [345]. The methylation status of DAPK1 contributed to stratifying colon cancer patients into subgroups with different prognosis [346]. The correlation between apoptotic machinery and DAPK expression has just recently been validated in the ovarian cancer cell line OVCAR-3 [347]. Taken together, DAPK1 might be a potent prognostic marker for predicting apoptotic activity and tumor progression in various cancer types including ovarian cancer.

Of the remaining top-ten candidate genes, PAX6, E2F6, SMARCB1 and PAX5 also regulate gene transcription, while SOCS3, RAF1 and P18INK4C play a role in signal transduction pathways. In summary, the identified genes modulate key elements of molecular pathways of tumorigenesis and have already been associated with the progression in various malignancies.

DOI:10.15774/PPKE.ITK.2016.008

3.3. Inference on hierarchical graphs: fast and sensitive alignment of microbial whole metagenome sequencing reads

Simply put, annotation of metagenomics sequencing reads consists of two essential parts: i) mapping reads to taxa and determining the taxon (strain, species, genus, etc.) composition, and ii) mapping reads to functions and determining the functional repertoire (functional composition) of the community. These two tasks can be carried out by the same computational step. Namely, a read is mapped to a specific location within a genome. The genome name is mapped to the taxon so we have the answer for i). The location within the genome is mapped to a function so we have the answer for ii). The problem is that the mappings are part of a large and complex data network represented in a data-warehouse, which may not guarantee a sufficiently fast access to the required information. The solution is to construct a small, dedicated database that only contains the required data. Importantly, both taxonomy and function are defined to hierarchical graphs, so by using a database of appropriate format we can reduce the problem to fast tree-operations. For this, we need to design and benchmark an efficient algorithm, as described in the subsequent chapters.

The work described in this chapter is a close collaboration with Lőrinc S. Pongor (Semmelweis University) and Roberto Vera (Pázmány University) [348]. The Taxoner algorithm was developed by Lőrinc, the database necessary for fast execution was developed by Roberto. My own contribution was to design the conceptual scheme for the underlying data-network, to develop test cases for the testing and the statistical evaluation. My work also included the selection and implementation of programs used for comparison [82, 85, 88, 349] .

3.3.1. Taxoner algorithm

The idea behind Taxoner comes from a technical problem. Running fast aligners such as Bowtie2 on a large number of microbial genomes is prohibitively time consuming, since – at least in principle – each of the small genomes has to be indexed separately. However, if we concatenate the small bacterial original genomes into larger units, i.e. concatenated FASTA files that are termed “artificial chromosomes”, the problem becomes more manageable. In such an artificial chromosome, a genome is a segment that is annotated by various identifiers including taxonomic name and GI identifier. As such the number of reads matching a particular genome can be counted at various taxonomic levels which corresponds to the well-known principle of taxonomic binning [350]. The only prerequisite is to know the starting and endpoints of the genomes and/or other

DOI:10.15774/PPKE.ITK.2016.008

segments incorporated into the “artificial chromosome”, which is solved by pre-calculated index files.

Importantly, this process is analogous to the mapping of reads to an annotated genome wherein the segments – i.e. the genes – are named according to such schemes as COG [351, 352], GO [353], etc. Namely, in both cases, a read is mapped to a large sequence consisting of annotated segments, and the segments are named according to various ontologies. As a consequence, this algorithm can be used both for taxon identification and for function prediction based on NGS datasets. It must be highlighted that mapping of counts to ontologies, sometimes also referred to as “ontology binning”, is a problem known in other fields of medical informatics as well [354]. Further analogies can be found in protein sequence similarity searching wherein BLAST hits (HSPs) are mapped to domains annotated within proteins [355-359].

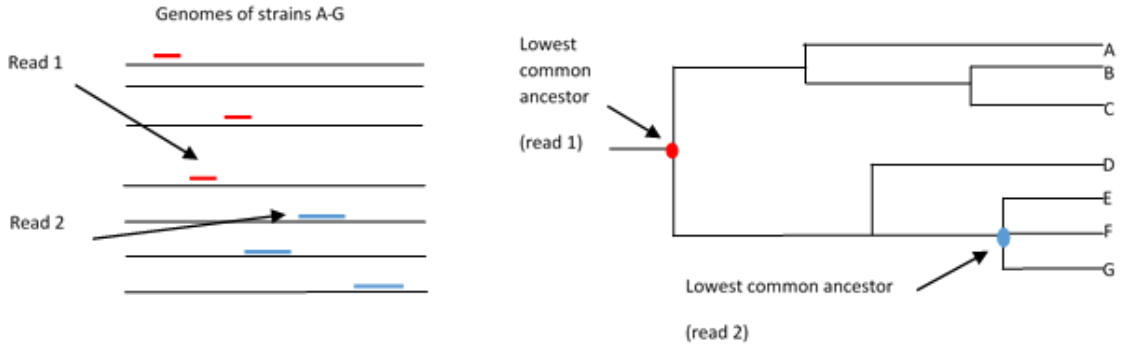

Figure 3.7. The Taxoner algorithm
In the first step, the short reads were mapped to the microbial genomes. Next, the alignments were preprocessed and only those hits were mapped to the taxonomy tree that were above a certain threshold. In the classification or the binning step, the read was assigned to the lowest common ancestor of the taxa it hit.

The algorithm has two computational phases. Firstly, as a preprocessing, a database of an arbitrary number of species is segmented into partitions (“artificial chromosomes”), then the Bowtie2-build program of the Bowtie2 package is used to index them [360]. Pre-built indices are available on the project site (http://code.google.com/p/taxoner). Phase II is the alignment itself for which the Bowtie2 was used. The lowest common ancestor algorithm provided the base for the identification of the taxa (**Figure 3.7**Hiba! A hivatkozási forrás nem található.). The output, which is a summary of the taxa and the alignments, is provided in SAMtools format.

DOI:10.15774/PPKE.ITK.2016.008

Furthermore, it is also possible to process the read alignments and the output with other programs such as MEGAN [85].

Table 3.6. Benchmark datasets

Dataset	ID	Sequencing platform and number of spots; average read length	Taxon	Note
A	SRR292150	454 GS 20 (183203;110.31)	*Staphylococcus aureus* subsp. *aureus* USA300_TCH959 (NCBI taxon id: 450394)	Randomly selected[1], Supplementary file[2]
B	ERR236069	Ion Torrent PGM (1338465;262.05)	*Staphylococcus aureus* (NCBI taxon id: 1280)	Randomly selected[1], Supplementary file[2]
C	SRR017390	Illumina Genome Analyzer II (26391487;76)	*Staphylococcus aureus* subsp. *aureus* 67-331 (NCBI taxon id: 585131)	Randomly selected[1], Supplementary file[2]
D	DRR000184	Illumina Genome Analyzer II (7631281;50)	*Bacillus anthracis* BA104 (NCBI taxon id: Not Available)	Randomly selected[1], Supplementary file[2]
E	DRR000184	Illumina Genome Analyzer II (7631281;50)	*Bacillus anthracis* BA104 (NCBI taxon id: Not Available)	Whole run[3]
F	AE017225	NA(104574;99.99)	*Bacillus anthracis* str. Sterne (NCBI taxon id: 260799)	Full genome sampling[4], supplementary file[2]
G	SRX055380	Illumina Genome Analyzer II (6562065; 75.00)	HMP Mock Community even sample	Whole genome sequencing

[1]Random selected datasets were produced by a Python script that uniformly sampled the read collection without replacement. The sample size was 100000. [2]Supplementary files are deposited at http://pongor.itk.ppke.hu/taxoner/examples/ [3]The whole run was analyzed. [4]The genome was sampled with overlapping reads. The read length was uniformly 100bp and the overlap between the adjacent reads was 50bp.

3.3.2. Execution times

The evaluation of Taxoner's performance was carried out by comparing it to MetaPhlAn [88], BLAST (legacy blast) [361], and BLAST+ (dc-megablast) [362]. In the last two cases, the

DOI:10.15774/PPKE.ITK.2016.008

MEGAN taxon assignment program was used [83, 85]. The former was chosen for the speed and accuracy it shows in the estimation of taxon composition, and the latter one for its outstanding performance in alignment. However, comparing the performance of the above-mentioned programs is not trivial for various reasons; for instance, MetaPhlAn has its own database with 367 million nucleotides including only bacteria, and it reads to this taxon-marker database. The other two, on the other hand, can run on more complex databases, like NCBI NT, which contains 52 000 million nucleotides of all species, or on a bacterial subset as well. These factors also impact the speed and the accuracy of the results. As a conclusion, it is clear that the size of the database, the number of threads used for the calculation, and the length and the number of the reads to be evaluated define the actual alignment times for the three programs. The typical results of the comparisons are shown in **Table 3.7**. As can be seen, Taxoner's performance is in between the running time of Blast and MetaPhlAn, meaning Taxoner being fast enough to be performed on an ordinary desktop PC or a laptop, although the indexing of the database requires more time.

Table 3.7. Average execution times of alignments

	Taxoner[1] nt database[4]	**MEGABLAST[2] nt database[4]**	**MetaPhlAn[3] unique marker db[5]**
1 thread[6]	2446 sec	37.6h	7 sec
12 threads[6]	1866 sec	9.4h	6 sec

[1]Running times include the read classifications; [2]Read classifications are not included; [3]Running times include the read classifications; [4]The nt database size is 1507MB (as of 11/07/2013); [5]The database size is 367MB; the measurement was carried out on an Intel(R) Xeon(R) CPU E5-2640 processor, the query reads: Dataset A, **Table 3.6**.

3.3.3. Compatibility with various sequencing platforms

The development of sequencing technologies has given rise to a number of sequencing platforms in recent years. The performance of read aligners often varies on the basis of reads produced by the various sequencing platforms. I compared the performance of aligner programs on read datasets selected from *Staphylococcus aureus* sequencing data (dataset A, B and C from **Table 3.6** with results shown in **Table 3.8**). BLAST aligners had the highest alignment rate, which is not surprising since BLAST is a sensitive local aligner. MetaPhlAn had the lowest alignment rates, which is again expected, since MetaPhlAn only aligns reads to a unique subset of the bacterial database (i.e. clade-specific markers). In general, all aligners had the best performance on the 454 dataset, since the 454 reads are longer and thus easier to analyze. All programs performed well at

DOI:10.15774/PPKE.ITK.2016.008

genus level. An important aspect of metagenome analysis is the identification of taxa at the lowest possible taxonomic levels, such as species or strains. At the species level, Taxoner showed the best performance with the exception of 454 reads where BLAST performed better. On the other hand, none of the programs identified strains very reliably (strain level identification is not even available with MetaPhlAn). The discrepancies may however be caused by the fact that the sequencing data used in this comparison were partly taken form uncharacterized strains (dataset B, Ion Torrent data) or from a strain not in the database of *Staphylococcus aureus* subsp. *aureus* 67-331 (NCBI taxon id: 585131, dataset C, **Table 3.6**, Illumina data).

Table 3.8. Read assignment for *Staphylococcus aureus* genome sequencing data

		Roche 454 (Dataset A)[1]			Ion Torrent (Dataset B)[1]			Illumina (Dataset C)[1]		
	Level:	**Genus**	**Species**	**Strain**	**Genus**	**Species**	**Strain**	**Genus**	**Species**	**Strain**
Taxoner	Total	93692	93692	93692	37175	37175	37175	27531	27531	27531
	Positive	93189	92728	62	36482	35919	0	26023	17019	0
	Negative	4	43	875	28	126	17174	29	121	1213
	FNR %[2]	0.004	0.046	0.934	0.075	0.339	46.198	0.105	0.440	4.406
Meta-PhlAn	Total	8525	8525	8525	2522	2522	2522	1692	1692	1692
	Positive	8209	8063	N/A[3]	2402	2399	0	1650	1613	N/A
	Negative	0		N/A	43	36	0	2	28	N/A
	FNR %	0.000	0.493	N/A	1.705	1.427	N/A	0.118	1.655	N/A
BLAST ALL + MEGAN	Total	86752	86752	86752	68696	68696	68696	45721	45721	45721
	Positive	83718	82951	156	65264	63801	0	41189	27375	0
	Negative	25	29	53	114	125	5310	408	441	3094
	FNR %	0.029	0.033	0.061	0.166	0.182	7.730	0.892	0.965	6.767
DC-MEGABLAST + MEGAN	Total	84211	84211	84211	64858	64858	64858	48677	48677	48677
	Positive	81000	80035	140	61662	60199	0	44161	29466	0
	Negative	48	68	131	94	128	3052	81	180	3421
	FNR %	0.057	0.081	0.156	0.145	0.197	4.706	0.166	0.370	7.028

[1]100,000 random selected reads from experimental data. [2]False Negative Rate. [3]Not available.

3.3.4. Detecting an unknown anthrax strain

As different bacterial strains may have quite similar sequences, it is essential to know whether the database contains the given genome or not. This question is also motivated by the fact that the microorganisms detected in environmental samples are mainly unknown strains. Being an unknown strain here means not having the genome or draft genome deposited in any well-known sequence database. To test how the algorithms handle the above-mentioned case, the detection

DOI:10.15774/PPKE.ITK.2016.008

probability of two *Bacillus anthracis* strains was compared to each other. I chose Sterne strain as the "known" strain (NCBI taxon id: 260799), which is used for vaccination. The dataset contained 100bp long overlapping segments ("artificial reads") of its own genome, offset by 50bp (dataset F). A Japanese isolate (*B. anthracis* strain BA104; NCBI taxon id: Not Available) represented the unknown strain being not included in the database at the time of the analysis. The dataset contained 7.7 million Illumina reads (dataset E). It is not surprising that the proportion of false negative results (number of misclassified reads) was higher for the unknown strain as shown in **Figure 3.8.**

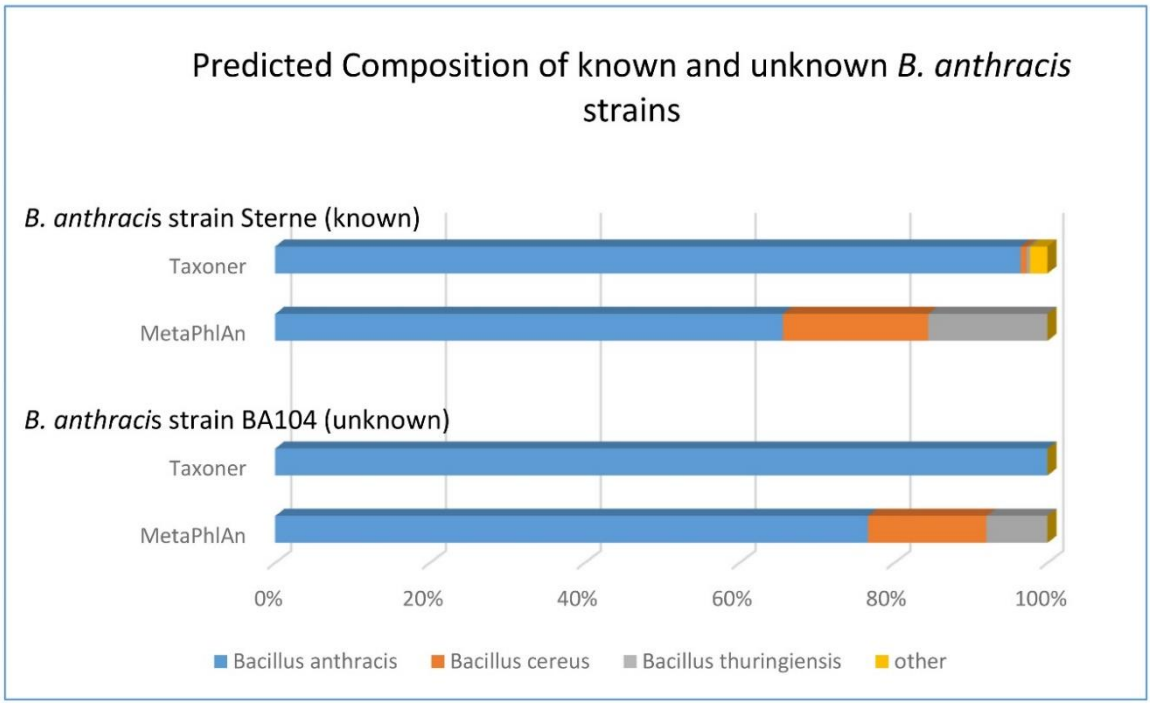

Figure 3.8. Analysis of *anthracis* strain not included in the database
The read classifications predicted by MetaPhlAn and Taxoner on an artificial dataset (**Table 3.6**, dataset F) made from a known anthrax strain and an unknown anthrax strain (**Table 3.6**, dataset E) was compared. Here, the percentages of classified reads on species level are reported. Both sample consists of exactly one bacterial species. The dataset F contains 104574 synthetic reads, while the Dataset F (**Table 3.6**) 7,379,118 reads.

As expected, Taxoner could perfectly detect the synthetic reads that were generated without introducing any error from a genome being included in the database. It is surprising, however, that MetaPhlAn identifies, even in the synthetic reads, a considerable amount of species that are not present in the samples. It must be noted that *B. anthracis* is a member of the *B. cereus* group. This includes three related species: *B. cereus*, *B. thuringiensis* and *B. anthracis*. These three are highly related, which is also supported by the fact that about 75% of the synthetic reads generated from the Sterne strain are 100% identical in them (data not shown). The difference between the three

DOI:10.15774/PPKE.ITK.2016.008

programs is that MetaPhlAn assigns these reads to the *Bacillus* genus, while the other two assign them to the *B. cereus* group. This illustrates that both the database and the taxonomy definitions of the given programs deeply influence the species % reported by them.

3.3.5. Detection of very low abundance reads

When it comes to detecting pathogens, one of the most important questions is the sensitivity of the test, which can be estimated from the number of reads safely detected by the analysis. I tried to estimate this value by randomly selecting a fixed number of reads from an experimental anthrax dataset (dataset E) in conjunction with the NT dataset (**Figure 3.9).** Number of reads necessary on average to detect an unknown species at various taxonomy levels. Error bars indicate standard deviation of the mean, calculated from 400 repetitions.

MetaPhlAn was used in conjunction with its own dataset. The analysis was repeated for each set of data. It is apparent that at species level, Taxoner/Bowtie2 performed better than BLAST or MetaPhlAn. At genus and higher levels, all methods performed well even though it was apparent that MetaPhlAn needed one-two orders of magnitude more reads to complete the identification than the other programs. Taxoner and BLAST were able to detect a genus essentially from one read while MetaPhlAn needed 200-350 reads on average for the identification. Species level identification was apparently more difficult for all programs, with Taxoner and BLAST needing 10 or 15 reads for the identification.

3.3.6. Analyzing metagenomic datasets

Analysis of metagenomic data from NGS reads has two goals: a) establishing the presence or absence of taxa at the lowest possible taxonomic level and if possible at species/strain level, and b) estimating the relative amounts of taxa.

I also investigated on the MOCK community samples published by the Human Microbiome Project. These are synthetic microbiomes consisting of 21 known microbial species with known composition that were used to develop sequencing protocol. There are two different composition types available: one is uniform, i.e. each strain concentration is the same, the other is staggered, where the concentrations are not constant. Each community was sequenced using various approaches and sequencing platforms (16S, WGS-Illumina).

DOI:10.15774/PPKE.ITK.2016.008

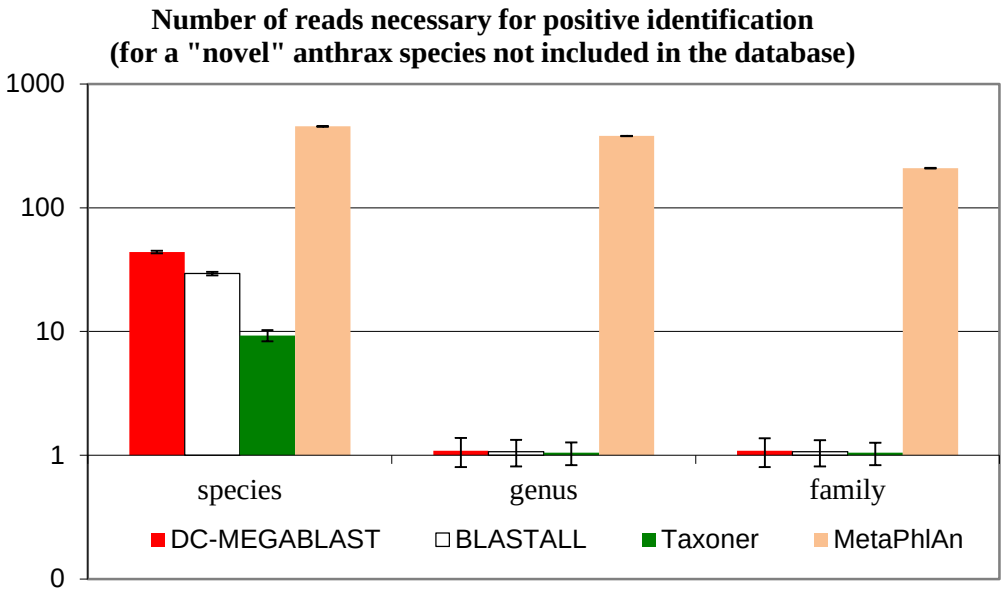

Figure 3.9. Detection of low abundance strains

Number of reads necessary on average to detect an unknown species at various taxonomy levels. Error bars indicate standard deviation of the mean, calculated from 400 repetitions.

Table 3.9. Identification of taxa in even MOCK community

		strain	**species**	**genus**	**family**
Number of positives (taxa present)[1]		22	22	19	18
Taxoner[6]	TP[2]	14	20	17	17
	FP[3]	7	2	2	1
	FN[4]	8	0	0	0
	F-measure[5]	**0.65**	**0.95**	**0.94**	**0.97**
MetaPhlAn	TP	NA	21	18	18
	FP	NA	1	1	0
	FN	NA	7	5	6
	F-measure	NA	0.84	0.86	0.86
WGSQUICKR	TP	1	9	13	13
	FP	20	13	6	5
	FN	79	67	45	29
	F-measure	0.02	0.18	0.34	0.43

[1]The analysis was run on dataset G, **Table 3.6**; [2]True positive; [3]False positive; [4]False negative; [5]The F-measure was calculated as 2TP/(2TP+FP+FN); [6]Positive taxa predicted by Taxoner are those that received at least 1000 hits.

DOI:10.15774/PPKE.ITK.2016.008

The analysis revealed that Taxoner is able to detect a taxon in a complex community (**Table 3.6**, dataset G) even in strain level (**Table 3.9**); however, MetaPhlAn (and WGSQUICKR) is not. The classification accuracy of Taxoner and MetaPhlAn is comparable in this task. In order to improve the accuracy in Taxoner a minimum threshold was applied for the number of reads needed for reporting a taxa as positive, omitting the filtering step leaded to high number of false positive identifications. MetaPhlAn and other programs, such as Megan, also use similar strategies. I also investigated a recently published tool, WGSQUICKER [349], that uses compressed sensing technique applied on genome composition information for detecting the taxa presenting in a sample. However, the approach is extremely fast, the number of false positive and negative identifications is large.

3.3.7. Discussion

Here I presented a novel approach for the analysis of whole genome shotgun metagenomics results. Taxoner utilizes a Burrow-Wheeler Transform (BWT) based fast aligner and a comprehensive sequence database along with a comprehensive taxonomy tree in order to interpret and efficiently analyze the massive sequencing data. The improved indexing strategies made it possible to evaluate metagenomics datasets in a single PC, without compromising the classification performance. Taxoner performed better in speed and accuracy than BLAST-based tools.

Detecting unknown strains is often a problematic task for aligners. It is proved that strains of the same species may have high variability in their genome. As a consequence, the analysis of, for example, soil bacteria may involve the detection of strains being largely unknown to the current databases. It follows from this that Taxoner using a comprehensive database has an advantage compared to other approaches using marker databases. It can be explained by the fact that new strains may not contain the unique sequences that are included in the marker database. The above-described problem is rather important because strain level identification is needed for the detection of hazardous pathogens. Only a few programs include this feature (strain level identification), and Taxoner is one of them. The fact that its sensitivity is much better than that of the pipelines based on BLAST is a result of the program using Bowtie2 instead of BLAST. These results implicate that combining fast alignment techniques with comprehensive databases can end up in a proper alternative for sensitive analysis of metagenomic samples. Lastly, it is shown that pathogen identification is a basically different task from metagenome analysis thus requiring specific treatment. There are three features in Taxoner that help pathogen identification: i) the feature of

DOI:10.15774/PPKE.ITK.2016.008

removing the reads originated from the host organism (by specifying a host genome);ii) an output in a format that is compatible with Megan and can be submitted to it, thus one can manually identify pathogens, and iii) the possibility of relying on dedicated databases. For example, after creating a specialized database one can directly use this for the analysis (or, naturally, download a standard database from our repository).

DOI:10.15774/PPKE.ITK.2016.008

4. Conclusions and new scientific results

In this work I investigated how different graph models can help in various bioinformatics problems. My research covered the following areas: i) finding novel drug combinations, ii) finding novel, unexpected biomarkers from literature, and iii) improving the classification performance of metagenomics reads.

In the first case the data network consists of protein-protein associations and drug-protein associations, and I used other types of data networks as well, like hierarchical data networks (GO and ATC classification system). In the second application the basic data network was built from literature information and protein-protein associations. Finally, in the third case, taxonomy tree and Clusters of Orthologous Groups (EggNOG) were used. Both are hierarchical networks; the former one more or less represents the phylogeny, the latter provides a resource for functional annotations for microorganisms.

However, there is a basic difference between the motivation of the above approaches: in the last case network-based methods are used to help interpreting our experimental data by reducing noise and finding reliable results by revealing the possible enriched functions in the diverse bacteria cocktail; while in the first two cases the goal is to find novel, unexpected relationships between various objects by applying the concept of network neighborhood on different data networks.

In the second application the network neighborhood was selected by using a static approach as described in **Section 1.6** and **3.2.** The genes being in the static neighborhood are considered to be related to the disease and represent a large body of new hypotheses. However, these predictions are noisy, but other data networks, such as protein-protein interactions, could be used efficiently to improve the prediction accuracy in terms of ranking. The task of reordering interconnected elements is a very common task in networks and the traditional searching algorithms such as PageRank and diffusion processes proved to be useful.

In spite of the similar goal, namely to reveal unexpected relationships between different objects, in the first case the network neighborhood was constructed not in a static, but in a dynamic way using also the diffusion processes as described in **Section 1.6** and **3.1.1**.

Despite of the different motivations, I attempted to address the problem with the help of hierarchical networks both in the first and the third case. It proves to be useful in annotation processes (i.e. in metagenomic problems) and in finding similarities, but does not help with ranking drug combinations.

DOI:10.15774/PPKE.ITK.2016.008

4.1. Network neighborhood analysis – revealing unexpected relationships

The idea underlying network neighborhood analysis is seemingly simple: a concept, such as a molecule or a pathway represented in a biological database, is not a single object but a subnetwork of interrelated concepts and relationships. From this, it trivially follows that subnetworks can overlap with each other so we can significantly broaden the scope of associations between concepts and extend the analysis of hidden, implicit links, which is the essence of new discoveries. What is not trivial is how to design a network in which associations will be useful; in other terms, we can answer practical questions with the help of it. The suggestion put forward in the first two subsections of **Section 3** is that we construct a data network dedicated for a given purpose. Namely, if we want to query associations between diseases, drugs and drug targets, we construct a network consisting from these items, by combining, say drug databases (STITCH [111], DrugBank [107], TTD [223], DCDB [113]), interaction databases (STRING [4], IntAct [137]), disease databases (OMIM [363]), and various resources such as ontologies [89, 90] and manually curated datasets. Alternatively, if we want associations based on text mining, we include a network composed of our useful terms (say, diseases, target genes) and text-mining based links between them. Such dedicated data networks take some expertise to construct, but the time of network construction and analysis are not prohibitively long. What is questionable, of course, is how good our data are. Here we have no guarantees for success, just the hope that the body of databases and the number of new database types will continue to increase as fast as it does today, and that novel types of integration methodologies will emerge. Currently, a bottleneck in the construction of data networks is data heterogeneity, namely the concepts are not uniformly defined across the various databases we want to integrate in a network. With these caveats in mind, these approaches should be considered as pilot studies into two seemingly unrelated directions, the prioritization of drug combinations, and prediction of potential biomarkers.

Since hypothesis generation based on genomic data is a key problem in life sciences today, this approach can also be used in other fields. The limitations of this approach follow by the probabilistic nature of the answers. For instance, I considered the prediction of a drug combination successful if the successful combination was in the toplist of say 10 best hits. Since the number of potential drug candidates is very high, such a ranking can be considered a partial success, since one can narrow down the experiments to a relatively small number of cases. On the other hand, we expect that semantic pruning of the network may improve the efficiency of predictions in the future.

DOI:10.15774/PPKE.ITK.2016.008

Namely, one may design useful rules regarding which links of the networks should be omitted from the analysis. In such a manner the size and complexity of the network could be decreased so more sophisticated algorithms could be used for the analysis. In view of the efforts invested into biomedical ontologies, we can trust that this development will broaden the scope of the network analysis technique proposed here.

4.1.1. Prediction of efficient drug combinations

In the first part of my thesis I showed that molecular interaction data can successfully predict known combinations of chemotherapeutic agents used to treat breast cancer. Here the prediction is a ranking, in which the efficient combinations are expected to be in the top of the list. The performance, namely how good a ranking is, was characterized with the AUC value. This score is 1 if the ranking is perfect (i.e. all efficient combination ranked at the top), 0.5 if it is random. Drug - drug interactions are often considered as harmful "negative combinations", since they increase the risk of side effects and may cause "overdose". On the other hand, drug combinations are considered to be desirable (positive) since they can be efficiently used in the treatment of complex diseases. We could show that a simple network overlap measure is well correlated with the intensity of positive and negative drug interactions as well as with clinical data.

Thesis group I.

Related publications of the author: [J1][J3][C1]

THESIS I.1. *I have developed a novel drug combination prediction method based on the assumption that a perturbation generated by multiple drugs propagates through an interaction network and the drugs may have unexpected effect on targets not directly targeted by either of them* ***(Figure 3.1****). I introduced a new index, the so-called Target Overlap Score (TOS), to capture this phenomenon. The score quantifies the potential amplification effect as the overlap between the affected subnetworks. The score is computed as the Jacquard or Tanimoto coefficient between the sets of nodes in the subnetworks (for details see* ***Section 3.1.1****).*

DOI:10.15774/PPKE.ITK.2016.008

THESIS I.2. *I have showed that using the TOS score it is possible to distinguish both the drug-drug interactions and the drug combinations from random combinations. I also presented that this measure is correlated with the known effects of beneficial and deleterious drug combinations taken from the DCDB, TTD and Drugs.com databases (**Figure 3.2**).*

THESIS I.3. *I have also investigated that combining two frequently used drug-drug similarity measures with TOS - namely the functional similarity of drugs computed based on their imminent targets, and their therapeutic similarity quantified by using the anatomical therapeutic chemical (ATC) classification system - does not improve the classification performance.*

THESIS I.4. *I have demonstrated the utility of TOS by correlating the score to the outcome of recent clinical trials evaluating trastuzumab, an effective anticancer agent used in combination with anthracycline- and taxane-based systemic chemotherapy in HER2-receptor (erb-b2 receptor tyrosine kinase 2) positive breast cancer.*

4.1.2. Prediction of cancer biomarkers by integrating text and data networks

In biomarker prediction I showed that novel biomarkers can be prioritized using a network built from text mining data as well as ovary cancer data. In particular, I found that new biomarkers discovered in a given period of time are correlated with gene names sporadically emerging in the oncological literature of the previous years.

Since many current medical hypotheses are formulated in terms of molecular entities and molecular mechanisms, here the methodology is extended to proteins and genes using a standardized vocabulary as well as a gene/protein network model. The proposed enhanced RaJoLink rare-term model combines text mining and gene prioritization approaches. Its utility is illustrated by finding known, as well as potential gene-disease associations in ovarian cancer using MEDLINE abstracts and the STRING database.

Thesis group II.

Related publications of the author: [J1][J6]

THESIS II.1. *I have improved the sensitivity of the RaJoLink rare term based algorithm by using network analysis algorithm such as personalized diffusion ranking and PageRank with Prior on the STRING protein-protein association network.*

DOI:10.15774/PPKE.ITK.2016.008

THESIS II.2. *Based on the enhanced prediction I have proposed 10 novel genes - RUNX2, SOCS3, BCL6, PAX6, DAPK1, SMARCB1, RAF1, E2F6, P18INK4C (CDKN2C), and PAX5 - that are likely to be related to the disease and at the time had not been described as such. Since 2012, two of them (RUNX2, BCL6) have been confirmed.*

4.2. Fast and sensitive characterization of microbial studies

Next generation sequencing (NGS) of metagenomic samples is becoming a standard approach to detect individual species or pathogenic strains of microorganisms. Computer programs used in the NGS community have to balance between speed and sensitivity and as a result, species or strain level identification is often inaccurate and low abundance pathogens can sometimes be missed. In metagenome analysis I showed that a dedicated database, consisting of taxonomic and gene classification data mapped to whole genome sequences, can be successfully used to quickly identify both species composition and functional repertoire in metagenomic samples. When used in conjunction with fast sequence alignment methods, Taxoner data network provides a more accurate identification of species compositions than standard marker-based methodologies such as MetaPhlAn.

In the recent years various sequencing platforms have been developed. I compared the classification performance of Taxoner, MetaPhlAn [88] and BLAST [361, 362] combined with Megan [83, 85] on whole genome sequencing datasets of *Staphylococcus aureus* produced by Roche 454, Ion Torrent and Illumina. The low false negative rates implicate that Taxoner is almost as reliable as BLAST+Megan, however, it requires much less computational power and time.

I analyzed the MOCK dataset representing 22 microbial strains and species in equal amounts provided by the Human Microbiome Project for validation purposes. The dataset consists of 6.5 Illumina short-reads. Taxoner was capable of confidently detecting most of the taxa (14/22) even in strain level.

In the application to pathogen detection the sensitivity of the analysis is a crucial question. The sensitivity was measured as the number of reads necessary for detecting a certain species. After randomly sampling an experimental anthrax dataset the analysis revealed that Taxoner could confidently identify the anthrax from 10 reads, while MetaPhlAn needed 200-350 reads.

DOI:10.15774/PPKE.ITK.2016.008

Sensitive detection of microorganisms with unknown sequence is a crucial question as well, since the majority of them are still unknown. In order to assess the classification performance on unknown species (their genome sequence is missing from the database) I analyzed an experimental anthrax dataset (*B. anthraci*s strain BA104; NCBI taxon id: Not Available). Taxoner classified the majority of reads (96.50%) as *Bacillus anthracis*, s small portion 1.2% was classified as other species from the *Bacillus* genus.

Thesis group III.

Related publications of the author: [J5]

THESIS III.1. *I have demonstrated that by using hierarchical networks such as taxonomy along with fast aligners, i.e. bowtie2, the evaluation of high-throughput sequencing data is feasible in a reasonable time with good classification accuracy. The algorithm assigns the individual reads to the common ancestor of the taxa having its genome hit by the short read (Hiba! A hivatkozási forrás nem található.).*

THESIS III.2. *I have illustrated the applicability of the Taxoner principles on whole genome shotgun sequencing of known or unknown pathogens (Staphylococcus aureus, Bacillus anthracis). The results suggested that the performance of Taxoner is as good as the state-of-the-art BLAST based methods, while it is faster by two orders of magnitude. Finally, it is also compatible with various sequencing platforms.*

THESIS III.3. *I have proved that using the Taxoner principles it is possible to characterize the microbial communities at the lowest taxonomic level, even in species or strain level.*

THESIS III.4. *Taxoner is sensitive and capable of identifying taxa being present only in small abundance; furthermore, it needs two orders of magnitude less reads to complete the identification than MetaPhlAn. In addition, the method is applicable in cases where the genome sequence of the studied microbe is unknown.*

DOI:10.15774/PPKE.ITK.2016.008

5. Publications

The author's publications include journal papers (marked as J) and conference proceedings papers (marked as C).

[J1] **Ligeti, B.,** Menyhárt, O., Petrič I., Győrffy, B.; Pongor, S. (2016). Propagation on Molecular Interaction Networks: Prediction of Effective Drug Combinations and Biomarkers in Cancer Treatment. *Current Pharmaceutical Design*, accepted.

[J2] **Ligeti, B**.; Vera, R.; Juhász, J.; Pongor, S. (2016). CX, DPX and PCW: Web servers for the visualization of interior and protruding regions of protein structures in 3D and 1D. *Springer Protocols: Methods in Molecular Biology*, in press.

[J3] **Ligeti, B.**; Pénzváltó, Z.; Vera, R.; Győrffy, B.; Pongor, S. (2015). A Network-Based Target Overlap Score for Characterizing Drug Combinations: High Correlation with Cancer Clinical Trial Results. *PLoS One*. **10** (9), e0129267.

[J4] Hudaiberdiev, S.; Choudhary, K.; Vera, R.; Gelencsér, Zs.; **Ligeti**, **B**.; Lamba, D.; Pongor, S. (2015); Census of solo LuxR genes in prokaryotic genomes. *Front. Cell. Infect. Microbiol*. 5:20. doi:10.3389/fcimb.2015.00020

[J5] Pongor, L. S.; Vera, R.; **Ligeti, B**. (2014). Fast and Sensitive Alignment of Microbial Whole Genome Sequencing Reads to Large Sequence Datasets on a Desktop PC: Application to Metagenomic Datasets and Pathogen Identification. *PLoS One*, published 31 Jul 2014, 10.1371/journal.pone.0103441

[J6] Petrič, I.; **Ligeti**, **B**.; Győrffy, B.; Pongor, S. (2014). Biomedical Hypothesis Generation by Text Mining and Gene Prioritization. *Protein Pept Lett*. 20, 1-1.

[C1] **Ligeti**, **B**.; Vera, R.; Lukács, G.; Győrffy, B.; Pongor, S. (2013). Predicting effective drug combinations via network propagation. *Biomedical Circuits and Systems Conference*, 378-381.

[J8] Vera, R.; Perez-Riverol, Y.; Perez, S.; **Ligeti, B**.; Kertész-Farkas, A.; Pongor, S. (2013). JBioWH: an open-source Java framework for bioinformatics data integration. *Database*. 2013, bat051.

DOI:10.15774/PPKE.ITK.2016.008

6. References

[1] P. Sonego, A. Kocsor, and S. Pongor, "ROC analysis: applications to the classification of biological sequences and 3D structures," *Brief Bioinform,* vol. 9, pp. 198-209, May 2008.

[2] A. Pavlopoulou, D. A. Spandidos, and I. Michalopoulos, "Human cancer databases (Review)," *Oncology reports,* vol. 33, pp. 3-18, 2015.

[3] I. Petric, B. Ligeti, B. Gyorffy, and S. Pongor, "Biomedical hypothesis generation by text mining and gene prioritization," *Protein and peptide letters,* vol. 21, pp. 847-857, 2014.

[4] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, *et al.*, "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Res,* vol. 41, pp. D808-15, Jan 2013.

[5] R. Page and C. Takimoto, "Principles of chemotherapy," *Cancer Management: A Multidisciplinary Approach Medical, Surgical & Radiation Oncology. Editors: R Pazdur, LR Coia, WJ Hoskins, LD Wagman. PRR, New York,* pp. 21-38, 2004.

[6] J. Hirsch, "An anniversary for cancer chemotherapy," *Jama,* vol. 296, pp. 1518-20, Sep 27 2006.

[7] L. Kelland, "The resurgence of platinum-based cancer chemotherapy," *Nat Rev Cancer,* vol. 7, pp. 573-584, 08//print 2007.

[8] Z. Penzvalto, A. Lanczky, J. Lenart, N. Meggyeshazi, T. Krenacs, N. Szoboszlai, *et al.*, "MEK1 is associated with carboplatin resistance and is a prognostic biomarker in epithelial ovarian cancer," *BMC Cancer,* vol. 14, p. 837, 2014.

[9] Z. Penzvalto, P. Surowiak, and B. Gyorffy, "Biomarkers for systemic therapy in ovarian cancer," *Curr Cancer Drug Targets,* vol. 14, pp. 259-73, 2014.

[10] G. Minotti, P. Menna, E. Salvatorelli, G. Cairo, and L. Gianni, "Anthracyclines: molecular advances and pharmacologic developments in antitumor activity and cardiotoxicity," *Pharmacol Rev,* vol. 56, pp. 185-229, Jun 2004.

[11] J. Jolivet, K. H. Cowan, G. A. Curt, N. J. Clendeninn, and B. A. Chabner, "The pharmacology and clinical use of methotrexate," *N Engl J Med,* vol. 309, pp. 1094-104, Nov 3 1983.

[12] E. A. Perez, "Microtubule inhibitors: Differentiating tubulin-inhibiting agents based on mechanisms of action, clinical activity, and resistance," *Mol Cancer Ther,* vol. 8, pp. 2086-95, Aug 2009.

[13] Y. Q. Liu, W. Q. Li, S. L. Morris-Natschke, K. Qian, L. Yang, G. X. Zhu, *et al.*, "Perspectives on biologically active camptothecin derivatives," *Med Res Rev,* vol. 35, pp. 753-89, Jul 2015.

[14] L. Gatti and F. Zunino, "Overview of tumor cell chemoresistance mechanisms," *Methods Mol Med,* vol. 111, pp. 127-48, 2005.

[15] B. Gyorffy, P. Surowiak, O. Kiesslich, C. Denkert, R. Schafer, M. Dietel, *et al.*, "Gene expression profiling of 30 cancer cell lines predicts resistance towards 11 anticancer drugs at clinically achieved concentrations," *Int J Cancer,* vol. 118, pp. 1699-712, Apr 1 2006.

[16] R. W. Johnstone, A. A. Ruefli, and S. W. Lowe, "Apoptosis: a link between cancer genetics and chemotherapy," *Cell,* vol. 108, pp. 153-64, Jan 25 2002.

[17] F. Zunino, P. Perego, S. Pilotti, G. Pratesi, R. Supino, and F. Arcamone, "Role of apoptotic response in cellular resistance to cytotoxic agents," *Pharmacol Ther,* vol. 76, pp. 177-85, Oct-Dec 1997.

[18] A. A. Stavrovskaya, "Cellular mechanisms of multidrug resistance of tumor cells," *Biochemistry (Mosc),* vol. 65, pp. 95-106, Jan 2000.

[19] S. G. Chaney and A. Sancar, "DNA repair: enzymatic mechanisms and relevance to drug response," *J Natl Cancer Inst,* vol. 88, pp. 1346-60, Oct 2 1996.

[20] G. Munkacsy, R. Abdul-Ghani, Z. Mihaly, B. Tegze, O. Tchernitsa, P. Surowiak, *et al.*, "PSMB7 is associated with anthracycline resistance and is a prognostic biomarker in breast cancer," *Br J Cancer,* vol. 102, pp. 361-8, Jan 19 2010.

[21] W. E. Evans, Y. Y. Hon, L. Bomgaars, S. Coutre, M. Holdsworth, R. Janco, *et al.*, "Preponderance of Thiopurine S-Methyltransferase Deficiency and Heterozygosity Among Patients Intolerant to Mercaptopurine or Azathioprine," *J Clin Oncol,* vol. 19, pp. 2293-2301, April 15, 2001 2001.

[22] C. M. Ulrich, K. Robien, and H. L. McLeod, "Cancer pharmacogenetics: polymorphisms, pathways and beyond," *Nat Rev Cancer,* vol. 3, pp. 912-20, Dec 2003.

[23] C. Ekhart, S. Rodenhuis, P. H. Smits, J. H. Beijnen, and A. D. Huitema, "An overview of the relations between polymorphisms in drug metabolising enzymes and drug transporters and survival after cancer drug treatment," *Cancer Treat Rev,* vol. 35, pp. 18-31, Feb 2009.

[24] B. Tegze, Z. Szallasi, I. Haltrich, Z. Penzvalto, Z. Toth, I. Liko, *et al.*, "Parallel evolution under chemotherapy pressure in 29 breast cancer cell lines results in dissimilar mechanisms of resistance," *PLoS One,* vol. 7, p. e30804, 2012.

[25] M. Ando, H. Saka, Y. Ando, H. Minami, T. Kuzuya, M. Yamamoto, *et al.*, "Sequence effect of docetaxel and carboplatin on toxicity, tumor response and pharmacokinetics in non-small-cell lung cancer patients: a phase I study of two sequences," *Cancer Chemother Pharmacol,* vol. 55, pp. 552-8, Jun 2005.

[26] C. E. DeSantis, S. A. Fedewa, A. Goding Sauer, J. L. Kramer, R. A. Smith, and A. Jemal, "Breast cancer statistics, 2015: Convergence of incidence rates between black and white women," *CA Cancer J Clin,* vol. 66, pp. 31-42, Jan 2016.

[27] C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, *et al.*, "Molecular portraits of human breast tumours," *Nature,* vol. 406, pp. 747-752, 08/17/print 2000.

[28] D. Hanahan and Robert A. Weinberg, "Hallmarks of Cancer: The Next Generation," *Cell,* vol. 144, pp. 646-674, 3/4/ 2011.

[29] Y. Chen, J. McGee, X. Chen, T. N. Doman, X. Gong, Y. Zhang, *et al.*, "Identification of druggable cancer driver genes amplified across TCGA datasets," *PLoS One,* vol. 9, p. e98293, 2014.

[30] A. Urruticoechea, R. Alemany, J. Balart, A. Villanueva, F. Vinals, and G. Capella, "Recent advances in cancer therapy: an overview," *Curr Pharm Des,* vol. 16, pp. 3-10, Jan 2010.

[31] G. Saglio, D. W. Kim, S. Issaragrisil, P. le Coutre, G. Etienne, C. Lobo, *et al.*, "Nilotinib versus imatinib for newly diagnosed chronic myeloid leukemia," *N Engl J Med,* vol. 362, pp. 2251-9, Jun 17 2010.

[32] B. J. Druker, F. Guilhot, S. G. O'Brien, I. Gathmann, H. Kantarjian, N. Gattermann, *et al.*, "Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia," *N Engl J Med,* vol. 355, pp. 2408-17, Dec 7 2006.

[33] F. Li, C. Zhao, and L. Wang, "Molecular-targeted agents combination therapy for cancer: developments and potentials," *Int J Cancer,* vol. 134, pp. 1257-69, Mar 15 2014.

[34] S. Kummar, M. Gutierrez, J. H. Doroshow, and A. J. Murgo, "Drug development in oncology: classical cytotoxics and molecularly targeted agents," *British Journal of Clinical Pharmacology,* vol. 62, pp. 15-26, 05/30

01/18/received

04/26/accepted 2006.

[35] S. R. Park, M. Davis, J. H. Doroshow, and S. Kummar, "Safety and feasibility of targeted agent combinations in solid tumours," *Nat Rev Clin Oncol,* vol. 10, pp. 154-168, 03//print 2013.

[36] D. J. Slamon, G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, and W. L. McGuire, "Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene," *Science,* vol. 235, pp. 177-82, Jan 9 1987.

[37] S. Dawood, K. Broglio, A. U. Buzdar, G. N. Hortobagyi, and S. H. Giordano, "Prognosis of women with metastatic breast cancer by HER2 status and trastuzumab treatment: an institutional-based review," *J Clin Oncol,* vol. 28, pp. 92-8, Jan 1 2010.

[38] A. Awada, I. Bozovic-Spasojevic, and L. Chow, "New therapies in HER2-positive breast cancer: a major step towards a cure of the disease?," *Cancer Treat Rev,* vol. 38, pp. 494-504, Aug 2012.

[39] C. L. Vogel, M. A. Cobleigh, D. Tripathy, J. C. Gutheil, L. N. Harris, L. Fehrenbacher, *et al.*, "Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer," *J Clin Oncol,* vol. 20, pp. 719-26, Feb 1 2002.

[40] F. Montemurro, A. Prat, V. Rossi, G. Valabrega, J. Sperinde, C. Peraldo-Neia, *et al.*, "Potential biomarkers of long-term benefit from single-agent trastuzumab or lapatinib in HER2-positive metastatic breast cancer," *Mol Oncol,* vol. 8, pp. 20-6, Feb 2014.

[41] D. Graus-Porta, R. R. Beerli, J. M. Daly, and N. E. Hynes, "ErbB-2, the preferred heterodimerization partner of all ErbB receptors, is a mediator of lateral signaling," *Embo j,* vol. 16, pp. 1647-55, Apr 1 1997.

[42] Z. Penzvalto, B. Tegze, A. M. Szasz, Z. Sztupinszki, I. Liko, A. Szendroi, *et al.*, "Identifying resistance mechanisms against five tyrosine kinase inhibitors targeting the ERBB/RAS pathway in 45 cancer cell lines," *PLoS One,* vol. 8, p. e59503, 2013.

[43] O. Menyhart, L. Santarpia, and B. Gyorffy, "A Comprehensive Outline of Trastuzumab Resistance Biomarkers in HER2 Overexpressing Breast Cancer," *Curr Cancer Drug Targets,* vol. 15, pp. 665-83, 2015.

[44] C. Dang, N. Iyengar, F. Datko, G. D'Andrea, M. Theodoulou, M. Dickler, *et al.*, "Phase II study of paclitaxel given once per week along with trastuzumab and pertuzumab in patients with human epidermal growth factor receptor 2-positive metastatic breast cancer," *J Clin Oncol,* vol. 33, pp. 442-7, Feb 10 2015.

[45] J. Baselga , J. Cortés , S.-B. Kim , S.-A. Im , R. Hegg , Y.-H. Im , *et al.*, "Pertuzumab plus Trastuzumab plus Docetaxel for Metastatic Breast Cancer," *New England Journal of Medicine*, vol. 366, pp. 109-119, 2012.

[46] S. Dhillon, "Trastuzumab emtansine: a review of its use in patients with HER2-positive advanced breast cancer previously treated with trastuzumab-based therapy," *Drugs,* vol. 74, pp. 675-86, Apr 2014.

[47] C. E. Geyer , J. Forster , D. Lindquist , S. Chan , C. G. Romieu , T. Pienkowski , *et al.*, "Lapatinib plus Capecitabine for HER2-Positive Advanced Breast Cancer," *New England Journal of Medicine,* vol. 355, pp. 2733-2743, 2006.

[48] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, *et al.*, "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012," *Int J Cancer,* vol. 136, pp. E359-86, Mar 1 2015.

[49] R. J. Kurman and M. Shih Ie, "The origin and pathogenesis of epithelial ovarian cancer: a proposed unifying theory," *Am J Surg Pathol,* vol. 34, pp. 433-43, Mar 2010.

[50] Y. Lee, A. Miron, R. Drapkin, M. R. Nucci, F. Medeiros, A. Saleemuddin, *et al.*, "A candidate precursor to serous carcinoma that originates in the distal fallopian tube," *J Pathol,* vol. 211, pp. 26-35, Jan 2007.

[51] "Integrated genomic analyses of ovarian carcinoma," *Nature,* vol. 474, pp. 609-615, 06/30/print 2011.

[52] D. Lim and E. Oliva, "Precursors and pathogenesis of ovarian carcinoma," *Pathology,* vol. 45, pp. 229-42, Apr 2013.

[53] E. I. Braicu, J. Sehouli, R. Richter, K. Pietzner, C. Denkert, and C. Fotopoulou, "Role of histological type on surgical outcome and survival following radical primary tumour debulking of epithelial ovarian, fallopian tube and peritoneal cancers," *Br J Cancer,* vol. 105, pp. 1818-1824, 12/06/print 2011.

[54] A. du Bois, A. Reuss, E. Pujade-Lauraine, P. Harter, I. Ray-Coquard, and J. Pfisterer, "Role of surgical outcome as prognostic factor in advanced epithelial ovarian cancer: A combined exploratory analysis of 3 prospectively randomized phase 3 multicenter trials," *Cancer,* vol. 115, pp. 1234-1244, 2009.

[55] R. F. Ozols, B. N. Bundy, B. E. Greer, J. M. Fowler, D. Clarke-Pearson, R. A. Burger, *et al.*, "Phase III Trial of Carboplatin and Paclitaxel Compared With Cisplatin and Paclitaxel in Patients With Optimally Resected Stage III Ovarian Cancer: A Gynecologic Oncology Group Study," *Journal of Clinical Oncology,* vol. 21, pp. 3194-3200, September 1, 2003 2003.

[56] J. P. Neijt, S. A. Engelholm, M. K. Tuxen, P. G. Sørensen, M. Hansen, C. Sessa, *et al.*, "Exploratory Phase III Study of Paclitaxel and Cisplatin Versus Paclitaxel and Carboplatin in Advanced Ovarian Cancer," *Journal of Clinical Oncology,* vol. 18, pp. 3084-3092, September 17, 2000 2000.

[57] S. Pignata, G. Scambia, G. Ferrandina, A. Savarese, R. Sorio, E. Breda, *et al.*, "Carboplatin Plus Paclitaxel Versus Carboplatin Plus Pegylated Liposomal Doxorubicin As First-Line Treatment for Patients With Ovarian Cancer: The MITO-2 Randomized Phase III Trial," *Journal of Clinical Oncology,* vol. 29, pp. 3628-3635, September 20, 2011 2011.

[58] P. A. Vasey, G. C. Jayson, A. Gordon, H. Gabra, R. Coleman, R. Atkinson, *et al.*, "Phase III Randomized Trial of Docetaxel–Carboplatin Versus Paclitaxel–Carboplatin as First-line Chemotherapy for Ovarian Carcinoma," *Journal of the National Cancer Institute,* vol. 96, pp. 1682-1691, November 17, 2004 2004.

[59] S. Vaughan, J. I. Coward, R. C. Bast, Jr., A. Berchuck, J. S. Berek, J. D. Brenton, *et al.*, "Rethinking ovarian cancer: recommendations for improving outcomes," *Nat Rev Cancer,* vol. 11, pp. 719-25, Oct 2011.

[60] D. D. Bowtell, S. Bohm, A. A. Ahmed, P.-J. Aspuria, R. C. Bast Jr, V. Beral, *et al.*, "Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer," *Nat Rev Cancer,* vol. 15, pp. 668-679, 11//print 2015.

[61] M. Shimada, J. Kigawa, Y. Ohishi, M. Yasuda, M. Suzuki, M. Hiura, *et al.*, "Clinicopathological characteristics of mucinous adenocarcinoma of the ovary," *Gynecol Oncol,* vol. 113, pp. 331-4, Jun 2009.

[62] S. C. Mok, D. A. Bell, R. C. Knapp, P. M. Fishbaugh, W. R. Welch, M. G. Muto, *et al.*, "Mutation of K-ras protooncogene in human ovarian epithelial tumors of borderline malignancy," *Cancer Res,* vol. 53, pp. 1489-92, Apr 1 1993.

[63] J. Brown and M. Frumovitz, "Mucinous tumors of the ovary: current thoughts on diagnosis and management," *Curr Oncol Rep,* vol. 16, p. 389, Jun 2014.

[64] S. Sato, H. Itamochi, J. Kigawa, T. Oishi, M. Shimada, S. Sato, *et al.*, "Combination chemotherapy of oxaliplatin and 5-fluorouracil may be an effective regimen for mucinous adenocarcinoma of the ovary: a potential treatment strategy," *Cancer Sci,* vol. 100, pp. 546-51, Mar 2009.

[65] M. Takano, Y. Kikuchi, N. Yaegashi, K. Kuzuya, M. Ueki, H. Tsuda, *et al.*, "Clear cell carcinoma of the ovary: a retrospective multicentre experience of 254 patients with complete surgical staging," *Br J Cancer,* vol. 94, pp. 1369-74, May 22 2006.

[66] K.-T. Kuo, T.-L. Mao, S. Jones, E. Veras, A. Ayhan, T.-L. Wang, *et al.*, "Frequent Activating Mutations of PIK3CA in Ovarian Clear Cell Carcinoma," *The American Journal of Pathology,* vol. 174, pp. 1597-1601, 01/22/accepted 2009.

[67] S. Jones, T. L. Wang, M. Shih Ie, T. L. Mao, K. Nakayama, R. Roden, *et al.*, "Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma," *Science,* vol. 330, pp. 228-31, Oct 8 2010.

[68] M. S. Anglesio, J. George, H. Kulbe, M. Friedlander, D. Rischin, C. Lemech, *et al.*, "IL6-STAT3-HIF signaling and therapeutic response to the angiogenesis inhibitor sunitinib in ovarian clear cell cancer," *Clin Cancer Res,* vol. 17, pp. 2538-48, Apr 15 2011.

[69] G. E. Konecny, C. Wang, H. Hamidi, B. Winterhoff, K. R. Kalli, J. Dering, *et al.*, "Prognostic and Therapeutic Relevance of Molecular Subtypes in High-Grade Serous Ovarian Cancer," *Journal of the National Cancer Institute,* vol. 106, October 1, 2014 2014.

[70] A. A. Ahmed, D. Etemadmoghadam, J. Temple, A. G. Lynch, M. Riad, R. Sharma, *et al.*, "Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary," *J Pathol,* vol. 221, pp. 49-56, May 2010.

[71] A. M. Patch, E. L. Christie, D. Etemadmoghadam, D. W. Garsed, J. George, S. Fereday, *et al.*, "Whole-genome characterization of chemoresistant ovarian cancer," *Nature,* vol. 521, pp. 489-94, May 28 2015.
[72] S. B. Kaye, "Progress in the treatment of ovarian cancer-lessons from homologous recombination deficiency-the first 10 years," *Ann Oncol,* vol. 27 Suppl 1, pp. i1-i3, Apr 2016.
[73] B. Gyorffy and R. Schafer, "Biomarkers downstream of RAS: a search for robust transcriptional targets," *Curr Cancer Drug Targets,* vol. 10, pp. 858-68, Dec 2010.
[74] J. A. Sosman , K. B. Kim , L. Schuchter , R. Gonzalez , A. C. Pavlick , J. S. Weber , *et al.*, "Survival in BRAF V600–Mutant Advanced Melanoma Treated with Vemurafenib," *New England Journal of Medicine,* vol. 366, pp. 707-714, 2012.
[75] A. T. Shaw, D.-W. Kim, K. Nakagawa, T. Seto, L. Crinó, M.-J. Ahn, *et al.*, "Crizotinib versus Chemotherapy in Advanced ALK-Positive Lung Cancer," *New England Journal of Medicine,* vol. 368, pp. 2385-2394, 2013.
[76] D. J. Slamon, B. Leyland-Jones, S. Shak, H. Fuchs, V. Paton, A. Bajamonde, *et al.*, "Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2," *N Engl J Med,* vol. 344, pp. 783-92, Mar 15 2001.
[77] M. Marty, F. Cognetti, D. Maraninchi, R. Snyder, L. Mauriac, M. Tubiana-Hulin, *et al.*, "Randomized phase II trial of the efficacy and safety of trastuzumab combined with docetaxel in patients with human epidermal growth factor receptor 2-positive metastatic breast cancer administered as first-line treatment: the M77001 study group," *J Clin Oncol,* vol. 23, pp. 4265-74, Jul 1 2005.
[78] A. Brufsky, "Trastuzumab-based therapy for patients with HER2-positive breast cancer: from early scientific development to foundation of care," *Am J Clin Oncol,* vol. 33, pp. 186-95, Apr 2010.
[79] P. Sharma and James P. Allison, "Immune Checkpoint Targeting in Cancer Therapy: Toward Combination Strategies with Curative Potential," *Cell,* vol. 161, pp. 205-214.
[80] G. Maravić, J. M. Bujnicki, M. Feder, S. Pongor, and M. Flögel, "Alanine-scanning mutagenesis of the predicted rRNA-binding domain of ErmC′ redefines the substrate-binding site and suggests a model for protein–RNA interactions," *Nucleic acids research,* vol. 31, pp. 4941-4949, 2003.
[81] A. Escobar-Zepeda, A. V.-P. de León, and A. Sanchez-Flores, "The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics," *Frontiers in genetics,* vol. 6, 2015.
[82] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, *et al.*, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res,* vol. 25, pp. 3389-402, Sep 1 1997.
[83] D. H. Huson, D. C. Richter, S. Mitra, A. F. Auch, and S. C. Schuster, "Methods for comparative metagenomics," *BMC Bioinformatics,* vol. 10 Suppl 1, p. S12, 2009.
[84] J. Droge and A. C. McHardy, "Taxonomic binning of metagenome samples generated by next-generation sequencing technologies," *Brief Bioinform,* vol. 13, pp. 646-55, Nov 2012.
[85] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "MEGAN analysis of metagenomic data," *Genome Res,* vol. 17, pp. 377-86, Mar 2007.
[86] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, *et al.*, "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Appl Environ Microbiol,* vol. 75, pp. 7537-41, Dec 2009.
[87] M. Monzoorul Haque, T. S. Ghosh, D. Komanduri, and S. S. Mande, "SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences," *Bioinformatics,* vol. 25, pp. 1722-30, Jul 15 2009.
[88] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower, "Metagenomic microbial community profiling using unique clade-specific marker genes," *Nat Methods,* vol. 9, pp. 811-4, Aug 2012.
[89] W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, *et al.*, "Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data," *Nucleic acids research,* vol. 43, pp. D1071-D1078, 2015.
[90] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic acids research,* vol. 32, pp. D267-D270, 2004.
[91] S. Burge, T. K. Attwood, A. Bateman, T. Z. Berardini, M. Cherry, C. O'Donovan, *et al.*, "Biocurators and biocuration: surveying the 21st century challenges," *Database,* vol. 2012, p. bar059, 2012.
[92] E. Krissinel and K. Henrick, "Inference of macromolecular assemblies from crystalline state," *Journal of molecular biology,* vol. 372, pp. 774-797, 2007.
[93] J. Janin and C. Chothia, "The structure of protein-protein recognition sites," *J. biol. Chem,* vol. 265, 1990.
[94] O. Vinogradova and J. Qin, "NMR as a unique tool in assessment and complex determination of weak protein–protein interactions," in *NMR of Proteins and Small Biomolecules*, ed: Springer, 2011, pp. 35-45.
[95] A. J. Wand and S. W. Englander, "Protein complexes studied by NMR spectroscopy," *Current opinion in biotechnology,* vol. 7, pp. 403-408, 1996.
[96] E. M. Phizicky and S. Fields, "Protein-protein interactions: methods for detection and analysis," *Microbiological reviews,* vol. 59, pp. 94-123, 1995.

[97] S. Lu, P. Deng, X. Liu, J. Luo, R. Han, X. Gu, *et al.*, "Solution structure of the major α-amylase inhibitor of the crop plant amaranth," *Journal of Biological Chemistry,* vol. 274, pp. 20473-20478, 1999.
[98] S. Fields and O.-k. Song, "A novel genetic system to detect protein protein interactions," 1989.
[99] Y. Xia, H. Yu, R. Jansen, M. Seringhaus, S. Baxter, D. Greenbaum, *et al.*, "Analyzing cellular biochemistry in terms of molecular networks," *Annual review of biochemistry,* vol. 73, pp. 1051-1087, 2004.
[100] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, *et al.*, "DrugBank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic acids research,* vol. 34, pp. D668-D672, 2006.
[101] U. Consortium, "UniProt: a hub for protein information," *Nucleic acids research,* p. gku989, 2014.
[102] A. Yates, W. Akanni, M. R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, *et al.*, "Ensembl 2016," *Nucleic acids research,* vol. 44, pp. D710-D716, 2016.
[103] D. R. Maglott, K. S. Katz, H. Sicotte, and K. D. Pruitt, "NCBI's LocusLink and RefSeq," *Nucleic acids research,* vol. 28, pp. 126-128, 2000.
[104] T. Tatusova, S. Ciufo, B. Fedorov, K. O'Neill, and I. Tolstoy, "RefSeq microbial genomes database: new representation and annotation strategy," *Nucleic acids research,* p. gkt1274, 2013.
[105] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, *et al.*, "GenBank," *Nucleic acids research,* p. gks1195, 2012.
[106] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, *et al.*, "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs," *Nucleic Acids Res,* vol. 39, pp. D1035-41, Jan 2011.
[107] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, *et al.*, "DrugBank 4.0: shedding new light on drug metabolism," *Nucleic acids research,* vol. 42, pp. D1091-D1097, 2014.
[108] X. Chen, Z. L. Ji, and Y. Z. Chen, "TTD: therapeutic target database," *Nucleic acids research,* vol. 30, pp. 412-415, 2002.
[109] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "STITCH: interaction networks of chemicals and proteins," *Nucleic acids research,* vol. 36, pp. D684-D688, 2008.
[110] M. Kuhn, D. Szklarczyk, A. Franceschini, C. von Mering, L. J. Jensen, and P. Bork, "STITCH 3: zooming in on protein-chemical interactions," *Nucleic Acids Res,* vol. 40, pp. D876-80, Jan 2012.
[111] M. Kuhn, D. Szklarczyk, S. Pletscher-Frankild, T. H. Blicher, C. von Mering, L. J. Jensen, *et al.*, "STITCH 4: integration of protein–chemical interactions with user data," *Nucleic acids research,* p. gkt1207, 2013.
[112] Y. Liu, B. Hu, C. Fu, and X. Chen, "DCDB: drug combination database," *Bioinformatics,* vol. 26, pp. 587-588, 2010.
[113] Y. Liu, Q. Wei, G. Yu, W. Gai, Y. Li, and X. Chen, "DCDB 2.0: a major update of the drug combination database," *Database,* vol. 2014, p. bau124, 2014.
[114] F. Home, "Orange Book: approved drug products with therapeutic equivalence evaluations," 2013.
[115] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs," *Molecular systems biology,* vol. 6, 2010.
[116] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The SIDER database of drugs and side effects," *Nucleic acids research,* p. gkv1075, 2015.
[117] N. P. Tatonetti, P. Y. Patrick, R. Daneshjou, and R. B. Altman, "Data-driven prediction of drug effects and interactions," *Science translational medicine,* vol. 4, pp. 125ra31-125ra31, 2012.
[118] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, "DIP: the database of interacting proteins," *Nucleic acids research,* vol. 28, pp. 289-291, 2000.
[119] G. D. Bader and C. W. Hogue, "BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways," *Bioinformatics,* vol. 16, pp. 465-477, 2000.
[120] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni, "MINT: a Molecular INTeraction database," *FEBS letters,* vol. 513, pp. 135-140, 2002.
[121] A. Chatr-Aryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli, *et al.*, "MINT: the Molecular INTeraction database," *Nucleic acids research,* vol. 35, pp. D572-D574, 2007.
[122] A. Ceol, A. C. Aryamontri, L. Licata, D. Peluso, L. Briganti, L. Perfetto, *et al.*, "MINT, the molecular interaction database: 2009 update," *Nucleic acids research,* p. gkp983, 2009.
[123] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, *et al.*, "MINT, the molecular interaction database: 2012 update," *Nucleic acids research,* vol. 40, pp. D857-D861, 2012.
[124] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, *et al.*, "IntAct: an open source molecular interaction database," *Nucleic acids research,* vol. 32, pp. D452-D455, 2004.
[125] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The database of interacting proteins: 2004 update," *Nucleic acids research,* vol. 32, pp. D449-D451, 2004.
[126] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic acids research,* vol. 30, pp. 303-305, 2002.
[127] L. Salwinski and D. Eisenberg, "The MiSink Plugin: Cytoscape as a graphical interface to the Database of Interacting Proteins," *Bioinformatics,* vol. 23, pp. 2193-2195, 2007.

[128] S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, *et al.*, "Development of human protein reference database as an initial platform for approaching systems biology in humans," *Genome research*, vol. 13, pp. 2363-2371, 2003.

[129] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, *et al.*, "IntAct—open source resource for molecular interaction data," *Nucleic acids research*, vol. 35, pp. D561-D565, 2007.

[130] B. Aranda, P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, *et al.*, "The IntAct molecular interaction database in 2010," *Nucleic acids research*, vol. 38, pp. D525-D531, 2010.

[131] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, *et al.*, "The IntAct molecular interaction database in 2012," *Nucleic acids research*, p. gkr1088, 2011.

[132] R. Côté, F. Reisinger, L. Martens, H. Barsnes, J. A. Vizcaino, and H. Hermjakob, "The ontology lookup service: bigger and better," *Nucleic acids research*, vol. 38, pp. W155-W160, 2010.

[133] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, *et al.*, "The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data," *Nature biotechnology*, vol. 22, pp. 177-183, 2004.

[134] S. Kerrien, S. Orchard, L. Montecchi-Palazzi, B. Aranda, A. F. Quinn, N. Vinod, *et al.*, "Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions," *BMC biology*, vol. 5, p. 44, 2007.

[135] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, *et al.*, "Gene Ontology: tool for the unification of biology," *Nature genetics*, vol. 25, pp. 25-29, 2000.

[136] N. R. Coordinators, "Database resources of the national center for biotechnology information," *Nucleic acids research*, vol. 43, p. D6, 2015.

[137] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, *et al.*, "The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases," *Nucleic acids research*, p. gkt1115, 2013.

[138] S. Orchard, S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, *et al.*, "Protein interaction data curation: the International Molecular Exchange (IMEx) consortium," *Nature methods*, vol. 9, pp. 345-350, 2012.

[139] A. Calderone, L. Castagnoli, and G. Cesareni, "Mentha: a resource for browsing integrated protein-interaction networks," *Nature methods*, vol. 10, pp. 690-691, 2013.

[140] R. Knüppel, P. Dietze, W. Lehnberg, K. Frech, and E. Wingender, "TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins," *Journal of Computational Biology*, vol. 1, pp. 191-198, 1994.

[141] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature genetics*, vol. 45, pp. 1113-1120, 2013.

[142] E. D. Pleasance, R. K. Cheetham, P. J. Stephens, D. J. McBride, S. J. Humphray, C. D. Greenman, *et al.*, "A comprehensive catalogue of somatic mutations from a human cancer genome," *Nature*, vol. 463, pp. 191-196, 2010.

[143] M. J. Ellis, M. Gillette, S. A. Carr, A. G. Paulovich, R. D. Smith, K. K. Rodland, *et al.*, "Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium," *Cancer discovery*, vol. 3, pp. 1108-1112, 2013.

[144] B. Zhang, J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, *et al.*, "Proteogenomic characterization of human colon and rectal cancer," *Nature*, vol. 513, pp. 382-387, 2014.

[145] T. J. Hudson, W. Anderson, A. Aretz, A. D. Barker, C. Bell, R. R. Bernabé, *et al.*, "International network of cancer genome projects," *Nature*, vol. 464, pp. 993-998, 2010.

[146] R. Shepherd, S. A. Forbes, D. Beare, S. Bamford, C. G. Cole, S. Ward, *et al.*, "Data mining using the catalogue of somatic mutations in cancer BioMart," *Database*, vol. 2011, p. bar018, 2011.

[147] C. Perez-Llamas, G. Gundem, and N. Lopez-Bigas, "Integrative cancer genomics (IntOGen) in Biomart," *Database*, vol. 2011, p. bar039, 2011.

[148] M. Goldman, B. Craft, T. Swatloski, K. Ellrott, M. Cline, M. Diekhans, *et al.*, "The UCSC cancer genomics browser: update 2013," *Nucleic acids research*, vol. 41, pp. D949-D954, 2013.

[149] J. Zhu, J. Z. Sanborn, S. Benz, C. Szeto, F. Hsu, R. M. Kuhn, *et al.*, "The UCSC cancer genomics browser," *Nature methods*, vol. 6, pp. 239-240, 2009.

[150] O. An, V. Pendino, M. D'Antonio, E. Ratti, M. Gentilini, and F. D. Ciccarelli, "NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes," *Database*, vol. 2014, p. bau015, 2014.

[151] M. D'Antonio, V. Pendino, S. Sinha, and F. D. Ciccarelli, "Network of Cancer Genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes," *Nucleic acids research*, vol. 40, pp. D978-D983, 2012.

[152] J. Zhang, R. P. Finney, W. Rowe, M. Edmonson, S. H. Yang, T. Dracheva, *et al.*, "Systematic analysis of genetic alterations in tumors using Cancer Genome WorkBench (CGWB)," *Genome research*, vol. 17, pp. 1111-1117, 2007.

[153] J. Feichtinger, R. J. McFarlane, and L. D. Larcombe, "CancerMA: a web-based tool for automatic meta-analysis of public cancer microarray data," *Database*, vol. 2012, p. bas055, 2012.

[154] S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, *et al.*, "COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer," *Nucleic acids research*, p. gkq929, 2010.
[155] Q. Cao, M. Zhou, X. Wang, C. A. Meyer, Y. Zhang, Z. Chen, *et al.*, "CaSNP: a database for interrogating copy number alterations of cancer genome from SNP array data," *Nucleic acids research*, p. gkq997, 2010.
[156] W.-C. Cheng, I.-F. Chung, C.-Y. Chen, H.-J. Sun, J.-J. Fen, W.-C. Tang, *et al.*, "DriverDB: an exome sequencing database for cancer driver gene identification," *Nucleic acids research*, vol. 42, pp. D1048-D1054, 2014.
[157] G. Gundem, C. Perez-Llamas, A. Jene-Sanz, A. Kedzierska, A. Islam, J. Deu-Pons, *et al.*, "IntOGen: integration and data mining of multidimensional oncogenomic data," *Nature methods*, vol. 7, pp. 92-93, 2010.
[158] C. J. Richardson, Q. Gao, C. Mitsopoulous, M. Zvelebil, L. H. Pearl, and F. M. Pearl, "MoKCa database—mutations of kinases in cancer," *Nucleic acids research*, vol. 37, pp. D824-D831, 2009.
[159] J. L. Hess, "The Cancer Genome Anatomy Project: Power Tools for Cancer Biologists: EDITORIAL," *Cancer investigation*, vol. 21, pp. 325-326, 2003.
[160] F. Mitelman, B. Johansson, and F. Mertens, "The impact of translocations and gene fusions on cancer causation," *Nature Reviews Cancer*, vol. 7, pp. 233-245, 2007.
[161] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, *et al.*, "A census of human cancer genes," *Nature Reviews Cancer*, vol. 4, pp. 177-183, 2004.
[162] J. Li, D. T. Duncan, and B. Zhang, "CanProVar: a human cancer proteome variation database," *Human mutation*, vol. 31, pp. 219-228, 2010.
[163] X. He, S. Chang, J. Zhang, Q. Zhao, H. Xiang, K. Kusonmano, *et al.*, "MethyCancer: the database of human DNA methylation and cancer," *Nucleic acids research*, vol. 36, pp. D836-D841, 2008.
[164] M. Krupp, T. Itzel, T. Maass, A. Hildebrandt, P. R. Galle, and A. Teufel, "CellLineNavigator: a workbench for cancer cell line analysis," *Nucleic acids research*, vol. 41, pp. D942-D948, 2013.
[165] M. Ongenaert, L. Van Neste, T. De Meyer, G. Menschaert, S. Bekaert, and W. Van Criekinge, "PubMeth: a cancer methylation database combining text-mining and expert annotation," *Nucleic acids research*, vol. 36, pp. D842-D846, 2008.
[166] D. Wang, J. Gu, T. Wang, and Z. Ding, "OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs," *Bioinformatics*, vol. 30, pp. 2237-2238, 2014.
[167] R. Kumar, K. Chaudhary, S. Gupta, H. Singh, S. Kumar, A. Gautam, *et al.*, "CancerDR: cancer drug resistance database," *Scientific reports*, vol. 3, 2013.
[168] !!! INVALID CITATION !!! [154, 155].
[169] J. Ahmed, T. Meinel, M. Dunkel, M. S. Murgueitio, R. Adams, C. Blasse, *et al.*, "CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge," *Nucleic acids research*, vol. 39, pp. D960-D967, 2011.
[170] I. Scheinin, S. Myllykangas, I. Borze, T. Böhling, S. Knuutila, and J. Saharinen, "CanGEM: mining gene copy number changes in cancer," *Nucleic acids research*, vol. 36, pp. D830-D835, 2008.
[171] J. N. Weinstein, K. W. Kohn, M. R. Grever, V. N. Viswanadhan, L. V. Rubinstein, A. P. Monks, *et al.*, "Neural computing in cancer drug development: predicting mechanism of action," *Science*, vol. 258, pp. 447-451, 1992.
[172] A. Monks, D. Scudiero, P. Skehan, R. Shoemaker, K. Paull, D. Vistica, *et al.*, "Feasibility of a high-flux anticancer drug screen using a diverse panel of cultured human tumor cell lines," *Journal of the National Cancer Institute*, vol. 83, pp. 757-766, 1991.
[173] J. Ramana, "RCDB: renal cancer gene database," *BMC research notes*, vol. 5, p. 246, 2012.
[174] B. F. Ganzfried, M. Riester, B. Haibe-Kains, T. Risch, S. Tyekucheva, I. Jazic, *et al.*, "curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome," *Database*, vol. 2013, p. bat013, 2013.
[175] R. J. Cutts, E. Gadaleta, N. R. Lemoine, and C. Chelala, "Using BioMart as a framework to manage and query pancreatic cancer data," *Database*, vol. 2011, p. bar024, 2011.
[176] L. Wang, Y. Xiong, Y. Sun, Z. Fang, L. Li, H. Ji, *et al.*, "HLungDB: an integrated database of human lung cancer research," *Nucleic acids research*, p. gkp945, 2009.
[177] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, pp. 27-30, 2000.
[178] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, *et al.*, "Reactome: a database of reactions, pathways and biological processes," *Nucleic acids research*, p. gkq1018, 2010.
[179] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, *et al.*, "HMDB: the human metabolome database," *Nucleic acids research*, vol. 35, pp. D521-D526, 2007.
[180] T. Korcsmáros, I. J. Farkas, M. S. Szalay, P. Rovó, D. Fazekas, Z. Spiró, *et al.*, "Uniformly curated signaling pathways reveal tissue-specific cross-talks and support drug target discovery," *Bioinformatics*, vol. 26, pp. 2042-2050, 2010.
[181] D. Fazekas, M. Koltai, D. Türei, D. Módos, M. Pálfy, Z. Dúl, *et al.*, "SignaLink 2–a signaling pathway resource with multi-layered regulatory networks," *BMC systems biology*, vol. 7, p. 1, 2013.
[182] D. Pratt, J. Chen, D. Welker, R. Rivas, R. Pillich, V. Rynkov, *et al.*, "NDEx, the Network Data Exchange," *Cell systems*, vol. 1, pp. 302-305, 2015.

[183] D. L. Wheeler, D. M. Church, A. E. Lash, D. D. Leipe, T. L. Madden, J. U. Pontius, *et al.*, "Database resources of the national center for biotechnology information," *Nucleic acids research,* vol. 29, pp. 11-16, 2001.
[184] A. Kastrin, T. C. Rindflesch, and D. Hristovski, "Large-scale structure of a network of co-occurring MeSH terms: statistical analysis of macroscopic properties," *PloS one,* vol. 9, p. e102188, 2014.
[185] K. D. Pruitt, G. R. Brown, S. M. Hiatt, F. Thibaud-Nissen, A. Astashyn, O. Ermolaeva, *et al.*, "RefSeq: an update on mammalian reference sequences," *Nucleic acids research,* vol. 42, pp. D756-D763, 2014.
[186] Y. Liu, B. Hu, C. Fu, and X. Chen, "DCDB: drug combination database," *Bioinformatics,* vol. 26, pp. 587-8, Feb 15 2010.
[187] C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel, "STRING: a database of predicted functional associations between proteins," *Nucleic Acids Research,* vol. 31, pp. 258-261, 2003.
[188] A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard, "JASPAR: an open-access database for eukaryotic transcription factor binding profiles," *Nucleic acids research,* vol. 32, pp. D91-D94, 2004.
[189] S. K. Kummerfeld and S. A. Teichmann, "DBD: a transcription factor prediction database," *Nucleic acids research,* vol. 34, pp. D74-D81, 2006.
[190] R. Caspi, H. Foerster, C. A. Fulcher, R. Hopkinson, J. Ingraham, P. Kaipa, *et al.*, "MetaCyc: a multiorganism database of metabolic pathways and enzymes," *Nucleic acids research,* vol. 34, pp. D511-D516, 2006.
[191] K. Kandasamy, S. S. Mohan, R. Raju, S. Keerthikumar, G. S. S. Kumar, A. K. Venugopal, *et al.*, "NetPath: a public resource of curated signal transduction pathways," *Genome biology,* vol. 11, pp. 1-9, 2010.
[192] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, *et al.*, "The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website," *British journal of cancer,* vol. 91, pp. 355-358, 2004.
[193] A. Hamosh, A. F. Scott, J. Amberger, D. Valle, and V. A. McKusick, "Online Mendelian inheritance in man (OMIM)," *Human mutation,* vol. 15, pp. 57-61, 2000.
[194] U.S. National Library of Medicine. (2001). *PubMed Overview*. Available: http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/index.html
[195] R. Vera, Y. Perez-Riverol, S. Perez, B. Ligeti, A. Kertész-Farkas, and S. Pongor, "JBioWH: an open-source Java framework for bioinformatics data integration," *Database,* vol. 2013, p. bat051, 2013.
[196] J. A. Bondy and U. S. R. Murty, *Graph theory with applications* vol. 290: Macmillan London, 1976.
[197] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: bringing order to the web," 1999.
[198] T. H. Haveliwala, "Topic-sensitive pagerank," in *Proceedings of the 11th international conference on World Wide Web*, 2002, pp. 517-526.
[199] G. Jeh and J. Widom, "Scaling personalized web search," in *Proceedings of the 12th international conference on World Wide Web*, 2003, pp. 271-279.
[200] T. Ito, M. Shimbo, T. Kudo, and Y. Matsumoto, "Application of kernels to link analysis," *Proceedings of the eleventh ...,* pp. 586-592, 2005.
[201] J. Kandola, J. Shawe-Taylor, and N. Cristianini, "On the application of diffusion kernel to text data," Technical report, Neurocolt, 2002. NeuroCOLT Technical Report NC-TR-02-1222002.
[202] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*: Cambridge university press, 2004.
[203] R. I. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete input spaces," in *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*, San Francisco, CA, USA, 2002, pp. 315-322.
[204] S. White and P. Smyth, "Algorithms for estimating relative importance in networks," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 266-275.
[205] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM),* vol. 46, pp. 604-632, 1999.
[206] R. I. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete structures," in *Proceedings of the 19th international conference on machine learning*, 2002, pp. 315-322.
[207] M. Eiermann and O. G. Ernst, "A restarted Krylov subspace method for the evaluation of matrix functions," *SIAM Journal on Numerical Analysis,* vol. 44, pp. 2481-2504, 2006.
[208] C. Moler and C. Van Loan, "Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later," *SIAM review,* vol. 45, pp. 3-49, 2003.
[209] Y. Saad, "Analysis of some Krylov subspace approximations to the matrix exponential operator," *SIAM Journal on Numerical Analysis,* vol. 29, pp. 209-228, 1992.
[210] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *Knowledge and data engineering, ieee transactions on,* vol. 19, pp. 355-369, 2007.
[211] J. Stoer and R. Bulirsch, *Introduction to numerical analysis* vol. 12: Springer Science & Business Media, 2013.
[212] M. Newman, *Networks: an introduction*: OUP Oxford, 2010.
[213] G. H. Golub and C. Greif, "An Arnoldi-type algorithm for computing page rank," *BIT Numerical Mathematics,* vol. 46, pp. 759-771, 2006.

[214] E. Cuthill and J. McKee, "Reducing the bandwidth of sparse symmetric matrices," in *Proceedings of the 1969 24th national conference*, 1969, pp. 157-172.
[215] Y. Saad and M. H. Schultz, "GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems," *SIAM Journal on scientific and statistical computing*, vol. 7, pp. 856-869, 1986.
[216] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems* vol. 49: NBS, 1952.
[217] T. K. Sarkar, "On the application of the generalized biconjugate gradient method," *Journal of Electromagnetic Waves and Applications*, vol. 1, pp. 223-242, 1987.
[218] R. Busa-Fekete, A. Kertész-Farkas, A. Kocsor, and S. Pongor, "Balanced ROC analysis (BAROC) protocol for the evaluation of protein similarities," *Journal of biochemical and biophysical methods*, vol. 70, pp. 1210-1214, 2008.
[219] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," presented at the Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1, Montreal, Quebec, Canada, 1995.
[220] C. Pesquita, D. Faria, A. O. Falcao, P. Lord, and F. M. Couto, "Semantic similarity in biomedical ontologies," *PLoS Comput Biol*, vol. 5, p. e1000443, Jul 2009.
[221] A. Gottlieb, G. Y. Stein, Y. Oron, E. Ruppin, and R. Sharan, "INDI: a computational framework for inferring drug interactions and their associated recommendations," *Mol Syst Biol*, vol. 8, p. 592, 2012.
[222] S. D, F. A, and K. M, "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research (Database issue)*, 2011.
[223] F. Zhu, Z. Shi, C. Qin, L. Tao, X. Liu, F. Xu, *et al.*, "Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery," *Nucleic Acids Res*, vol. 40, pp. D1128-36, Jan 2012.
[224] P. Therasse, S. G. Arbuck, E. A. Eisenhauer, J. Wanders, R. S. Kaplan, L. Rubinstein, *et al.*, "New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada," *J Natl Cancer Inst*, vol. 92, pp. 205-16, Feb 2 2000.
[225] E. Sayers and D. Wheeler. (2004). *Building Customized Data Pipelines Using the Entrez Programming Utilities (eUtils)*. Available: http://www.ncbi.nlm.nih.gov/books/NBK1056/
[226] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, *et al.*, "Biopython: freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, pp. 1422-1423, 2009.
[227] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery," *Nat Chem Biol*, vol. 4, pp. 682-90, Nov 2008.
[228] P. Imming, C. Sinning, and A. Meyer, "Drugs, their targets and the nature and number of drug targets," *Nat Rev Drug Discov*, vol. 5, pp. 821-34, Oct 2006.
[229] H. Kitano, "A robustness-based approach to systems-oriented drug design," *Nat Rev Drug Discov*, vol. 6, pp. 202-10, Mar 2007.
[230] V. Agoston, P. Csermely, and S. Pongor, "Multiple weak hits confuse complex systems: a transcriptional regulatory network as an example," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 71, p. 051909, May 2005.
[231] P. Csermely, V. Agoston, and S. Pongor, "The efficiency of multi-target drugs: the network approach might help drug design," *Trends Pharmacol Sci*, vol. 26, pp. 178-82, Apr 2005.
[232] J. Lehar, A. S. Krueger, W. Avery, A. M. Heilbut, L. M. Johansen, E. R. Price, *et al.*, "Synergistic drug combinations tend to improve therapeutically relevant selectivity," *Nat Biotechnol*, vol. 27, pp. 659-66, Jul 2009.
[233] C. T. Keith, A. A. Borisy, and B. R. Stockwell, "Multicomponent therapeutics for networked systems," *Nat Rev Drug Discov*, vol. 4, pp. 71-8, Jan 2005.
[234] G. R. Zimmermann, J. Lehar, and C. T. Keith, "Multi-target therapeutics: when the whole is greater than the sum of the parts," *Drug Discov Today*, vol. 12, pp. 34-42, Jan 2007.
[235] E. A. Perez, E. H. Romond, V. J. Suman, J. H. Jeong, N. E. Davidson, C. E. Geyer, Jr., *et al.*, "Four-year follow-up of trastuzumab plus adjuvant chemotherapy for operable human epidermal growth factor receptor 2-positive breast cancer: joint analysis of data from NCCTG N9831 and NSABP B-31," *J Clin Oncol*, vol. 29, pp. 3366-73, Sep 1 2011.
[236] I. Smith, M. Procter, R. D. Gelber, S. Guillaume, A. Feyereislova, M. Dowsett, *et al.*, "2-year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: a randomised controlled trial," *Lancet*, vol. 369, pp. 29-36, Jan 6 2007.
[237] A. A. Borisy, P. J. Elliott, N. W. Hurst, M. S. Lee, J. Lehar, E. R. Price, *et al.*, "Systematic discovery of multicomponent therapeutics," *Proc Natl Acad Sci U S A*, vol. 100, pp. 7977-82, Jun 24 2003.
[238] C. I. Bliss, "The Toxicity of Poisons Applied Jointly1," *Annals of Applied Biology*, vol. 26, pp. 585-615, 1939.
[239] S. Loewe and H. Muischnek, "Effect of combinations: mathematical basis of problem," *Arch Exp Pathol Pharmakol*, vol. 114, pp. 313-326, 1926.
[240] W. R. Greco, G. Bravo, and J. C. Parsons, "The search for synergy: a critical review from a response surface perspective," *Pharmacol Rev*, vol. 47, pp. 331-85, Jun 1995.

[241] P. K. Wong, F. Yu, A. Shahangian, G. Cheng, R. Sun, and C. M. Ho, "Closed-loop control of cellular functions using combinatory drugs guided by a stochastic search algorithm," *Proc Natl Acad Sci U S A*, vol. 105, pp. 5105-10, Apr 1 2008.

[242] D. Calzolari, S. Bruschi, L. Coquin, J. Schofield, J. D. Feala, J. C. Reed, *et al.*, "Search algorithms as a framework for the optimization of drug combinations," *PLoS Comput Biol*, vol. 4, p. e1000249, Dec 2008.

[243] K. Yang, H. Bai, Q. Ouyang, L. Lai, and C. Tang, "Finding multiple target optimal intervention in disease-related molecular network," *Mol Syst Biol*, vol. 4, p. 228, 2008.

[244] G. Jin, H. Zhao, X. Zhou, and S. T. Wong, "An enhanced Petri-net model to predict synergistic effects of pairwise drug combinations from gene microarray data," *Bioinformatics*, vol. 27, pp. i310-6, Jul 1 2011.

[245] Z. Wu, X. M. Zhao, and L. Chen, "A systems biology approach to identify effective cocktail drugs," *BMC Syst Biol*, vol. 4 Suppl 2, p. S7, 2010.

[246] X. M. Zhao, M. Iskar, G. Zeller, M. Kuhn, V. van Noort, and P. Bork, "Prediction of drug combinations by integrating molecular and pharmacological data," *PLoS Comput Biol*, vol. 7, p. e1002323, Dec 2011.

[247] S. Li, B. Zhang, and N. Zhang, "Network target for screening synergistic drug combinations with application to traditional Chinese medicine," *BMC Syst Biol*, vol. 5 Suppl 1, p. S10, 2011.

[248] M. Cokol, H. N. Chua, M. Tasan, B. Mutlu, Z. B. Weinstein, Y. Suzuki, *et al.*, "Systematic exploration of synergistic drug pairs," *Mol Syst Biol*, vol. 7, p. 544, 2011.

[249] J. Xiong, J. Liu, S. Rayner, Z. Tian, Y. Li, and S. Chen, "Pre-clinical drug prioritization via prognosis-guided genetic interaction networks," *PLoS One*, vol. 5, p. e13937, 2010.

[250] B. G. M. S. B. T. A. J. Katzung, *Basic & clinical pharmacology*. New York; London: McGraw-Hill Medical ; McGraw-Hill [distributor], 2012.

[251] K. J. Xu, J. Song, and X. M. Zhao, "The drug cocktail network," *BMC Syst Biol*, vol. 6 Suppl 1, p. S5, Jul 16 2012.

[252] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*: Wiley. com, 2013.

[253] J. Hardy and A. Singleton, "Genomewide association studies and human disease," *The New England Journal of Medicine*, vol. 360, pp. 1759–1768, 2009.

[254] L. C. Tranchevent, F. B. Capdevila, D. Nitsch, B. De Moor, P. De Causmaecker, and M. Y., "A guide to web tools to prioritize candidate genes," *Briefings in Bioinformatics*, vol. 12, pp. 22-32, 2011.

[255] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, pp. 107–117, 1998.

[256] M. Krallinger, A. Valencia, and L. Hirschman, "Linking genes to literature: text mining, information extraction, and retrieval applications for biology.," *Genome Biology*, vol. 9, p. S8, 2008.

[257] G. Grimes, T. Wen, M. Mewissen, R. Baxter, S. Moodie, J. Beattie, *et al.*, "PDQ Wizard: automated prioritization and characterization of gene and protein lists using biomedical literature," *Bioinformatics*, vol. 22, pp. 2055-2057, 2006.

[258] R. A.-A. Erhardt, R. Schneider, and C. Blaschke, "Status of text-mining techniques applied to biomedical text," *Drug Discovery Today*, vol. 11, pp. 315-325, 2006.

[259] L. J. Jensen, J. Saric, and P. Bork, "Literature mining for the biologist: from information retrieval to biological discovery," *Nature Reviews Genetics*, vol. 7, pp. 119-129, 2006.

[260] D. Hristovski, A. Kastrin, B. Peterlin, and T. Rindflesch, "Combining semantic relations and DNA microarray data for novel hypotheses generation," in *Linking Literature, Information, and Knowledge for Biology*. vol. 6004, C. Blaschke and H. Shatkay, Eds., ed: Springer, 2010, pp. 53–61.

[261] R. Frijters, M. van Vugt, R. Smeets, R. van Schaik, J. de Vlieg, and W. Alkema, "Literature mining for the discovery of hidden connections between drugs, genes and diseases," *PLoS Computational Biology*, vol. 6, p. e1000943, 2010.

[262] I. Petrič, T. Urbančič, B. Cestnik, and M. Macedoni-Lukšič, "Literature mining method RaJoLink for uncovering relations between biomedical concepts," *J Biomed Inform*, vol. 42, pp. 219-227, 2009.

[263] B. Gyorffy, A. Lánczky, and Z. Szállási, "Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients," *Endocrine-Related Cancer*, vol. 19, pp. 197-208, 2012.

[264] B. C. Cooper, A. K. Sood, C. S. Davis, J. M. Ritchie, J. I. Sorosky, B. Anderson, *et al.*, "Preoperative CA 125 levels: an independent prognostic factor for epithelial ovarian cancer," *Obstet Gynecol*, vol. 100, pp. 59-64, Jul 2002.

[265] A. Gadducci, S. Cosio, A. Fanucchi, S. Negri, R. Cristofani, and A. R. Genazzani, "The predictive and prognostic value of serum CA 125 half-life during paclitaxel/platinum-based chemotherapy in patients with advanced ovarian carcinoma," *Gynecol Oncol*, vol. 93, pp. 131-6, Apr 2004.

[266] A. Gadducci, P. Zola, F. Landoni, T. Maggino, E. Sartori, T. Bergamino, *et al.*, "Serum half-life of CA 125 during early chemotherapy as an independent prognostic variable for patients with advanced epithelial ovarian cancer: results of a multicentric Italian study," *Gynecol Oncol*, vol. 58, pp. 42-7, Jul 1995.

[267] J. M. Riedinger, J. Wafflart, G. Ricolleau, N. Eche, H. Larbre, J. P. Basuyau, *et al.*, "CA 125 half-life and CA 125 nadir during induction chemotherapy are independent predictors of epithelial ovarian cancer outcome: results of a French multicentric study," *Ann Oncol*, vol. 17, pp. 1234-8, Aug 2006.

[268] D. Katsaros, W. Cho, R. Singal, S. Fracchioli, I. A. Rigault De La Longrais, R. Arisio, *et al.*, "Methylation of tumor suppressor gene p16 and prognosis of epithelial ovarian cancer," *Gynecol Oncol*, vol. 94, pp. 685-92, Sep 2004.

[269] S. Kommoss, A. du Bois, R. Ridder, M. J. Trunk, D. Schmidt, J. Pfisterer, *et al.*, "Independent prognostic significance of cell cycle regulator proteins p16(INK4a) and pRb in advanced-stage ovarian carcinoma including optimally debulked patients: a translational research subprotocol of a randomised study of the Arbeitsgemeinschaft Gynaekologische Onkologie Ovarian Cancer Study Group," *Br J Cancer*, vol. 96, pp. 306-13, Jan 29 2007.

[270] L. Sui, Y. Dong, M. Ohno, Y. Watanabe, K. Sugimoto, and M. Tokuda, "Survivin expression and its correlation with cell proliferation and prognosis in epithelial ovarian tumors," *Int J Oncol*, vol. 21, pp. 315-20, Aug 2002.

[271] A. Gadducci, M. Ferdeghini, S. Cosio, A. Fanucchi, R. Cristofani, and A. R. Genazzani, "The clinical relevance of serum CYFRA 21-1 assay in patients with ovarian cancer," *Int J Gynecol Cancer*, vol. 11, pp. 277-82, Jul-Aug 2001.

[272] C. Tempfer, L. Hefler, H. Heinzl, A. Loesch, G. Gitsch, H. Rumpold, *et al.*, "CYFRA 21-1 serum levels in women with adnexal masses and inflammatory diseases," *Br J Cancer*, vol. 78, pp. 1108-12, Oct 1998.

[273] A. Bali, P. M. O'Brien, L. S. Edwards, R. L. Sutherland, N. F. Hacker, and S. M. Henshall, "Cyclin D1, p53, and p21Waf1/Cip1 expression is predictive of poor clinical outcome in serous epithelial ovarian cancer," *Clin Cancer Res*, vol. 10, pp. 5168-77, Aug 1 2004.

[274] G. Ferrandina, A. Stoler, A. Fagotti, F. Fanfani, R. Sacco, A. De Pasqua, *et al.*, "p21WAF1/CIP1 protein expression in primary ovarian cancer," *Int J Oncol*, vol. 17, pp. 1231-5, Dec 2000.

[275] J. Plisiecka-Halasa, G. Karpinska, T. Szymanska, I. Ziolkowska, R. Madry, A. Timorek, *et al.*, "P21WAF1, P27KIP1, TP53 and C-MYC analysis in 204 ovarian carcinomas treated with platinum-based regimens," *Ann Oncol*, vol. 14, pp. 1078-85, Jul 2003.

[276] H. Brustmann, "Immunohistochemical detection of human telomerase reverse transcriptase (hTERT) and c-kit in serous ovarian carcinoma: a clinicopathologic study," *Gynecol Oncol*, vol. 98, pp. 396-402, Sep 2005.

[277] E. P. Diamandis, A. Scorilas, S. Fracchioli, M. Van Gramberen, H. De Bruijn, A. Henrik, *et al.*, "Human kallikrein 6 (hK6): a new potential serum biomarker for diagnosis and prognosis of ovarian carcinoma," *J Clin Oncol*, vol. 21, pp. 1035-43, Mar 15 2003.

[278] P. Korkolopoulou, I. Vassilopoulos, A. E. Konstantinidou, H. Zorzos, E. Patsouris, E. Agapitos, *et al.*, "The combined evaluation of p27Kip1 and Ki-67 expression provides independent information on overall survival of ovarian carcinoma patients," *Gynecol Oncol*, vol. 85, pp. 404-14, Jun 2002.

[279] V. Masciullo, G. Ferrandina, B. Pucci, F. Fanfani, S. Lovergine, J. Palazzo, *et al.*, "p27Kip1 expression is associated with clinical outcome in advanced epithelial ovarian cancer: multivariate analysis," *Clin Cancer Res*, vol. 6, pp. 4816-22, Dec 2000.

[280] E. W. Newcomb, M. Sosnow, R. I. Demopoulos, A. Zeleniuch-Jacquotte, J. Sorich, and J. L. Speyer, "Expression of the cell cycle inhibitor p27KIP1 is a new prognostic marker associated with survival in epithelial ovarian tumors," *Am J Pathol*, vol. 154, pp. 119-25, Jan 1999.

[281] A. Schmider-Ross, O. Pirsig, E. Gottschalk, C. Denkert, W. Lichtenegger, and A. Reles, "Cyclin-dependent kinase inhibitors CIP1 (p21) and KIP1 (p27) in ovarian cancer," *J Cancer Res Clin Oncol*, vol. 132, pp. 163-70, Mar 2006.

[282] I. Skirnisdottir, T. Seidal, and B. Sorbe, "A new prognostic model comprising p53, EGFR, and tumor grade in early stage epithelial ovarian carcinoma and avoiding the problem of inaccurate surgical staging," *Int J Gynecol Cancer*, vol. 14, pp. 259-70, Mar-Apr 2004.

[283] A. Psyrri, M. Kassar, Z. Yu, A. Bamias, P. M. Weinberger, S. Markakis, *et al.*, "Effect of epidermal growth factor receptor expression level on survival in patients with epithelial ovarian cancer," *Clin Cancer Res*, vol. 11, pp. 8637-43, Dec 15 2005.

[284] L. Y. Luo, D. Katsaros, A. Scorilas, S. Fracchioli, R. Piccinno, I. A. Rigault de la Longrais, *et al.*, "Prognostic value of human kallikrein 10 expression in epithelial ovarian carcinoma," *Clin Cancer Res*, vol. 7, pp. 2372-9, Aug 2001.

[285] Y. Dong, M. D. Walsh, M. A. McGuckin, M. C. Cummings, B. G. Gabrielli, G. R. Wright, *et al.*, "Reduced expression of retinoblastoma gene product (pRB) and high expression of p53 are associated with poor prognosis in ovarian cancer," *International Journal of Cancer*, vol. 74, pp. 407-15, Aug 22 1997.

[286] A. E. Konstantinidou, P. Korkolopoulou, I. Vassilopoulos, A. Tsenga, I. Thymara, E. Agapitos, *et al.*, "Reduced retinoblastoma gene protein to Ki-67 ratio is an adverse prognostic indicator for ovarian adenocarcinoma patients," *Gynecol Oncol*, vol. 88, pp. 369-78, Mar 2003.

[287] H. Lassus, A. Leminen, A. Vayrynen, G. Cheng, J. A. Gustafsson, J. Isola, *et al.*, "ERBB2 amplification is superior to protein expression status in predicting patient outcome in serous ovarian carcinoma," *Gynecol Oncol*, vol. 92, pp. 31-9, Jan 2004.

[288] G. Scambia, U. Testa, P. Benedetti Panici, E. Foti, R. Martucci, A. Gadducci, *et al.*, "Prognostic significance of interleukin 6 serum levels in patients with ovarian cancer," *Br J Cancer*, vol. 71, pp. 354-6, Feb 1995.

[289] D. S. Suh, M. S. Yoon, K. U. Choi, and J. Y. Kim, "Significance of E2F-1 overexpression in epithelial ovarian cancer," *Int J Gynecol Cancer*, vol. 18, pp. 492-8, May-Jun 2008.

[290] K. Sawada, A. R. Radjabi, N. Shinomiya, E. Kistner, H. Kenny, A. R. Becker, *et al.*, "c-Met overexpression is a prognostic factor in ovarian cancer and an effective target for inhibition of peritoneal dissemination and invasion," *Cancer Res*, vol. 67, pp. 1670-9, Feb 15 2007.

[291] A. J. Lambeck, A. P. Crijns, N. Leffers, W. J. Sluiter, K. A. ten Hoor, M. Braid, *et al.*, "Serum cytokine profiling as a diagnostic and prognostic tool in ovarian cancer: a potential role for interleukin 7," *Clin Cancer Res*, vol. 13, pp. 2385-91, Apr 15 2007.

[292] D. Reimer, S. Sadr, A. Wiedemair, S. Stadlmann, N. Concin, G. Hofstetter, *et al.*, "Clinical relevance of E2F family members in ovarian cancer--an evaluation in a training set of 77 patients," *Clin Cancer Res*, vol. 13, pp. 144-51, Jan 1 2007.

[293] P. L. Torng, T. L. Mao, W. Y. Chan, S. C. Huang, and C. T. Lin, "Prognostic significance of stromal metalloproteinase-2 in ovarian adenocarcinoma and its relation to carcinoma progression," *Gynecol Oncol*, vol. 92, pp. 559-67, Feb 2004.

[294] C. Marth, H. Fiegl, A. G. Zeimet, E. Muller-Holzner, M. Deibl, W. Doppler, *et al.*, "Interferon-gamma expression is an independent prognostic factor in ovarian cancer," *Am J Obstet Gynecol*, vol. 191, pp. 1598-605, Nov 2004.

[295] S. Sillanpaa, M. Anttila, K. Voutilainen, K. Ropponen, T. Turpeenniemi-Hujanen, U. Puistola, *et al.*, "Prognostic significance of matrix metalloproteinase-9 (MMP-9) in epithelial ovarian cancer," *Gynecol Oncol*, vol. 104, pp. 296-303, Feb 2007.

[296] L. Hefler, K. Mayerhofer, A. Nardi, A. Reinthaller, C. Kainz, and C. Tempfer, "Serum soluble Fas levels in ovarian cancer," *Obstet Gynecol*, vol. 96, pp. 65-9, Jul 2000.

[297] R. Konno, T. Takano, S. Sato, and A. Yajima, "Serum soluble fas level as a prognostic factor in patients with gynecological malignancies," *Clin Cancer Res*, vol. 6, pp. 3576-80, Sep 2000.

[298] F. Buttitta, A. Marchetti, A. Gadducci, S. Pellegrini, M. Morganti, V. Carnicelli, *et al.*, "p53 alterations are predictive of chemoresistance and aggressiveness in ovarian carcinomas: a molecular and immunohistochemical study," *Br J Cancer*, vol. 75, pp. 230-5, 1997.

[299] A. Reles, W. H. Wen, A. Schmider, C. Gee, I. B. Runnebaum, U. Kilian, *et al.*, "Correlation of p53 mutations with resistance to platinum-based chemotherapy and shortened survival in ovarian cancer," *Clin Cancer Res*, vol. 7, pp. 2984-97, Oct 2001.

[300] A. A. Kamat, M. Fletcher, L. M. Gruman, P. Mueller, A. Lopez, C. N. Landen, Jr., *et al.*, "The clinical relevance of stromal matrix metalloproteinase expression in ovarian cancer," *Clin Cancer Res*, vol. 12, pp. 1707-14, Mar 15 2006.

[301] L. A. Hefler, R. Zeillinger, C. Grimm, A. K. Sood, W. F. Cheng, A. Gadducci, *et al.*, "Preoperative serum vascular endothelial growth factor as a prognostic parameter in ovarian cancer," *Gynecol Oncol*, vol. 103, pp. 512-7, Nov 2006.

[302] K. Becker, P. Pancoska, N. Concin, K. Vanden Heuvel, N. Slade, M. Fischer, *et al.*, "Patterns of p73 N-terminal isoform expression and p53 status have prognostic value in gynecological cancers," *Int J Oncol*, vol. 29, pp. 889-902, Oct 2006.

[303] A. A. Secord, P. S. Lee, K. M. Darcy, L. J. Havrilesky, L. A. Grace, J. R. Marks, *et al.*, "Maspin expression in epithelial ovarian cancer and associations with poor prognosis: a Gynecologic Oncology Group study," *Gynecol Oncol*, vol. 101, pp. 390-7, Jun 2006.

[304] F. Barbieri, P. Lorenzi, N. Ragni, G. Schettini, C. Bruzzo, F. Pedulla, *et al.*, "Overexpression of cyclin D1 is associated with poor survival in epithelial ovarian cancer," *Oncology*, vol. 66, pp. 310-5, 2004.

[305] I. Skirnisdottir, B. Sorbe, and T. Seidal, "P53, bcl-2, and bax: their relationship and effect on prognosis in early stage epithelial ovarian carcinoma," *Int J Gynecol Cancer*, vol. 11, pp. 147-58, Mar-Apr 2001.

[306] Y. T. Tai, S. Lee, E. Niloff, C. Weisman, T. Strobel, and S. A. Cannistra, "BAX protein expression and clinical outcome in epithelial ovarian cancer," *J Clin Oncol*, vol. 16, pp. 2583-90, Aug 1998.

[307] M. Thrall, H. H. Gallion, R. Kryscio, M. Kapali, D. K. Armstrong, and J. A. DeLoia, "BRCA1 expression in a large series of sporadic ovarian carcinomas: a Gynecologic Oncology Group study," *Int J Gynecol Cancer*, vol. 16 Suppl 1, pp. 166-71, Jan-Feb 2006.

[308] G. Levidou, P. Korkolopoulou, I. Thymara, I. Vassilopoulos, A. A. Saetta, H. Gakiopoulou, *et al.*, "Expression and prognostic significance of cyclin D3 in ovarian adenocarcinomas," *Int J Gynecol Pathol*, vol. 26, pp. 410-7, Oct 2007.

[309] V. Materna, P. Surowiak, E. Markwitz, M. Spaczynski, M. Drag-Zalesinska, M. Zabel, *et al.*, "Expression of factors involved in regulation of DNA mismatch repair- and apoptosis pathways in ovarian cancer patients," *Oncol Rep*, vol. 17, pp. 505-16, Mar 2007.

[310] K. M. Darcy, C. Tian, and E. Reed, "A Gynecologic Oncology Group study of platinum-DNA adducts and excision repair cross-complementation group 1 expression in optimal, stage III epithelial ovarian cancer treated with platinum-taxane chemotherapy," *Cancer Res,* vol. 67, pp. 4474-81, May 1 2007.

[311] I. Bedrosian, C. Lee, S. L. Tucker, S. L. Palla, K. Lu, and K. Keyomarsi, "Cyclin E-associated kinase activity predicts response to platinum-based chemotherapy," *Clin Cancer Res,* vol. 13, pp. 4800-6, Aug 15 2007.

[312] J. Farley, L. M. Smith, K. M. Darcy, E. Sobel, D. O'Connor, B. Henderson, *et al.*, "Cyclin E expression is a significant predictor of survival in advanced, suboptimally debulked ovarian epithelial cancers: a Gynecologic Oncology Group study," *Cancer Res,* vol. 63, pp. 1235-41, Mar 15 2003.

[313] D. G. Rosen, G. Yang, M. T. Deavers, A. Malpica, J. J. Kavanagh, G. B. Mills, *et al.*, "Cyclin E expression is correlated with tumor progression and predicts a poor prognosis in patients with ovarian carcinoma," *Cancer,* vol. 106, pp. 1925-32, May 1 2006.

[314] L. Sui, Y. Dong, M. Ohno, K. Sugimoto, Y. Tai, T. Hando, *et al.*, "Implication of malignancy and prognosis of p27(kip1), Cyclin E, and Cdk2 expression in epithelial ovarian tumors," *Gynecol Oncol,* vol. 83, pp. 56-63, Oct 2001.

[315] A. Psyrri, Z. Yu, A. Bamias, P. M. Weinberger, S. Markakis, D. Kowalski, *et al.*, "Evaluation of the prognostic value of cellular inhibitor of apoptosis protein in epithelial ovarian cancer using automated quantitative protein analysis," *Cancer Epidemiol Biomarkers Prev,* vol. 15, pp. 1179-83, Jun 2006.

[316] K. Huhtinen, P. Suvitie, J. Hiissa, J. Junnila, J. Huvila, H. Kujari, *et al.*, "Serum HE4 concentration differentiates malignant ovarian tumours from ovarian endometriotic cysts," *Br J Cancer,* vol. 100, pp. 1315-9, Apr 21 2009.

[317] R. G. Moore, M. Jabre-Raughley, A. K. Brown, K. M. Robison, M. C. Miller, W. J. Allard, *et al.*, "Comparison of a novel multiple marker assay vs the Risk of Malignancy Index for the prediction of epithelial ovarian cancer in patients with a pelvic mass," *Am J Obstet Gynecol,* vol. 203, pp. 228 e1-6, Sep 2010.

[318] R. G. Moore, D. S. McMeekin, A. K. Brown, P. DiSilvestro, M. C. Miller, W. J. Allard, *et al.*, "A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass," *Gynecol Oncol,* vol. 112, pp. 40-6, Jan 2009.

[319] K. Kudoh, Y. Ichikawa, S. Yoshida, M. Hirai, Y. Kikuchi, I. Nagata, *et al.*, "Inactivation of p16/CDKN2 and p15/MTS2 is associated with prognosis and response to chemotherapy in ovarian cancer," *International Journal of Cancer,* vol. 99, pp. 579-82, Jun 1 2002.

[320] J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga, "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization," *Nucleic Acids Res,* vol. 37, pp. W305-11, Jul 2009.

[321] L. C. Tranchevent, R. Barriot, S. Yu, S. Van Vooren, P. Van Loo, B. Coessens, *et al.*, "ENDEAVOUR update: a web resource for gene prioritization in multiple species," *Nucleic Acids Res,* vol. 36, pp. W377-84, Jul 1 2008.

[322] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, *et al.*, "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.," *Nucleic Acids Research,* vol. 39, pp. 561-568, 2011.

[323] W. Li, S. Xu, S. Lin, and W. Zhao, "Overexpression of runt-related transcription factor-2 is associated with advanced tumor progression and poor prognosis in epithelial ovarian cancer," *BioMed Research International,* vol. 2012, 2012.

[324] W. Li, Z. Liu, L. Chen, L. Zhou, and Y. Yao, "MicroRNA-23b is an independent prognostic marker and suppresses ovarian cancer progression by targeting runt-related transcription factor-2," *FEBS letters,* vol. 588, pp. 1608-1615, 2014.

[325] Y.-Q. Wang, M.-D. Xu, W.-W. Weng, P. Wei, Y.-S. Yang, and X. Du, "BCL6 is a negative prognostic factor and exhibits pro-oncogenic activity in ovarian cancer," *American journal of cancer research,* vol. 5, p. 255, 2015.

[326] W. Shan, J. Li, Y. Bai, and X. Lu, "miR-339-5p inhibits migration and invasion in ovarian cancer cell lines by targeting NACC1 and BCL6," *Tumor Biology,* vol. 37, pp. 5203-5211, 2016.

[327] A. Goel, C. Arnold, P. Tassone, D. Chang, D. Niedzwiecki, J. Dowell, *et al.*, "Epigenetic inactivation of RUNX3 in microsatellite unstable sporadic colon cancers," *International Journal of Cancer,* vol. 112, pp. 754-759, 2004.

[328] G. Little, H. Noushmehr, S. Baniwal, B. Berman, G. Coetzee, and B. Frenkel, "Genome-wide Runx2 occupancy in prostate cancer cells suggests a role in regulating secretion," *Nucleic Acids Research* vol. 40, pp. 3538-3547, 2012.

[329] M. Tandon, K. Gokul, S. Ali, Z. Chen, J. Lian, G. Stein, *et al.*, "Runx2 mediates epigenetic silencing of the bone morphogenetic protein-3B (BMP-3B/GDF10) in lung cancer cells," *Molecular Cancer,* vol. 11, 2012.

[330] N. Chimge, S. Baniwal, G. Little, Y. Chen, M. Kahn, D. Tripathy, *et al.*, "Regulation of breast cancer metastasis by Runx2 and estrogen signaling: the role of SNAI2," *Breast Cancer Research,* vol. 13, p. R127, 2011.

[331] J. Martin, M. Zielenska, G. Stein, A. van Wijnen, and J. Squire, "The Role of RUNX2 in Osteosarcoma Oncogenesis," *Sarcoma,* vol. 282745, 2011.

[332] L. Dalle Carbonare, A. Frigo, G. Francia, M. Davì, L. Donatelli, C. Stranieri, *et al.*, "Runx2 mRNA Expression in the Tissue, Serum, and Circulating Non-Hematopoietic Cells of Patients with Thyroid Cancer," *The Journal of Clinical Endocrinology & Metabolism*, 2012.

[333] M. van der Deen, J. Akech, T. Wang, T. FitzGerald, D. Altieri, L. Languino, *et al.*, "The cancer-related Runx2 protein enhances cell growth and responses to androgen and TGFbeta in prostate cancer cells," *Journal of Cellular Biochemistry*, vol. 109, pp. 828-837, 2010.

[334] J. Pratap, Wixted, JJ, Gaur, T, Zaidi, SK, Dobson, J, Gokul, KD, Hussain, S, van Wijnen, AJ, Stein, JL, Stein, GS, Lian, JB., "Runx2 transcriptional activation of Indian Hedgehog and a downstream bone metastatic pathway in breast cancer cells," *Cancer Research*, vol. 68, pp. 7795-7802, 2008.

[335] T. Fekete, E. Rásó, I. Pete, B. Tegze, I. Liko, G. Munkácsy, *et al.*, "Meta-analysis of gene expression profiles associated with histological classification and survival in 829 ovarian cancer samples," *International Journal of Cancer*, vol. 131, pp. 95-105, 2012.

[336] Z.-Q. Wang, M. Keita, M. Bachvarova, S. Gobeil, C. Morin, M. Plante, *et al.*, "Inhibition of RUNX2 transcriptional activity blocks the proliferation, migration and invasion of epithelial ovarian carcinoma cells," *PloS one*, vol. 8, p. e74384, 2013.

[337] S. Wagner, M. Ahearne, and P. Ko Ferrigno, "The role of BCL6 in lymphomas and routes to therapy," *British Journal of Haematology*, vol. 152, pp. 3-12, 2011.

[338] F. Pellicano and T. Holyoake, "Assembling defenses against therapy-resistant leukemic stem cells: Bcl6 joins the ranks," *The Journal of Experimental Medicine*, vol. 208, pp. 2155-2158, 2011.

[339] A. Pinto, S. André, G. Silva, S. Vieira, A. Santos, S. Dias, *et al.*, "BCL-6 oncoprotein in breast cancer: loss of expression in disease progression," *Pathobiology*, vol. 76, pp. 235-242, 2009.

[340] Y. Hirata, N. Ogasawara, M. Sasaki, T. Mizushima, T. Shimura, T. Mizoshita, *et al.*, "BCL6 degradation caused by the interaction with the C-terminus of pro-HB-EGF induces cyclin D2 expression in gastric cancers," *British Journal of Cancer*, vol. 100, pp. 1320-1329, 2009.

[341] A. Dmitriev, V. Kashuba, K. Haraldson, V. Senchenko, T. Pavlova, A. Kudryavtseva, *et al.*, "Genetic and epigenetic analysis of non-small cell lung cancer with NotI-microarrays," *Epigenetics*, vol. 7, pp. 502-513, 2012.

[342] T. Tran, F. Utama, J. Lin, N. Yang, A. Sjolund, A. Ryder, *et al.*, "Prolactin inhibits BCL6 expression in breast cancer through a Stat5a-dependent mechanism," *Cancer Research*, vol. 70, pp. 1711-1721, 2010.

[343] N. Charoenphandhu, J. Teerapornpuntakit, M. Methawasin, K. Wongdee, K. Thongchote, and N. Krishnamra, "Prolactin decreases expression of Runx2, osteoprotegerin, and RANKL in primary osteoblasts derived from tibiae of adult female rats," *Canadian Journal of Physiology and Pharmacology*, vol. 86, pp. 240-248, 2008.

[344] M. Aberg, M. Johnell, M. Wickström, and A. Siegbahn, "Tissue Factor/ FVIIa prevents the extrinsic pathway of apoptosis by regulation of the tumor suppressor Death-Associated Protein Kinase 1 (DAPK1)," *Thrombosis Research*, vol. 127, pp. 141-148, 2011.

[345] W. Tapper, V. Hammond, S. Gerty, S. Ennis, P. Simmonds, and E. Collins A; Prospective study of Outcomes in Sporadic versus Hereditary breast cancer (POSH) Steering Group, D., "The influence of genetic variation in 30 selected genes on the clinical characteristics of early onset breast cancer," *Breast Cancer Research*, vol. 10, p. R108, 2008.

[346] F. de Fraipont, G. Levallet, C. Creveuil, E. Bergot, M. Beau-Faller, M. Mounawar, *et al.*, "An Apoptosis Methylation Prognostic Signature for Early Lung Cancer in the IFCT-0002 Trial," *Clinical Cancer Research*, vol. 18, pp. 2976-2986, 2012.

[347] H. Yoo, H. Byun, B. Kim, K. Lee, S. Park, and S. Rho, "DAPk1 inhibits NF-κB activation through TNF-α and INF-γ-induced apoptosis," *Cell Signalling*, vol. 24, pp. 1471-1477, 2012.

[348] L. S. Pongor, R. Vera, and B. Ligeti, "Fast and sensitive alignment of microbial whole genome sequencing reads to large sequence datasets on a desktop PC: application to metagenomic datasets and pathogen identification," *PloS one*, vol. 9, p. e103441, 2014.

[349] D. Koslicki, S. Foucart, and G. Rosen, "WGSQuikr: Fast Whole-Genome Shotgun Metagenomic Classification," *PloS one*, vol. 9, p. e91784, 2014.

[350] J. Dröge and A. C. McHardy, "Taxonomic binning of metagenome samples generated by next-generation sequencing technologies," *Brief Bioinform*, vol. 13, pp. 646-655, November 1, 2012 2012.

[351] S. Powell, K. Forslund, D. Szklarczyk, K. Trachana, A. Roth, J. Huerta-Cepas, *et al.*, "eggNOG v4.0: nested orthology inference across 3686 organisms," *Nucleic Acids Res*, vol. 42, pp. D231-9, Jan 2014.

[352] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, *et al.*, "The COG database: an updated version includes eukaryotes," *BMC Bioinformatics*, vol. 4, p. 41, Sep 11 2003.

[353] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, *et al.*, "Gene Ontology: tool for the unification of biology," *Nat Genet*, vol. 25, pp. 25-29, 2000.

[354] C. B. Moore, J. R. Wallace, A. T. Frase, S. A. Pendergrass, and M. D. Ritchie, "BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge," *BMC Med Gen*, vol. 6, p. S6, 2013.

[355] S. Dhir, M. Pacurar, D. Franklin, Z. Gáspári, A. Kertész-Farkas, A. Kocsor, *et al.*, "Detecting atypical examples of known domain types by sequence similarity searching: The SBASE domain library approach," *Curr Protein Pept Sci*, vol. 11, pp. 538-549, 2010.

[356] J. Murvai, K. Vlahovicek, E. Barta, S. Parthasarathy, H. Hegyi, F. Pfeiffer, *et al.*, "The domain-server: direct prediction of protein domain-homologies from BLAST search," *Bioinformatics*, vol. 15, pp. 343-344, 1999.

[357] S. Lu, P. Deng, X. Liu, J. Luo, R. Han, X. Gu, *et al.*, "Solution structure of the major alpha-amylase inhibitor of the crop plant amaranth," *J Biol Chem*, vol. 274, pp. 20473-8, Jul 16 1999.

[358] G. Maravic, J. M. Bujnicki, M. Feder, S. Pongor, and M. Flogel, "Alanine-scanning mutagenesis of the predicted rRNA-binding domain of ErmC' redefines the substrate-binding site and suggests a model for protein-RNA interactions," *Nucleic Acids Res*, vol. 31, pp. 4941-9, Aug 15 2003.

[359] G. Maravic, M. Feder, S. Pongor, M. Flogel, and J. M. Bujnicki, "Mutational analysis defines the roles of conserved amino acid residues in the predicted catalytic pocket of the rRNA:m6A methyltransferase ErmC'," *J Mol Biol*, vol. 332, pp. 99-109, Sep 5 2003.

[360] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nat Methods*, vol. 9, pp. 357-9, Apr 2012.

[361] S. A. Shiryev, J. S. Papadopoulos, A. A. Schaffer, and R. Agarwala, "Improved BLAST searches using longer words for protein seeding," *Bioinformatics*, vol. 23, pp. 2949-51, Nov 1 2007.

[362] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller, "A greedy algorithm for aligning DNA sequences," *J Comput Biol*, vol. 7, pp. 203-14, Feb-Apr 2000.

[363] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic acids research*, vol. 33, pp. D514-D517, 2005.