Propagation on Molecular Interaction Networks: Prediction of Effective Drug Combinations and Biomarkers in Cancer Treatment

Article *in* Current Pharmaceutical Design · October 2016

DOI: 10.2174/1381612822666161021162727

5 authors, including:

Balázs Ligeti
Pázmány Péter Catholic University
17 PUBLICATIONS **102** CITATIONS
SEE PROFILE

Otilia Menyhart
Hungarian Academy of Sciences
33 PUBLICATIONS **529** CITATIONS
SEE PROFILE

Ingrid Petric
University of Nova Gorica
16 PUBLICATIONS **142** CITATIONS
SEE PROFILE

Balázs Györffy
Hungarian Academy of Sciences
571 PUBLICATIONS **9,340** CITATIONS
SEE PROFILE

Send Orders for Reprints to reprints@benthamscience.ae

Current Pharmaceutical Design, 2017, *23*, 5-28

REVIEW ARTICLE

Propagation on Molecular Interaction Networks: Prediction of Effective Drug Combinations and Biomarkers in Cancer Treatment

Balázs Ligeti[1,*], Otilia Menyhárt[2], Ingrid Petrič[3], Balázs Győrffy[2,4] and Sándor Pongor[1]

[1]*Faculty of Information Technology, Pázmány Péter Catholic University, Budapest, Hungary;* [2]*MTA TTK Lendület Cancer Biomarker Research Group, Budapest, Hungary;* [3]*Centre for Systems and Information Technologies, University of Nova Gorica, Nova Gorica, Slovenia;* [4]*2nd Department of Pediatrics, Semmelweis University, Budapest, Hungary*

Abstract: ***Background*:** Biomedical sciences use a variety of data sources on drug molecules, genes, proteins, 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.

***Methods*:** The method we applied was to test whether novel drug combinations or novel biomarkers can be predicted from a network of existing oncological databases. We start from the hypothesis that novel, implicit links can be discovered between the network neighborhoods of data items.

***Results*:** We 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. In a second case study we employed this strategy to the discovery of novel biomarkers based on text analysis. In 2012 we prioritized 10 potential biomarkers for ovarian cancers, 2 of which were in fact described as such in the subsequent years.

***Conclusion*:** The strategy seems to hold promises for prioritizing new drug combinations or new biomarkers for experimental testing. Its use is naturally limited by the sparsity and the quality of experimental data, however both of these aspects are expected to improve given the development of current databases.

Balázs Ligeti

ARTICLE HISTORY

Received: August 4, 2016
Accepted: October 17, 2016

DOI: 10.2174/1381612822666161021162727

Keywords: Drug-drug combinations, drug-drug interactions, combination chemotherapy, ovarian cancer biomarkers, breast cancer.

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 looked 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 are just taken over from homologous proteins of various organisms. Also, we cannot be sure whether or not 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 *etc*. - manually curated parts which contain reliable information on a given problem. This is a tedious and labor-intensive solution which is justified only in very important fields. Cancer-specific data networks, the subject of this work, are an example of this approach. In addition, there are two major information sources that can help data-sparsity problems. On the one hand, various high-throughput experimental methods (two hybrid systems, 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. 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.

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 [1, 2]. Oncologists use a variety of traditional databases, but there are a number of data-collection efforts dedicated to the collection 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 we are concerned with two types of questions that can be ad-

*Address correspondence to this author at the Faculty of Information Technology, Pázmány Péter Catholic University 1083 Budapest, Práter str. 50/A, Budapest, Hungary; E-mail: ligeti.balazs@itk.ppke.hu

1381-6128/17 $58.00+.00

dressed via integrated data networks: i) finding drug combinations potentially useful for cancer therapy. We tackle this problem by using a simple network overlap measure applied to data networks. And ii) finding novel gene-disease associations in ovarian cancer for generating a list of potential biomarkers. We approach this problem using a text mining approach applied to MEDLINE abstracts [3] as well as the STRING database [4]. Section 2 is an introduction to the problem of chemotherapy. Section 3 describes the main database types used in this project. Section 4 contains the mathematical and computational background. Section 5 describes the identification of drug combinations via a network overlap measure. Section 6 describes the principle of hypothesis generation via network overlap analysis. Section 7 discusses conclusions and future trends.

2. CHEMOTHERAPY

Chemotherapy is the most frequently used first-line treatment of cancer. Chemoterapeutic 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, 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 consist 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. 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 cyclophosphamids. 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]. 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. Antimetabolites are structurally similar analogues of naturally occurring molecules. They interfere with metabolic processes either by 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 pyrimidin synthesis, while mercaptopurin, azathioprin, pentostatin and thioguanin hamper purin production. Vinca alcaloids and taxanes consist of cell-cycle-phase specific antimicrotubule blocking chemotherapy agents. During the S phase vinca alcaloids 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]. Topoisomerase inhibitors, such as camptothecin analogs (irinotecan) inhibit DNA elongation by blocking topoisomerase I in the S phase of the cell cycle [13]. Anthracyclines inhibit both topoisomerase I and II [5].

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].

2.1. 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 require 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 down-stage 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, capboplatin, docetaxel, albumin-bound paclitaxel, cisplatin, epirubicin, ixabepilone. Chemotherapy combinations are listed in Table **1**.

2.2. 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

Table 1. Chemotherapy combinations for recurrent or metastatic breast cancer adapted from the Guidelines of the US National Comprehensive Cancer Network.

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

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 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].

2.3 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].

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 by-passing 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 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.

2.4 Ovarian Cancer

Ovarian cancer is the fifth most common cancer in women worldwide and the most deadly gynecologic malignancy, as less than 30% of patients with advanced disease reach 5 year survival [48]. It is characterized with extreme heterogeneity, as a consider-

able 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 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]. Germline mutations, and other genetic/epigenetic events, such as promoter methylation related to BRCA1/2 genes render homologous recombination (HR) DNA repair pathways defective in about 50% of HGSOCs [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].

2.5. 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 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 inactivation only in a fraction of patients [80].

3. CANCER AND DRUG-RELATED DATA-NETWORKS

Biological databases, including the ones related to cancer chemotherapy, 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 [81, 82], 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. 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 [83]. The current solution is to build partial networks that allow one to answer a few questions related to a given project. 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 that act on proteins that are in physical contacts. Such a network allows one to find out if two drugs are likely to act on related or interacting targets.

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 **2**). Such projects contain data on cancer mutations, and are often di-

vided 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 are used to build cancer data network consisting 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 interactions and try to gather fine details by studying the interacting partners with methods like x-ray crystallography [84, 85], nuclear magnetic resonance [86, 87], 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 coimmunoprecipitation [80, 88, 89]. 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 [90]. 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 partners 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 [91]. 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 [92] 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*.

General purpose databases such as Uniprot [93], Ensembl [94] or GenBank [95] 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 [92] 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 [96, 97]. 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 drugcard of Drugbank could be rich source of text mining.

TTD database [98] 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.

STITCH [99-101] 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 [102, 103] 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 [104], clinical trials (https://clinicaltrials.gov/), and publications. It also holds information about the individual drug components such as ATC codes, target and cross references. 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 [105, 106], where the side effects are extracted (using controlled vocabulary such as UMLS [82]) from the drug labels. A well-maintained collection of drug side effects are provided by the Tatonetti Lab [107].

Experimental results of protein-protein interaction measurements are deposited in various primary databases such as the Database of Interacting Proteins (DIP) [108], Biomolecular Interaction Network Database (BIND) [109], Molecular Interactions Database (MINT) [110-113], Biological General Repository for Interaction Datasets (BioGRID), Human Protein Reference Database (HPRD), IntAct Molecular Interaction Database [114].

The DIP database contains large number of manually curated and reviewed interactions from numerous species [108, 115]. It also provides some services and visualization tools for the available data [116] and a cytoscape plugin (MiSink) [117]. Different evidences for the interactions were integrated and considered manually.

Human Protein Reference Database (HPRD) [118] 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 also have 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 [114, 119-121] developed and maintained by the European Bioinformatics Institute (EBI), and is 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 [122] (PSI-MI [123, 124], gene ontology [125] and NCBI taxonomy terms [126]) 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 representation and data integration. The MiNTAct [127], Imex [128], Mentha [129] 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-occurence, co-expression, experiments, databases, text mining, 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 [130]. 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 **2**.

Special types of molecular interactions are metabolic and signal transduction molecular interactions. One of the oldest pathway databases is KEGG [165]. 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 [166], similarly to KEGG, is a comprehensive, manually curated, high quality pathway database with support of enrichment analysis and data visualization.

The Human Metabolome Database [167], 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 [168]. The Signalink [169] and NDEx databases [170] 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 [171] 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 [172]). 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 **3**.

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 [83]. The current solution is to build partial networks that allow one to answer a few questions related to a given project. 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 that act on proteins that are in physical contacts. Such a network allows one to find out, for instance, if two drugs are likely to act on related or interacting targets. Construction of such data networks is largely facilitated by database frameworks capable to handle an arbitrary set of biological entities and relationships [182]. 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 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, 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 [182].

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 **2**.

Table 2. Cancer-related databases and resources.

Database	Description	URL	Refs.
Comprehensive Databases and Resources			
TCGA	The Cancer Genome Atlas	http://cancergenome.nih.gov/	[131]
CGP	Cancer Genome Project	http://www.sanger.ac.uk/research/projects/cancergenome/	[132]
CPTAC	Clinical Proteomic Tumor Analysis Consortium	http://proteomics.cancer.gov/programs/cptacnetwork	[133, 134]
ICGC	International Cancer Genome Consortium	https://www.icgc.org/	[135]
Data mining resources			
COSMICMart	BioMart tool for COSMIC	https://cancer.sanger.ac.uk/cosmic/login	[136]
G-2-O	Linking genotype alterations to clinical outcome	http://www.g-2-o.com/	Pmid: 26474971
KM plotter	Survival analysis using multiple gene chip datasets	http://kmplot.com/analysis/	Pmid: 20020197
IntOGen Biomart	BioMart tool for IntOGen	http://biomart.intogen.org/	[137]
UCSC Cancer Genomics Browser	A visualization and analysis tool specialized to cancer data	https://genome-cancer.ucsc.edu/	[138, 139]
ICPS	An Integrative Cancer Profiler System	http://server.bioicps.org/	[133]
NCG 4.0	Network of Cancer Genes	http://ncg.kcl.ac.uk/	[140, 141]
CGWB	The Cancer Genome WorkBench	http://cgap.nci.nih.gov/cgap.html	[142]
CancerMA	A web-based tool for analyzing microarray data	http://www.cancerma.org.uk/information.html	[143]
ICPS	An Integrative Cancer Profiler System	http://server.bioicps.org/	[133]
Databases of genetic variations in cancer			
COSMIC	Catalogue of Somatic Mutations in Cancer	http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/	[144]
CaSNP	Cancer SNP data on CNAs	http://cistrome.dfci.harvard.edu/CaSNP/	[145]
DriverDB	Cancer driver genes and mutation database	http://driverdb.ym.edu.tw/DriverDB/intranet/init.do	[146]
IntOGen	Integrative Oncogenomics	http://www.intogen.org/	[147]
MoKCa	Mutations of Kinases in Cancer database	http://strubiol.icr.ac.uk/extra/mokca/	[148]
CGAP	Cancer Genome Anatomy Project	http://cgap.nci.nih.gov/	[149]
Databases of genetic variations in cancer			
Mitelman Database	Database of chromosome aberrations and gene fusions in cancer	http://cgap.nci.nih.gov/Chromosomes/Mitelman	[150]
CGC	The Cancer Gene Census	http://cancer.sanger.ac.uk/cancergenome/projects/census/	[151]

(Table 2) Contd....

Database	Description	URL	Refs.
Databases of epigenetic, proteomic and transcriptome variations in cancer			
CanProVar	Human Cancer Proteome Variation Database	http://bioinfo.vanderbilt.edu/canprovar/	
MethyCancer	A database of human DNA methylation and cancer	http://methycancer.psych.ac.cn/	
CellLineNavigator	Expression profiles of cancer cell lines	http://www.medicalgenomics.org/celllinenavigator/	[152]
ITTACA	Integrated Tumor Transcriptome Array and Clinical data Analysis	http://bioinfo.curie.fr/ittaca	[137]
PubMeth	Cancer methylation database based on text-mining of PubMed	http://matrix.ugent.be/pubmeth/	[153]
OncomiRDB	A database of experimentally verified oncomiRs	http://bioinfo.au.tsinghua.edu.cn/member/jgu/oncomirdb/	[154]
Cancer specific clinical and drug resources			
CancerDR	Cancer Drug Resistance Database	http://crdd.osdd.net/raghava/cancerdr/	[155]
HPtaa	The Human Potential Tumor Associated Antigen database	http://www.bioinfo.org.cn/hptaa/	[156]
CancerResource	A resource of cancer-relevant compound and protein interactions	http://bioinf-data.charite.de/cancerresource/	[157]
CanGEM	Cancer Genome Mine	http://www.cangem.org/	[158]
DTP	Anti-cancer agent database	http://dtp.nci.nih.gov/docs/cancer/searches/standard_mechanism.html	[159, 160]
ITTACA	Integrated Tumor Transcriptome Array and Clinical data Analysis	http://bioinfo.curie.fr/ittaca	[137]
Cancer-type specific resources and databases			
RCDB	RCDB	http://www.juit.ac.in/attachments/jsr/rcdb/homenew.html	[161]
curatedOvarianData	Clinically annotated data for the ovarian cancer transcriptome	http://bcb.dfci.harvard.edu/ovariancancer/	[162]
PED	Pancreatic Expression Database	http://www.pancreasexpression.org/	[163]
HLungDB	Human Lung Cancer Database	http://www.megabionet.org/bio/hlung/	[164]

5. COMPUTATIONAL BACKGROUND

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*. [183].

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 indeces (i.e. k, v_k or x_k).

Table 3. Representative examples of molecular and molecular interaction databases relevant to cancer therapy.

	Contents	URL	V
1 General purpose databases			
Uniprot	Comprehensive database of protein sequences	http://www.uniprot.org/	[93]
RefSeq	Human genome sequences	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/	[173]
GenBank	Comprehensive database of genetic sequences	http://www.ncbi.nlm.nih.gov/genbank/	[95]
Ensembl	Comprehensive database of sequences with data mining tools	http://www.ensembl.org/index.html	[94]
2 Drug-related databases			
DrugBank	Drug data and drug-drug interactions	http://www.drugbank.ca/	[92]
Therapeutic Target Database (TTD)	Therapeutic Target Database	http://bidd.nus.edu.sg/group/cjttd/TTD_HOME.asp	[98]
STITCH	Drug molecular interactions	http://stitch.embl.de/	[99]
DCDB	Drug Combination Database	http://www.cls.zju.edu.cn/dcdb/	[174]
Offsides, TwoSides	Drug adverse effects and drug-drug interactions	http://tatonettilab.org/resources/tatonetti-stm.html	[107]
Drugs.com	FDA approved drugs linked to diseases and target proteins/genes	www.drugs.com	
SIDER	Drug adverse effects	http://sideeffects.embl.de/	[105]
3 Protein /protein interaction databases			
DIP	Experimentally and manually validated molecular interactions	http://dip.doe-mbi.ucla.edu/dip/Main.cgi	[108]
HPRD (Human Protein Reference Database)	Experimentally and manually validated molecular interactions	http://www.hprd.org/	[118]
Intact	Manually curated molecular interaction	http://www.ebi.ac.uk/intact/	[114]
MIntAct	Manually curated Integrated database	http://www.ebi.ac.uk/intact	[127]
STRING	Protein/protein interactions as well as connections derived from other databases.	http://string-db.org/	[175]
4 Transcription factor databases			
TRANSFAC	Transcription factors and binding sites	http://www.gene-regulation.com/pub/databases.html	[130]
JASPAR	Transcription Factor Binding Profile Database	http://jaspar.genereg.net/	[176]
DBD	Transcription factor prediction database	http://www.transcriptionfactor.org/	[177]
5 Metabolic pathways			
KEGG	Kyoto Encyclopedia of Genes and Genomes	http://www.genome.jp/kegg/	[165]
Reactome	Curated Pathway Database	http://www.reactome.org/	[166]
MetaCyc	Metabolic Pathway Database	http://metacyc.org/	[178]
HMDB	Human Metabolome Database	http://www.hmdb.ca/	[167]

(Table 3) Contd....

	Contents	URL	V
6 Signal transduction databases			
NetPath	Manually curated signal transduction pathways	http://www.netpath.org/	[179]
Signalink	Manually curated signal transduction pathways	http://signalink.org/	[169]
NDEx	Intergrated Network Database	http://www.ndexbio.org	[170]
7 Mutation databases			
COSMIC	Somatic mutations found human cancer	http://cancer.sanger.ac.uk/cosmic	[180]
OMIM	Disease gene and mutation database of humans	http://www.omim.org/	
8 Literature databases			
PubMed/Medline	PubMed/Medline	http://www.ncbi.nlm.nih.gov/pubmed	[181]
EMBASE	Biomedical and pharmacological bibliographic database	http://store.elsevier.com/embase	
Scopus	Bibliographic database of peer-reviewed literature	http://www.scopus.com/	

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 which 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.

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 $N = |V|$:

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

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 also often used. The Laplacian matrix is an $N \times N$ matrix:

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

The Laplacian matrix can be formulated by the matrices:

$$L = D - A \quad (3)$$

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

$$D = diag(d_1, d_2, \cdots, d_N) \quad (4)$$

where $d_i = \sum_j^N A_{ij}$. If the graph is directed then the d_i elements are the outdegrees of the node. Other possibility to generate the Laplacian matrix is if we normalize the entries of the adjacency matrix in the following way: $\sum_i^N A_{ij} = 1$

$$L_m = I - D^{-1}A \quad (5)$$

Where I denotes the identity matrix with dimension $N \times N$.

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 [184-186] and diffusion [187-190].

5.1. Page Rank

This algorithm is a special case of random walk on data network: a walker starts at a certain data node, 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 it can move to any other nodes with a certain probability ("restart probability"). If the walker is only allowed to move to specific set of nodes or to the neighboring nodes, then this is the PageRank with prior algorithm [185, 186, 191]. If there is prior knowledge available about which nodes are 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 , then we use this information in the random walk in the following way:

$$P(v)^{i+1} = (1-\beta)\left(\sum_{u=1}^{d_{in}(v)} p(u,v)P(v)^{i}\right) + \beta\, pr(v) \tag{6}$$

$P(v)^{i}$ 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 . Other well-known algorithms based on random walks include k-step Markov [191], HITS [192], HITS with Prior [191].

5.2. Diffusion on Graphs

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 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 $x(t)$:

$$\frac{dx(t)}{dt} = -Lx(t) \tag{7}$$

The solution of this differential equation is:

$$x(t) = e^{-Lt} x(0) \tag{8}$$

where $x(0)$ is the initial vector at time 0.

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 [187]. 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:

$$L_{\mu,\gamma} = QD - WAW \tag{9}$$

Where the W and Q matrices are defined as follows:

$$w_{ij} = \begin{cases} \gamma & \text{if } i = j \text{ and } \mathrm{x}_i(0) \neq 0 \\ 1 & \text{if } i = j \text{ and } \mathrm{x}_i(0) = 0 \\ 0 & \text{if } i \neq j \end{cases} \tag{10}$$

$$q_{ij} = \begin{cases} \mu & \text{if } i = j \text{ and } \mathrm{x}_0^i \neq 0 \\ 1 & \text{if } i = j \text{ and } \mathrm{x}_0^i = 0 \\ 0 & \text{if } i \neq j \end{cases} \tag{11}$$

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

The evaluation of an extremely large system of ordinary differential equations could be a challenging task, however by using sparse linear algebra and leveraging the sparseness of a typical data network, the solution could be computed in reasonable time. Instead of 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 by using iterative methods such as Arnoldi algorithm [193-195].

5.3. Background Probabilities

Both PageRank and diffusion methods require the estimation of a threshold value below which the nodes (and their respective edges) are omitted from the network neighborhood. This can be carried out by standard Monte-Carlo simulations in which a large number - say 10 thousand - of iterations are started from randomly selected nodes of the network, and values significantly - say $p \ll 0.05$ - higher than the background are selected as members of the neighborhood.

5.4. Subnetwork Overlap Measures

Subnetwork overlap is a crucial concept of our approach. It indicates a part of a data network where interesting phenomena are expected, such as a drug synergism (in gene/drug networks) or a novel hypothesis (text-gene networks). 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 we use as a target overlap score (TOS). The TOS of two subnetworks net_1 and net_2 is:

$$TOS(net_1, net_2) = \frac{\left|V_{net_1} \cap V_{net_2}\right|}{\left|V_{net_1} \cup V_{net_2}\right|} \tag{12}$$

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) de-

rived from Monte-Carlo simulations, in this case the overlap coefficient will have a statistical interpretation.

Another approach to characterize the neighborhood overlap is to search for interesting subgraphs. For instance, a subnetwork can be considered interesting if it substantially overlaps with a specific metabolic or signaling pathway. In other words, neighborhood overlap can be considered „interesting" if it overlaps with a pathway that has not been considered, or vice versa, a pathway of biological interest. Sophisticated questions can be answered in this manner: if a drug perturbs pathway X, we can systematically test pairwise drug combinations whose overlaps perturb the same pathway but neither participant does in itself. In this way we can look for drug combinations that potentially replace a given single pharmacon.

5.5. Prediction of Drug Combinations Via a Target Overlap Score

5.5.1. Principle

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 [196-198]. Multi-target drugs and drug combinations have been proposed as a general strategy to change this trend [199, 200]. Combinations of approved pharmacons are an especially attractive solution since they have higher therapeutic success, less toxicity and lower development costs as compared to single pharmacons [201]. There are a large number of methods and protocols for identifying novel combinations, and as a result, the number of approved drug combinations is on the increase, even though most of the approved combinations were established by experience and intuition [202, 203].

Given the fast growth of drug-related databases, network neighborhood analysis seems a promising avenue for predicting drug combinations for experimental testing. The underlying assumption is 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 point to unexpected synergies at target genes that are not in the immediate vicinity of the original drug targets. Pinpointing synergies is conceptually different from the traditional classification of drug interactions as positive or negative. Namely, the traditional view pharmacology concentrates on a hypothetical single target, a receptor, and two drugs can mutually increase or decrease the effect on this receptor. In the clinical practice, approved drug combinations (positive interactions) are those that exert some positive therapeutic effect, while the terms "drug interactions" or "interacting drugs" refer to deleterious side effect emerging when two drugs are administered together - i.e. the clinical view defines positive or negative with respect to the patient. The network view, on the other hand, defines synergy as the emergence of additional targets. In this view, the positivity or negativity of the effect is not included, only the fact that the joint effect of two pharmacons will reach new regions of the network. Cancer chemotherapy represents a special case, chemotherapeutic agents target all dividing cells in the body i.e. their effect is negative in the biochemical sense, while positive for the patient since tumor cells that proliferate fast will be selectively damaged. In this section we show that the network neighborhood analysis can pinpoint useful combinations for cancer therapy. For the analysis we defined the network as a combination of protein-protein interactions included in the STRING database, and drug-protein interactions included in the STICH, Drugbank TTD and JBioWH database. The generation of such a dataset is schematically shown in Fig. (**1**). In addition to protein-protein associations, the combined network links proteins and pathways targeted by the same drugs.

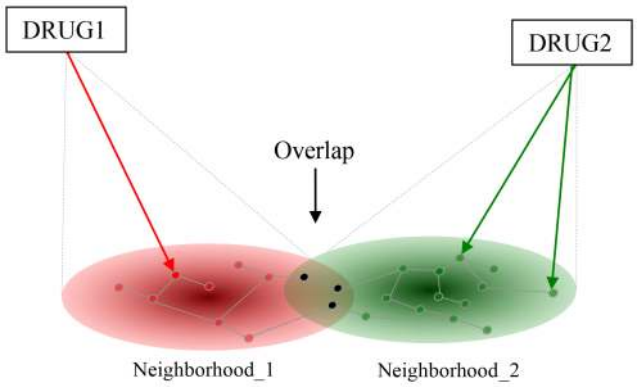

Fig. (1). The network-interaction hypothesis. The effects of two drugs (Drug1, Drug2) reach their imminent targets first (arrows) and the effects will then propagate to their network neighborhoods (subnetworks) indicated in red and green, respectively. Targets in the overlap are affected by both drugs, 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.

The principle of network overlap analysis is shown in Fig. (**1**). We 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 we 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. 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:

$$S_{DCM}(D_i) = e^{-L_{\mu,\gamma}t}x(0) \tag{13}$$

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

$$x_j(0) = \begin{cases} 1 & \text{if protein } j \text{ is drug target} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

The j th element of $S_{DCM}(\mathrm{D}_i)$ measures the disruption effect of D_i on protein j .

We used the parameters $\mu = 0.1$ and á $= 0.005$ throughout this study.

Then the network neighborhood or subnetwork of drug D_i , net_{D_i} consists of the significantly perturbed network elements:

$$net_{D_i} = \left\{ v_i \mid v_i \in V, p_{v_i} < 0.05 \right\} \tag{15}$$

The target overlap score is calculated as in equation 13.

Furthermore, this measure can be generalized to handle non-binary drug combinations. For this purpose we determined the number of nodes that were significantly perturbed by at least two drugs, divided by the size of the affected subnetwork.

$$TOS(net_1, net_2, \cdots net_M) = \frac{\left| \bigcup_{i,j=1\cdots M; i\neq j} V_{net_i} \cap V_{net_j} \right|}{\left| \bigcup_{i=1\cdots M; i\neq j} V_{net_i} \right|} \quad (16)$$

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 prescreened and drugs with identical targets should be excluded from the analysis, either a priori, or by omitting their combination from the results.

The process of calculation is as follows:

Calculating the perturbation values for each drug.

Assessing the significance of each protein for each drug.

Defining the subnetwork of the drugs.

Computing the TOS values for the drug combinations.

5.5.2. TOS is Correlated with the Strength of Both Beneficial and Deleterious Drug Combinations

For the evaluation we chose a simple ranking test, i.e. we 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 [204].

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 we used the STRING interaction network [4] and the first question we asked was whether or not the evaluation system fulfils these fundamental criteria. For this purpose we used the database of FDA-approved drugs [96] 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 (Fig. **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 network, the randomly chosen FDA-approved drug pairs indeed behave as random. We 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 we found 271 such drug pairs. Also, drugs having closely identical chemical structures are likely to affect similar targets. We found 179 such drug pairs but only 8 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. So, for the statistical evaluation described below we left out these drug pairs from both datasets.

Next, we 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 [205]. 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. In the selection we 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 10323 strongly, 92958 moderately and 17193 weakly interacting drugs from the database, denoted as sets A, B and C, respectively (Table **4**). The results show that the interacting drug pairs show remarkably higher AUC values than the randomly selected drug pairs, moreover these values qualitatively follow the strength of the interaction (Fig. **3**). Namely, strongly interacting drug pairs show substantially higher AUC values than the moderately interacting ones *etc*.

We 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) [174] as well as in the Therapeutic Target Database TTD [207], along with the specific mechanism of their interaction. Using the same filtering criteria we selected 293 combinations (dataset D, Table **4**). The results in Fig. (**2**), right show that therapeutic drug combinations yield AUC values substantially different from the random combinations.

Next we carried out the same comparisons for cancer related drugs. In this case the datasets were naturally smaller, we found 817 strongly, 5700 moderately and 241 weakly interacting drugs from the database, and denoted them as sets E, F and G, respectively

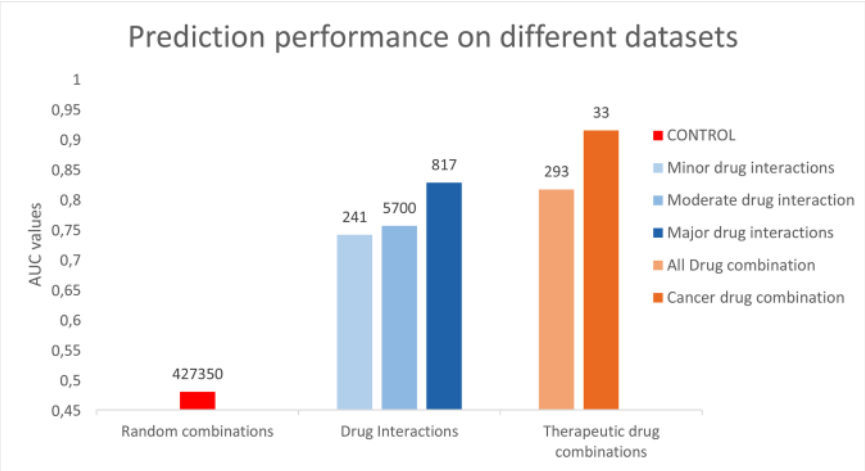

Fig. (2). Ranking performance of the TOS score on known drug interactions and therapeutic combinations. The ranking performance was measured via ROC analysis The standard deviation of AUC values (not shown) are between 0.0001 and 0.006 for the different datasets. Note that the tendencies of drug combination groups are the same between cancer-related and not cancer-related drugs. Also, combinations of drugs with identical targets or with similar chemical structures give high TOS scores. These combinations were left out from the statistics of the other groups so they do not influence the AUC values of the other groups.

Table 4. Datasets.

	Dataset	Original size	Size after filtering[1]	Data source
I. Datasets of all drugs				
Detrimental drug interactions[2]				
Severe	A	21831	10323	Drugs.com
Moderate	B	112976	92958	Drugs.com
Minor	C	13143	17973	Drugs.com
Beneficial drug interactions[3, 4]	D	429	293	DCDB, TTD
II. Cancer-related datasets				
Detrimental drug interactions[2]				
Severe	E	1053	817	Drugs.com
Moderate	F	6857	5700	Drugs.com
Minor	G	273	241	Drugs.com
Beneficial drug interactions[3, 5]	H	55	33	DCDB, TTD
III. Negative datasets used in ROC analysis				
All FDA-approved drugs[6]	I	848253	733542	Drugbank
Random drugs[7]	J	427350	426425	-

[1]We 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. The drugs with no available targets were also discarded; [2]Taken from Drugs.com (November 11, 2013); [3]Taken from the Drug Combination Database (March 8, 2012) and the Therapeutic Target Database (July 23, 2012); [4]All approved drug combinations were included; [5]All approved drug combinations that are used in cancer treatment. [6]We made all possible binary combinations of FDA-approved drugs (taken from DrugBank, 12th September of 2012), and then leaved out all pairs that were listed as beneficial or detrimental combinations. [7]We constructed random drugs corresponding to the number of targets of all individual drugs. We generated 25 random drugs for each target count [37]. From this pool we made all the possible binary combinations. In each case, we randomly selected a negative set of the size which was 5 times greater than the positive dataset [206].

(Table **4**). The set of beneficial combinations included 33 combinations specifically suggested for cancer (dataset H, Table **4**). The results presented in Fig. (**3**) shows the same general tendencies as seen in the case of all drug combinations (Fig. **3**). 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 both panels of Fig. (**3**), the beneficial interactions show higher AUC values. We have no ready explanation for this phenomenon, however we speculate that one of the reason could be that therapeutic combinations are usually optimized via careful clinical studies.

5.6.2. 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 [208]. 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 [209]. The list of drugs tested in clinical trials included bevacizumab, capecitabine, carboplatin, cyclophosphamide, docetaxel, doxorubicin, epirubicin, fluorouracil, gemcitabine, ixabepilone, lapatinib, oxaliplatin, paclitaxel, pertuzumab, sunitinib.

First we analyzed 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 score shows substantial correlation with the overall response (OR) (r=0.64; p=0.0028). 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).

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

6. 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 [210]. Computational methods of "gene prioritization" were developed for this purpose [211]. 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

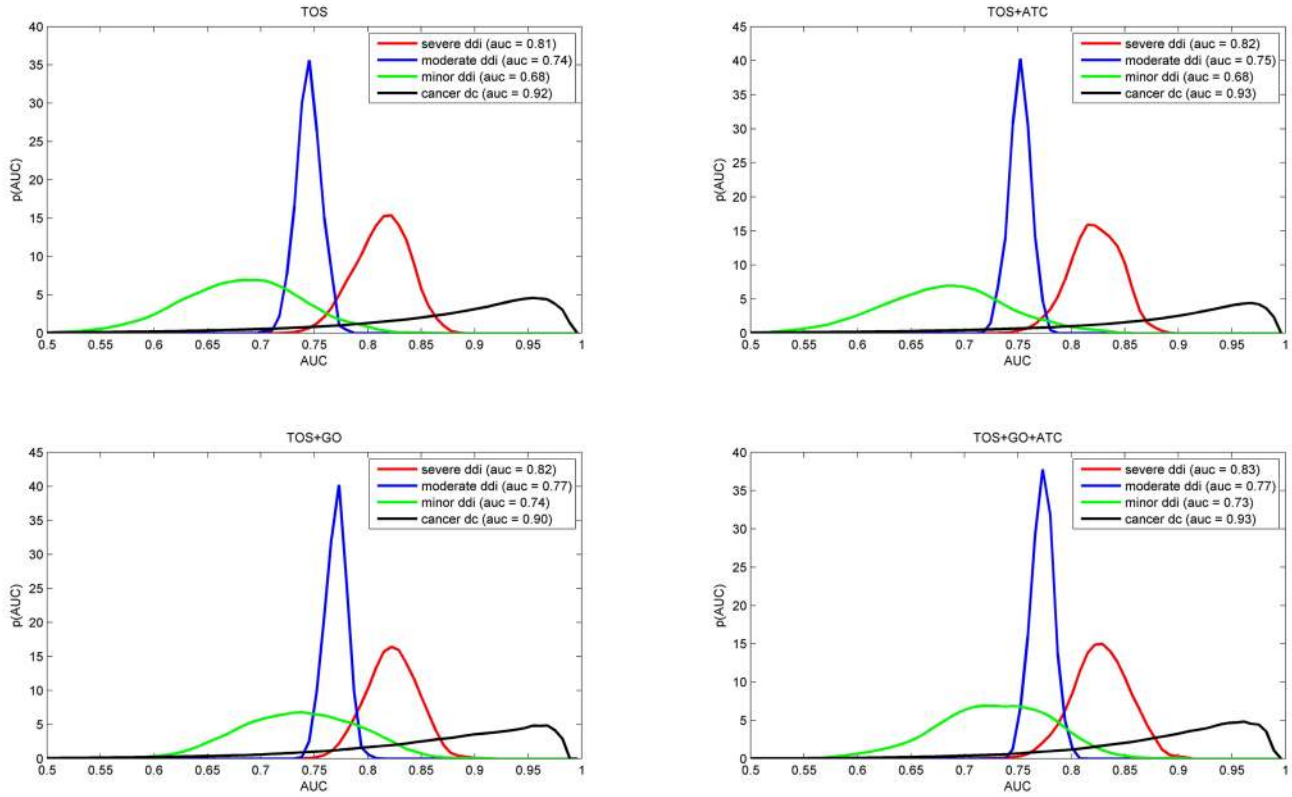

Fig. (3). Performance of combined predictors on different training sets. The short titles TOS, TOS+ATC, TOS+GO or TOS+GO+ATC refer to the combination used. The curves represent the AUC value distribution (as a probability density function) obtained via a kernel density estimation (KDE) approach. The 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.

path), on a propagation algorithm such as the popular PageRank [212] or on diffusion methods [190], 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, 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 [213], such as suggesting disease marker genes from MEDLINE records and ranking (prioritizing) genes based on biomedical literature [214]. Reviews of the earlier work are found in [215] and [216]. More recently, Hristovski and associates combined DNA microarray data and semantic relations extracted from MEDLINE, for generating novel hypotheses [217]. 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 [218].

6.1. 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 Petric et al [219] 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 may 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 as drugs). In this section, we add a new type of link, "co-occurrence", 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 Fig. (**4**). In this section we used ovarian cancer as the test case.

6.2. Constructing a Data-Network with Molecular and Literature-Based Links

The document sets in our experiments were acquired from the MEDLINE database through its PubMed system [181] using the

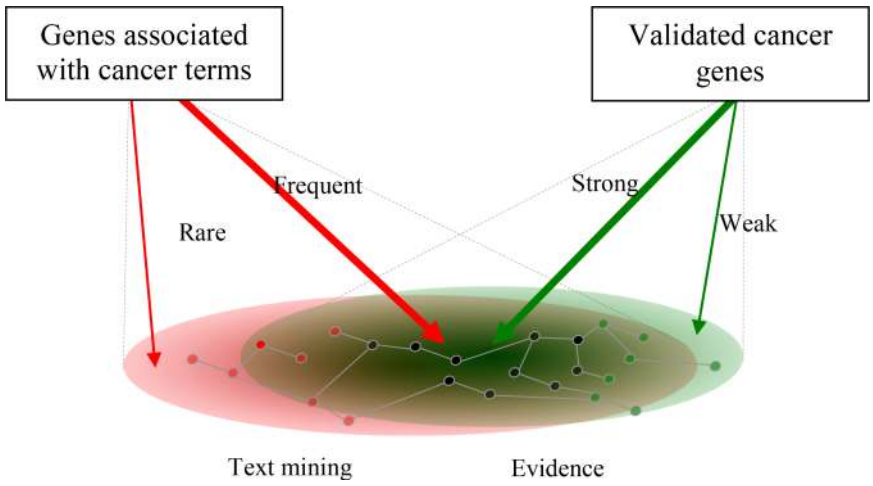

Fig. (4). The principle of biomarker prediction using terms rarely associated with cancer and a set of validated genes. A "hypothesis" is a gene worth to be experimentally tested. Such a gene (network node) is expected to rarely - but appreciably - associate with cancer terms in scientific papers (light 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.

Entrez Programming Utilities [220]. 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 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, we needed a corpus of abstracts published before a certain biological role was confirmed. We chose ovarian cancer as the model disease and used a recently published list of 37 ovarian cancer biomarker (OC biomarkers) genes [221] (Table **5**) as test cases. We 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, we selected the year 2007 as a separating line. A total of 10 OC biomarkers have been proposed after this date.

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 tumour of ovary" OR "ovarian malignancy" OR "ovarian carcinoma"). This search resulted in 4,878 abstracts published before the year 2007. We defined this set as the OC biomarker test corpus. Separately, 26,979 abstracts about the known OC biomarker genes [221] (Table **5**) published up until May 14th 2012 were obtained and these formed the OC biomarker prediction corpus. We used the HGNC gene symbols, names and their synonyms (downloaded on December 23rd 2011). Such HGNC nomenclature was then applied to the terms that we automatically extracted from collections of MEDLINE abstracts.

6.3. 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. We 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) [278, 279] and b) propagation-based methods that were implemented on the STRING database [280], as briefly described in section 5.1 and 5.2.

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

$$pr_i = \begin{cases} \frac{1}{M} & \text{if protein i is a known OC biomarker} \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

Where M is the number of validated biomarkers.

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

$$x_i(0) = \begin{cases} 1 & \text{if protein i is a known OC biomarker} \\ 0 & \text{otherwise} \end{cases} \quad (18)$$

6.4. Testing the Methods on the Rediscovery of Known OC Biomarker Genes

We sought 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. We considered genes suggested as biomarkers as those that co-occurred with the term "marker" or "biomarker" in MEDLINE abstracts. We used MEDLINE abstracts, MeSH and HUGO terms published before 2007 and used standard propagation algorithms (PageRank or diffusion kernel methods [190, 212]) 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 we 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 **5**, 27 are mentioned together with "marker" or "biomarker" in MEDLINE articles published before 2007. The remaining 10 genes (our 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 **6**).

Table 5. List of ovarian cancer biomarker genes published before May, 2012.

	Symbol	Gene		Symbol	Gene		Symbol	Gene
1	CA125	CA 125 [222-225]	14	P16	p16 [226, 227]	26	BIRC5	Survivin [228]
2	KRT19	Cytokeratin 19 [229, 230]	15	CDKN1A	p21 [231-233]	27	TERT	hTERT [234]
3	KLK6	Kallikrein 6 [235]	16	CDKN1B	p27 [236-239]	28	EGFR	ERBB1 [240, 241]
4	KLK10	Kallikrein 10 [242]	17	RB1	pRB [243, 244]	29	ERBB2	ERBB2 [245]
5	IL6	Interleukin-6 [246]	18	E2F1	E2F1 [247]	30	MET	c-Met [248]
6	IL7	Interleukin-7 [249]	19	E2F2	E2F2 [250]	31	MMP2	MMP-2 [251]
7	IFNG	γ-interferon [252]	20	E2F4	E2F4 [250]	32	MMP9	MMP-9 [253]
8	FAS	sFas [254, 255]	21	TP53	p53 [256, 257]	33	MMP14	MT1-MMP [258]
9	VEGFR	VEGFR [259]	22	TP73	p73 [260]	34	WFDC2 (HE4)	Epididymis protein 4 [261-263]
10	CCND1	Cyclin D1 [231, 264]	23	BAX	Bax [265, 266]	35	SERPINB5	Maspin [267]
11	CCND3	Cyclin D3 [268]	24	BCL2L1	Bcl-xl [269]	36	BRCA1	BRCA1 [270]
12	CCNE	Cyclin E [271-274]	25	BIRC2	cIAP [275]	37	ERCC1	ERCC1 [276]
13	P15	p15 [277]						

Table 6. 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

6.5. Prediction of New OC Biomarkers

We 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 [221] (Table **5**) 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 [190, 212]), and also with the gene prioritization servers ToppGene and Endeavour [278, 279]. It was apparent that a number of well-known cancer-related genes appear in the top of these lists (data not shown).

For a better overview, we compared the top of the lists and picked 10 genes that ranked highly in most of the rankings (Table **7**). 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

Table 7. Predicted OC biomarker genes.

Genes predicted as ovarian cancer biomarkers		Validation after May, 2012
Gene symbol	Gene	
RUNX2	Runt-related transcription factor 2	[281, 282]
SOCS3	Suppressor of cytokine signaling 3	NA
BCL6	B-cell lymphoma 6 protein	[283, 284]
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	
E2F6	Transcription factor E2F6	NA
P18INK4C (CDKN2C)	Cyclin-dependent kinase 4 inhibitor C	NA
PAX5	Paired box protein Pax-5	NA

proposed as OC biomarkers and have not been mentioned in literature sources together with ovarian cancer. These may represent genetic markers upon which hypotheses can be formulated in relation to ovarian cancer.

7. DISCUSSION AND CONCLUSION

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 extending the analysis of hidden, implicit links which is the essence of new discoveries What is not trivial is how we can design a network in which associations will be useful, in other terms, we can answer practical questions. The suggestion put forward in this chapter is that we construct a data network dedicated for a given purpose. Namely, if we want to query associations between diseases, drugs and drug target, we construct a network consisting of these items, by combining, say drug databases (STITCH [101], DrugBank [97], TTD [207], DCDB [103]), interaction databases (STRING [4], IntAct [127]), disease databases (OMIM [285]) and various resources such as ontologies [81, 82] and manually curated datasets. Or, 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, we consider our approaches as pilot studies into two seemingly unrelated directions, the prioritization of drug combinations, and prediction of potential biomarkers.

In drug combinations we showed that molecular interaction data can successfully predict known combinations of chemotherapeutic agents used to treat breast cancer. In particular, 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. In biomarker prediction we showed that novel biomarkers can be prioritized using a network built from text mining data as well as ovary cancer data. In particular, we found that new biomarkers discovered in a given period of time are correlated with genes sporadically emerging in the oncological literature of the previous years. 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, we considered the prediction of a drug combinations 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. 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. Seen the efforts invested into biomedical ontologies, we can trust that this development will broaden the scope of the network analysis technique proposed here.

DISCLOSURE

Part of this article has been previously published in PLoS ONE 10(6): e0129267. doi:10.1371/journal.pone.0129267.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Pavlopoulou A, Spandidos DA, Michalopoulos I. Human cancer databases (Review). Oncol Rep 2015; 33: 3-18.

[2] Ligeti B, Pénzváltó Z, Vera R, Győrffy B, Pongor S. A network-based target overlap score for characterizing drug combinations: high correlation with cancer clinical trial results. PloS One 2015; 10: e0129267.

[3] Petric I, Ligeti B, Gyorffy B, Pongor S. Biomedical hypothesis generation by text mining and gene prioritization. Protein Pept Lett 2014; 21: 847-57.
[4] Franceschini A, Szklarczyk D, Frankild S, *et al*. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 2013; 41: D808-15.
[5] Page R, Takimoto C. Principles of chemotherapy. In: Pazdur R, Coia LR, Hoskins WJ. Eds. Cancer Management: A Multidisciplinary Approach Medical, Surgical & Radiation Oncology. Editors: LD Wagman. New York: PRR 2004: 21-38.
[6] Hirsch J. An anniversary for cancer chemotherapy. JAMA 2006; 296: 1518-20.
[7] Kelland L. The resurgence of platinum-based cancer chemotherapy. Nat Rev Cancer 2007; 7: 573-84.
[8] Penzvalto Z, Lanczky A, Lenart J, *et al*. MEK1 is associated with carboplatin resistance and is a prognostic biomarker in epithelial ovarian cancer. BMC Cancer 2014; 14: 837.
[9] Penzvalto Z, Surowiak P, Gyorffy B. Biomarkers for systemic therapy in ovarian cancer. Curr Cancer Drug Targets 2014; 14: 259-73.
[10] Minotti G, Menna P, Salvatorelli E, Cairo G, Gianni L. Anthracyclines: molecular advances and pharmacologic developments in antitumor activity and cardiotoxicity. Pharmacol Rev 2004; 56: 185-229.
[11] Jolivet J, Cowan KH, Curt GA, Clendeninn NJ, Chabner BA. The pharmacology and clinical use of methotrexate. N Engl J Med 1983; 309: 1094-104.
[12] Perez EA. Microtubule inhibitors: Differentiating tubulin-inhibiting agents based on mechanisms of action, clinical activity, and resistance. Mol Cancer Ther 2009; 8: 2086-95.
[13] Liu YQ, Li WQ, Morris-Natschke SL, *et al*. Perspectives on biologically active camptothecin derivatives. Med Res Rev 2015; 35: 753-89.
[14] Gatti L, Zunino F. Overview of tumor cell chemoresistance mechanisms. Methods Mol Med 2005; 111: 127-48.
[15] Gyorffy B, Surowiak P, Kiesslich O, *et al.* Gene expression profiling of 30 cancer cell lines predicts resistance towards 11 anticancer drugs at clinically achieved concentrations. Int J Cancer 2006; 118: 1699-712.
[16] Johnstone RW, Ruefli AA, Lowe SW. Apoptosis: a link between cancer genetics and chemotherapy. Cell 2002; 108: 153-64.
[17] Zunino F, Perego P, Pilotti S, Pratesi G, Supino R, Arcamone F. Role of apoptotic response in cellular resistance to cytotoxic agents. Pharmacol Ther 1997; 76: 177-85.
[18] Stavrovskaya AA. Cellular mechanisms of multidrug resistance of tumor cells. Biochemistry (Mosc) 2000; 65: 95-106.
[19] Chaney SG, Sancar A. DNA repair: enzymatic mechanisms and relevance to drug response. J Natl Cancer Inst 1996; 88: 1346-60.
[20] Munkacsy G, Abdul-Ghani R, Mihaly Z, *et al*. PSMB7 is associated with anthracycline resistance and is a prognostic biomarker in breast cancer. Br J Cancer 2010; 102: 361-8.
[21] Evans WE, Hon YY, Bomgaars L, *et al*. Preponderance of thiopurine S-methyltransferase deficiency and heterozygosity among patients intolerant to mercaptopurine or azathioprine. J Clin Oncol 2001; 19: 2293-301.
[22] Ulrich CM, Robien K, McLeod HL. Cancer pharmacogenetics: polymorphisms, pathways and beyond. Nat Rev Cancer 2003; 3: 912-20.
[23] Ekhart C, Rodenhuis S, Smits PH, Beijnen JH, Huitema AD. An overview of the relations between polymorphisms in drug metabolising enzymes and drug transporters and survival after cancer drug treatment. Cancer Treat Rev 2009; 35: 18-31.
[24] Tegze B, Szallasi Z, Haltrich I, *et al.* Parallel evolution under chemotherapy pressure in 29 breast cancer cell lines results in dissimilar mechanisms of resistance. PLoS One 2012; 7: e30804.
[25] Ando M, Saka H, Ando Y, 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 2005; 55: 552-8.
[26] DeSantis CE, Fedewa SA, Goding Sauer A, Kramer JL, Smith RA, Jemal A. Breast cancer statistics 2015: Convergence of incidence rates between black and white women. CA Cancer J Clin 2016; 66: 31-42.
[27] Perou CM, Sorlie T, Eisen MB, *et al*. Molecular portraits of human breast tumours. Nature 2000; 406: 747-52.
[28] Hanahan D, Weinberg Robert A. Hallmarks of cancer: the next generation. Cell 2011; 144: 646-74.
[29] Chen Y, McGee J, Chen X, *et al*. Identification of druggable cancer driver genes amplified across TCGA datasets. PLoS One 2014; 9: e98293.
[30] Urruticoechea A, Alemany R, Balart J, Villanueva A, Vinals F, Capella G. Recent advances in cancer therapy: an overview. Curr Pharm Des 2010; 16: 3-10.
[31] Saglio G, Kim DW, Issaragrisil S, *et al*. Nilotinib versus imatinib for newly diagnosed chronic myeloid leukemia. N Engl J Med 2010; 362: 2251-9.
[32] Druker BJ, Guilhot F, O'Brien SG, *et al*. Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. N Engl J Med 2006; 355: 2408-17.
[33] Li F, Zhao C, Wang L. Molecular-targeted agents combination therapy for cancer: developments and potentials. Int J Cancer 2014; 134: 1257-69.
[34] Kummar S, Gutierrez M, Doroshow JH, Murgo AJ. Drug development in oncology: classical cytotoxics and molecularly targeted agents. Br J Clin Pharmacol 2006; 62: 15-26.
[35] Park SR, Davis M, Doroshow JH, Kummar S. Safety and feasibility of targeted agent combinations in solid tumours. Nat Rev Clin Oncol 2013; 10: 154-68.
[36] Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 1987; 235: 177-82.
[37] Dawood S, Broglio K, Buzdar AU, Hortobagyi GN, Giordano SH. Prognosis of women with metastatic breast cancer by HER2 status and trastuzumab treatment: an institutional-based review. J Clin Oncol 2010; 28: 92-8.
[38] Awada A, Bozovic-Spasojevic I, Chow L. New therapies in HER2-positive breast cancer: a major step towards a cure of the disease? Cancer Treat Rev 2012; 38: 494-504.
[39] Vogel CL, Cobleigh MA, Tripathy D, *et al*. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. J Clin Oncol 2002; 20: 719-26.
[40] Montemurro F, Prat A, Rossi V, *et al*. Potential biomarkers of long-term benefit from single-agent trastuzumab or lapatinib in HER2-positive metastatic breast cancer. Mol Oncol 2014; 8: 20-6.
[41] Graus-Porta D, Beerli RR, Daly JM, Hynes NE. ErbB-2, the preferred heterodimerization partner of all ErbB receptors, is a mediator of lateral signaling. EMBO J 1997; 16: 1647-55.
[42] Penzvalto Z, Tegze B, Szasz AM, *et al*. Identifying resistance mechanisms against five tyrosine kinase inhibitors targeting the ERBB/RAS pathway in 45 cancer cell lines. PLoS One 2013; 8: e59503.
[43] Menyhart O, Santarpia L, Gyorffy B. A comprehensive outline of trastuzumab resistance biomarkers in HER2 overexpressing breast cancer. Curr Cancer Drug Targets 2015; 15: 665-83.
[44] Dang C, Iyengar N, Datko F, *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 2015; 33: 442-7.
[45] Baselga J, Cortés J, Kim S-B, *et al*. Pertuzumab plus trastuzumab plus docetaxel for metastatic breast cancer. N Engl J Med 2012; 366: 109-19.
[46] Dhillon S. Trastuzumab emtansine: a review of its use in patients with HER2-positive advanced breast cancer previously treated with trastuzumab-based therapy. Drugs 2014; 74: 675-86.
[47] Geyer CE, Forster J, Lindquist D, *et al*. Lapatinib plus capecitabine for HER2-positive advanced breast cancer. N Engl J Med 2006; 355: 2733-43.
[48] Ferlay J, Soerjomataram I, Dikshit R, *et al*. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015; 136: E359-86.
[49] Kurman RJ, Shih Ie M. The origin and pathogenesis of epithelial ovarian cancer: a proposed unifying theory. Am J Surg Pathol 2010; 34: 433-43.
[50] Lee Y, Miron A, Drapkin R, *et al*. A candidate precursor to serous carcinoma that originates in the distal fallopian tube. J Pathol 2007; 211: 26-35.
[51] Integrated genomic analyses of ovarian carcinoma. Nature 2011; 474: 609-15.

[52] Lim D, Oliva E. Precursors and pathogenesis of ovarian carcinoma. Pathology 2013; 45: 229-42.
[53] Braicu EI, Sehouli J, Richter R, Pietzner K, Denkert C, Fotopoulou C. 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 2011; 105: 1818-24.
[54] du Bois A, Reuss A, Pujade-Lauraine E, Harter P, Ray-Coquard I, Pfisterer J. 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 2009; 115: 1234-44.
[55] Ozols RF, Bundy BN, Greer BE, *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. J Clin Oncol 2003; 21: 3194-200.
[56] Neijt JP, Engelholm SA, Tuxen MK, *et al*. Exploratory phase III study of paclitaxel and cisplatin versus paclitaxel and carboplatin in advanced ovarian cancer. J Clin Oncol 2000; 18: 3084-92.
[57] Pignata S, Scambia G, Ferrandina G, *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. J Clin Oncol 2011; 29: 3628-35.
[58] Vasey PA, Jayson GC, Gordon A, *et al*. Phase III randomized trial of docetaxel-carboplatin versus paclitaxel-carboplatin as first-line chemotherapy for ovarian carcinoma. J Natl Cancer Institute 2004; 96: 1682-91.
[59] Vaughan S, Coward JI, Bast RC, Jr., *et al*. Rethinking ovarian cancer: recommendations for improving outcomes. Nat Rev Cancer 2011; 11: 719-25.
[60] Bowtell DD, Bohm S, Ahmed AA, *et al*. Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer. Nat Rev Cancer 2015; 15: 668-79.
[61] Shimada M, Kigawa J, Ohishi Y, *et al*. Clinicopathological characteristics of mucinous adenocarcinoma of the ovary. Gynecol Oncol 2009; 113: 331-4.
[62] Mok SC, Bell DA, Knapp RC, *et al*. Mutation of K-ras protooncogene in human ovarian epithelial tumors of borderline malignancy. Cancer Res 1993; 53: 1489-92.
[63] Brown J, Frumovitz M. Mucinous tumors of the ovary: current thoughts on diagnosis and management. Curr Oncol Rep 2014; 16: 389.
[64] Sato S, Itamochi H, Kigawa J, *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 2009; 100: 546-51.
[65] Takano M, Kikuchi Y, Yaegashi N, *et al*. Clear cell carcinoma of the ovary: a retrospective multicentre experience of 254 patients with complete surgical staging. Br J Cancer 2006; 94: 1369-74.
[66] Kuo K-T, Mao T-L, Jones S, *et al*. Frequent activating mutations of PIK3CA in Ovarian clear cell carcinoma. Am J Pathol 2009; 174: 1597-601.
[67] Jones S, Wang TL, Shih Ie M, *et al*. Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. Science 2010; 330: 228-31.
[68] Anglesio MS, George J, Kulbe H, *et al*. IL6-STAT3-HIF signaling and therapeutic response to the angiogenesis inhibitor sunitinib in ovarian clear cell cancer. Clin Cancer Res 2011; 17: 2538-48.
[69] Konecny GE, Wang C, Hamidi H, *et al*. Prognostic and therapeutic relevance of molecular subtypes in high-grade serous ovarian cancer. J Nat Cancer Inst 2014; 106.
[70] Ahmed AA, Etemadmoghadam D, Temple J, *et al*. Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary. J Pathol 2010; 221: 49-56.
[71] Patch AM, Christie EL, Etemadmoghadam D, *et al*. Whole-genome characterization of chemoresistant ovarian cancer. Nature 2015; 521: 489-94.
[72] Kaye SB. Progress in the treatment of ovarian cancer-lessons from homologous recombination deficiency-the first 10 years. Ann Oncol 2016; 27 (Suppl 1): i1-i3.
[73] Gyorffy B, Schafer R. Biomarkers downstream of RAS: a search for robust transcriptional targets. Curr Cancer Drug Targets 2010; 10: 858-68.
[74] Sosman JA, Kim KB, Schuchter L, *et al*. Survival in BRAF V600-mutant advanced melanoma treated with vemurafenib. N Engl J Med 2012; 366: 707-14.
[75] Shaw AT, Kim D-W, Nakagawa K, *et al*. Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. N Engl J Med 2013; 368: 2385-94.
[76] Slamon DJ, Leyland-Jones B, Shak S, *et al*. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med 2001; 344: 783-92.
[77] Marty M, Cognetti F, Maraninchi D, *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 2005; 23: 4265-74.
[78] Brufsky A. Trastuzumab-based therapy for patients with HER2-positive breast cancer: from early scientific development to foundation of care. Am J Clin Oncol 2010; 33: 186-95.
[79] Sharma P, Allison James P. Immune checkpoint targeting in cancer therapy: Toward combination strategies with curative potential. Cell; 161: 205-14.
[80] Maravić G, Bujnicki JM, Feder M, Pongor S, Flögel M. 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 2003; 31: 4941-9.
[81] Kibbe WA, Arze C, Felix V, *et al*. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res 2015; 43: D1071-8.
[82] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004; 32: D267-70.
[83] Burge S, Attwood TK, Bateman A, *et al*. Biocurators and biocuration: surveying the 21st century challenges. Database 2012; 2012: bar059.
[84] Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol 2007; 372: 774-97.
[85] Janin J, Chothia C. The structure of protein-protein recognition sites. J Biol Chem 1990; 265.
[86] Vinogradova O, Qin J. NMR as a unique tool in assessment and complex determination of weak protein-protein interactions. Top Curr Chem 2012; 236: 35-45.
[87] Wand AJ, Englander SW. Protein complexes studied by NMR spectroscopy. Curr Opin Biotechnol 1996; 7: 403-8.
[88] Phizicky EM, Fields S. Protein-protein interactions: methods for detection and analysis. Microbiol Rev 1995; 59: 94-123.
[89] Lu S, Deng P, Liu X, *et al*. Solution structure of the major α-amylase inhibitor of the crop plant amaranth. J Biol Chem 1999; 274: 20473-8.
[90] Fields S, Song OK. A novel genetic system to detect protein protein interactions. Nature 1989; 340(6230): 245-6.
[91] Xia Y, Yu H, Jansen R, *et al*. Analyzing cellular biochemistry in terms of molecular networks. Annu Rev Biochem 2004; 73: 1051-87.
[92] Wishart DS, Knox C, Guo AC, *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 2006; 34: D668-72.
[93] Consortium U. UniProt: a hub for protein information. Nucleic Acids Res 2014: gku989.
[94] Yates A, Akanni W, Amode MR, *et al*. Ensembl 2016. Nucleic Acids Res 2016; 44: D710-6.
[95] Benson DA, Cavanaugh M, Clark K, *et al*. GenBank. Nucleic Acids Res 2012: gks1195.
[96] Knox C, Law V, Jewison T, *et al*. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res 2011; 39: D1035-41.
[97] Law V, Knox C, Djoumbou Y, *et al*. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 2014; 42: D1091-7.
[98] Chen X, Ji ZL, Chen YZ. TTD: therapeutic target database. Nucleic Acids Res 2002; 30: 412-5.
[99] Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P. STITCH: interaction networks of chemicals and proteins. Nucleic Acids Res 2008; 36: D684-8.
[100] Kuhn M, Szklarczyk D, Franceschini A, von Mering C, Jensen LJ, Bork P. STITCH 3: zooming in on protein-chemical interactions. Nucleic Acids Res 2012; 40: D876-80.
[101] Kuhn M, Szklarczyk D, Pletscher-Frankild S, *et al*. STITCH 4: integration of protein-chemical interactions with user data. Nucleic Acids Res 2014; 42(Database Issue): D401-7.

[102] Liu Y, Hu B, Fu C, Chen X. DCDB: drug combination database. Bioinformatics 2010; 26: 587-8.
[103] Liu Y, Wei Q, Yu G, Gai W, Li Y, Chen X. DCDB 2.0: a major update of the drug combination database. Database 2014; 2014: bau124.
[104] Home F. Orange Book: approved drug products with therapeutic equivalence evaluations. USA: US Food & Drug Administration 2013.
[105] Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 2010; 6: 343.
[106] Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res 2016; 44(Database issue): D1075-9.
[107] Tatonetti NP, Patrick PY, Daneshjou R, Altman RB. Data-driven prediction of drug effects and interactions. Sci Trans Med 2012; 4: 125ra31.
[108] Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP: the database of interacting proteins. Nucleic Acids Res 2000; 28: 289-91.
[109] Bader GD, Hogue CW. BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. Bioinformatics 2000; 16: 465-77.
[110] Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: a Molecular INTeraction database. FEBS Lett 2002; 513: 135-40.
[111] Chatr-Aryamontri A, Ceol A, *et al*. MINT: the Molecular INTeraction database. Nucleic Acids Res 2007; 35: D572-4.
[112] Ceol A, Aryamontri AC, Licata L, *et al*. MINT, the molecular interaction database: 2009 update. Nucleic Acids Res 2010; 38(Database issue): D532-9.
[113] Licata L, Briganti L, Peluso D, *et al*. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 2012; 40: D857-61.
[114] Hermjakob H, Montecchi‐Palazzi L, Lewington C, *et al*. IntAct: an open source molecular interaction database. Nucleic Acids Res 2004; 32: D452-5.
[115] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. Nucleic Acids Res 2004; 32: D449-51.
[116] Xenarios I, Salwinski L, Duan XJ, Higney P, Kim S-M, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 2002; 30: 303-5.
[117] Salwinski L, Eisenberg D. The MiSink Plugin: Cytoscape as a graphical interface to the database of interacting proteins. Bioinformatics 2007; 23: 2193-5.
[118] Peri S, Navarro JD, Amanchy R, *et al*. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 2003; 13: 2363-71.
[119] Kerrien S, Alam-Faruque Y, Aranda B, *et al*. IntAct—open source resource for molecular interaction data. Nucleic Acids Res 2007; 35: D561-65.
[120] Aranda B, Achuthan P, Alam-Faruque Y, *et al*. The IntAct molecular interaction database in 2010. Nucleic Acids Res 2010; 38: D525-31.
[121] Kerrien S, Aranda B, Breuza L, *et al*. The IntAct molecular interaction database in 2012. Nucleic Acids Res 2012; 40(Database issue): D841-6.
[122] Côté R, Reisinger F, Martens L, Barsnes H, Vizcaino JA, Hermjakob H. The ontology lookup service: bigger and better. Nucleic Acids Res 2010; 38: W155-60.
[123] Hermjakob H, Montecchi-Palazzi L, Bader G, *et al*. The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. Nat Biotechnol 2004; 22: 177-83.
[124] Kerrien S, Orchard S, Montecchi-Palazzi L, *et al*. Broadening the horizon-level 2.5 of the HUPO-PSI format for molecular interactions. BMC Biol 2007; 5: 44.
[125] Ashburner M, Ball CA, Blake JA, *et al*. Gene Ontology: tool for the unification of biology. Nat Genet 2000; 25: 25-9.
[126] Coordinators NR. Database resources of the national center for biotechnology information. Nucleic Acids Res 2015; 43: D6.
[127] Orchard S, Ammari M, Aranda B, *et al*. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 2014; 42(Database): D358-63.
[128] Orchard S, Kerrien S, Abbani S, *et al*. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. Nat Methods 2012; 9: 345-50.
[129] Calderone A, Castagnoli L, Cesareni G. Mentha: a resource for browsing integrated protein-interaction networks. Nat Methods 2013; 10: 690-1.
[130] Knüppel R, Dietze P, Lehnberg W, Frech K, Wingender E. TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. J Comput Biol 1994; 1: 191-8.
[131] Weinstein JN, Collisson EA, Mills GB, *et al*. The cancer genome atlas pan-cancer analysis project. Nat Genet 2013; 45: 1113-20.
[132] Pleasance ED, Cheetham RK, Stephens PJ, *et al*. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 2010; 463: 191-6.
[133] Ellis MJ, Gillette M, Carr SA, *et al*. Connecting genomic alterations to cancer biology with proteomics: the NCI clinical proteomic tumor analysis consortium. Cancer Discov 2013; 3: 1108-12.
[134] Zhang B, Wang J, Wang X, *et al*. Proteogenomic characterization of human colon and rectal cancer. Nature 2014; 513: 382-7.
[135] Hudson TJ, Anderson W, Aretz A, *et al*. International network of cancer genome projects. Nature 2010; 464: 993-8.
[136] Shepherd R, Forbes SA, Beare D, *et al*. Data mining using the catalogue of somatic mutations in cancer BioMart. Database 2011; 2011: bar018.
[137] Perez-Llamas C, Gundem G, Lopez-Bigas N. Integrative cancer genomics (IntOGen) in Biomart. Database 2011; 2011: bar039.
[138] Goldman M, Craft B, Swatloski T, *et al*. The UCSC cancer genomics browser: update 2013. Nucleic Acids Res 2013; 41: D949-54.
[139] Zhu J, Sanborn JZ, Benz S, *et al*. The UCSC cancer genomics browser. Nat Methods 2009; 6: 239-40.
[140] An O, Pendino V, D'Antonio M, Ratti E, Gentilini M, Ciccarelli FD. NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. Database 2014; 2014: bau015.
[141] D'Antonio M, Pendino V, Sinha S, Ciccarelli FD. Network of Cancer Genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes. Nucleic Acids Res 2012; 40: D978-83.
[142] Zhang J, Finney RP, Rowe W, *et al*. Systematic analysis of genetic alterations in tumors using Cancer Genome WorkBench (CGWB). Genome Res 2007; 17: 1111-7.
[143] Feichtinger J, McFarlane RJ, Larcombe LD. CancerMA: a web-based tool for automatic meta-analysis of public cancer microarray data. Database 2012; 2012: bas055.
[144] Forbes SA, Bindal N, Bamford S, *et al*. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res 2011; 39(Database issue): D945-50.
[145] Cao Q, Zhou M, Wang X, *et al*. CaSNP: a database for interrogating copy number alterations of cancer genome from SNP array data. Nucleic Acids Res 2011; 39(Database issue) D968-74.
[146] Cheng W-C, Chung I-F, Chen C-Y, *et al*. DriverDB: an exome sequencing database for cancer driver gene identification. Nucleic Acids Res 2014; 42: D1048-54.
[147] Gundem G, Perez-Llamas C, Jene-Sanz A, *et al*. IntOGen: integration and data mining of multidimensional oncogenomic data. Nat Methods 2010; 7: 92-3.
[148] Richardson CJ, Gao Q, Mitsopoulous C, Zvelebil M, Pearl LH, Pearl FM. MoKCa database—mutations of kinases in cancer. Nucleic Acids Res 2009; 37: D824-31.
[149] Hess JL. The cancer genome anatomy project: power tools for cancer biologists: EDITORIAL. Cancer Invest 2003; 21: 325-6.
[150] Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer 2007; 7: 233-45.
[151] Futreal PA, Coin L, Marshall M, *et al*. A census of human cancer genes. Nat Rev Cancer 2004; 4: 177-83.
[152] Krupp M, Itzel T, Maass T, Hildebrandt A, Galle PR, Teufel A. CellLineNavigator: a workbench for cancer cell line analysis. Nucleic Acids Res 2013; 41: D942-8.
[153] Ongenaert M, Van Neste L, De Meyer T, Menschaert G, Bekaert S, Van Criekinge W. PubMeth: a cancer methylation database combining text-mining and expert annotation. Nucleic Acids Res 2008; 36: D842-6.

[154] Wang D, Gu J, Wang T, Ding Z. OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs. Bioinformatics 2014; 30: 2237-8.
[155] Kumar R, Chaudhary K, Gupta S, *et al*. CancerDR: cancer drug resistance database. Sci Rep 2013; 3: 1445.
[156] Wang X, Zhao H, Xu Q, Jin W, Liu C, Zhang H, Huang Z, Zhang X, Zhang Y, Xin D. HPtaa database-potential target genes for clinical diagnosis and immunotherapy of human carcinoma. Nucleic Acids Res 2006; 34: D607-D612.
[157] Ahmed J, Meinel T, Dunkel M, *et al*. CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge. Nucleic Acids Res 2011; 39: D960-7.
[158] Scheinin I, Myllykangas S, Borze I, Böhling T, Knuutila S, Saharinen J. CanGEM: mining gene copy number changes in cancer. Nucleic Acids Res 2008; 36: D830-5.
[159] Weinstein JN, Kohn KW, Grever MR, *et al*. Neural computing in cancer drug development: predicting mechanism of action. Science 1992; 258: 447-51.
[160] Monks A, Scudiero D, Skehan P, *et al*. Feasibility of a high-flux anticancer drug screen using a diverse panel of cultured human tumor cell lines. J Nat Cancer Inst 1991; 83: 757-66.
[161] Ramana J. RCDB: renal cancer gene database. BMC Res Notes 2012; 5: 246.
[162] Ganzfried BF, Riester M, Haibe-Kains B, *et al*. curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. Database 2013; 2013: bat013.
[163] Cutts RJ, Gadaleta E, Lemoine NR, Chelala C. Using BioMart as a framework to manage and query pancreatic cancer data. Database 2011; 2011: bar024.
[164] Wang L, Xiong Y, Sun Y, Fang Z, Li L, Ji H, Shi T. HLungDB: an integrated database of human lung cancer research. Nucleic Acids Res 2010; 38(Database issue): D665-9.
[165] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000; 28: 27-30.
[166] Croft D, O'Kelly G, Wu G, *et al*. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 2011: 39(Database issue): D691-7.
[167] Wishart DS, Tzur D, Knox C, *et al*. HMDB: the human metabolome database. Nucleic Acids Res 2007; 35: D521-6.
[168] Korcsmáros T, Farkas IJ, Szalay MS, *et al*. Uniformly curated signaling pathways reveal tissue-specific cross-talks and support drug target discovery. Bioinformatics 2010; 26: 2042-50.
[169] Fazekas D, Koltai M, Türei D, *et al*. SignaLink 2-a signaling pathway resource with multi-layered regulatory networks. BMC Syst Biol 2013; 7: 1.
[170] Pratt D, Chen J, Welker D, *et al*. NDEx, the Network Data Exchange. Cell Syst 2015; 1: 302-5.
[171] Wheeler DL, Church DM, Lash AE, *et al*. Database resources of the national center for biotechnology information. Nucleic Acids Res 2001; 29: 11-6.
[172] Kastrin A, Rindflesch TC, Hristovski D. Large-scale structure of a network of co-occurring MeSH terms: statistical analysis of macroscopic properties. PloS One 2014; 9: e102188.
[173] Pruitt KD, Brown GR, Hiatt SM, *et al*. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 2014; 42: D756-63.
[174] Liu Y, Hu B, Fu C, Chen X. DCDB: drug combination database. Bioinformatics 2010; 26: 587-8.
[175] von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res 2003; 31: 258-61.
[176] Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 2004; 32: D91-4.
[177] Kummerfeld SK, Teichmann SA. DBD: a transcription factor prediction database. Nucleic Acids Res 2006; 34: D74-81.
[178] Caspi R, Foerster H, Fulcher CA, *et al*. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 2006; 34: D511-6.
[179] Kandasamy K, Mohan SS, Raju R, *et al*. NetPath: a public resource of curated signal transduction pathways. Genome Biol 2010; 11: 1-9.
[180] Bamford S, Dawson E, Forbes S, *et al*. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer 2004; 91: 355-8.
[181] U.S. National Library of Medicine. PubMed Overview. In: ed.^eds. U.S. National Library of Medicine 2001.
[182] Vera R, Perez-Riverol Y, Perez S, Ligeti B, Kertész-Farkas A, Pongor S. JBioWH: an open-source Java framework for bioinformatics data integration. Database 2013; 2013: bat051.
[183] Bondy JA, Murty USR. Graph theory with applications. London: Macmillan 1976.
[184] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: bringing order to the web. 1999; Available at: ilpubs.standford.edu.8090/422/1/199-66pdf.
[185] Haveliwala TH. Topic-sensitive pagerank. Proceedings of the 11th international conference on World Wide Web. ACM 2002; pp. 517-26.
[186] Jeh G, Widom J. Scaling personalized web search. In: ed.^eds., Proceedings of the 12th international conference on World Wide Web. ACM 2003; pp. 271-279.
[187] Ito T, Shimbo M, Kudo T, Matsumoto Y. Application of kernels to link analysis. Ikoma, Nara Japan: Nara Institute of Science and Technology 2005; pp. 586-92.
[188] Kandola J, Shawe-Taylor J, Cristianini N. On the application of diffusion kernel to text data. Technical report, Neurocolt 2002. NeuroCOLT Technical Report NC-TR-02-122 2002.
[189] Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis. UK: Cambridge University Press 2004.
[190] Kondor RI, Lafferty J. Diffusion kernels on graphs and other discrete input spaces. ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann 2002; pp. 315-22.
[191] White S, Smyth P. Algorithms for estimating relative importance in networks. In: ed.^eds., Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM 2003; pp. 266-75.
[192] Kleinberg JM. Authoritative sources in a hyperlinked environment. J ACM (JACM) 1999; 46: 604-32.
[193] Eiermann M, Ernst OG. A restarted Krylov subspace method for the evaluation of matrix functions. SIAM J Num Anal 2006; 44: 2481-504.
[194] Moler C, Van Loan C. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Rev 2003; 45: 3-49.
[195] Saad Y. Analysis of some Krylov subspace approximations to the matrix exponential operator. SIAM J Num Anal 1992; 29: 209-28.
[196] Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 2008; 4: 682-90.
[197] Imming P, Sinning C, Meyer A. Drugs, their targets and the nature and number of drug targets. Nat Rev Drug Discov 2006; 5: 821-34.
[198] Kitano H. A robustness-based approach to systems-oriented drug design. Nat Rev Drug Discov 2007; 6: 202-10.
[199] Agoston V, Csermely P, Pongor S. Multiple weak hits confuse complex systems: a transcriptional regulatory network as an example. Phys Rev E Stat Nonlin Soft Matter Phys 2005; 71: 051909.
[200] Csermely P, Agoston V, Pongor S. The efficiency of multi-target drugs: the network approach might help drug design. Trends Pharmacol Sci 2005; 26: 178-82.
[201] Lehar J, Krueger AS, Avery W, *et al*. Synergistic drug combinations tend to improve therapeutically relevant selectivity. Nat Biotechnol 2009; 27: 659-66.
[202] Keith CT, Borisy AA, Stockwell BR. Multicomponent therapeutics for networked systems. Nat Rev Drug Discov 2005; 4: 71-8.
[203] Zimmermann GR, Lehar J, Keith CT. Multi-target therapeutics: when the whole is greater than the sum of the parts. Drug Discov Today 2007; 12: 34-42.
[204] Sonego P, Kocsor A, Pongor S. ROC analysis: applications to the classification of biological sequences and 3D structures. Brief Bioinform 2008; 9: 198-209.
[205] Katzung BGMSBTAJ. Basic & clinical pharmacology. New York; London: McGraw-Hill 2012.
[206] Busa-Fekete R, Kertész-Farkas A, Kocsor A, Pongor S. Balanced ROC analysis (BAROC) protocol for the evaluation of protein similarities. J Biochem Biophys Methods 2008; 70: 1210-4.
[207] Zhu F, Shi Z, Qin C, *et al*. Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. Nucleic Acids Res 2012; 40: D1128-36.
[208] Jin G, Zhao H, Zhou X, Wong ST. An enhanced Petri-net model to predict synergistic effects of pairwise drug combinations from gene microarray data. Bioinformatics 2011; 27: i310-6.

[209] Therasse P, Arbuck SG, Eisenhauer EA, *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 2000; 92: 205-16.

[210] Hardy J, Singleton A. Genomewide association studies and human disease. N Engl J Med 2009; 360: 1759-68.

[211] Tranchevent LC, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, Y. M. A guide to web tools to prioritize candidate genes. Briefings Bioinform 2011; 12: 22-32.

[212] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Systems 1998; 30: 107-17.

[213] Krallinger M, Valencia A, Hirschman L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol 2008; 9: S8.

[214] Grimes G, Wen T, Mewissen M, *et al*. PDQ Wizard: automated prioritization and characterization of gene and protein lists using biomedical literature. Bioinformatics 2006; 22: 2055-7.

[215] Erhardt RA-A, Schneider R, Blaschke C. Status of text-mining techniques applied to biomedical text. Drug Discovery Today 2006; 11: 315-325.

[216] Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 2006; 7: 119-29.

[217] Hristovski D, Kastrin A, Peterlin B, Rindflesch T. Combining semantic relations and DNA microarray data for novel hypotheses generation. In: Blaschke C, Shatkay H, Eds. Linking Literature, Information, and Knowledge for Biology. USA: Springer 2010; pp. 53-61.

[218] Frijters R, van Vugt M, Smeets R, van Schaik R, de Vlieg J, Alkema W. Literature mining for the discovery of hidden connections between drugs, genes and diseases. PLoS Comput Biol 2010; 6: e1000943.

[219] Petrič I, Urbančič T, Cestnik B, Macedoni-Lukšič M. Literature mining method RaJoLink for uncovering relations between biomedical concepts. J Biomed Inform 2009; 42: 219-27.

[220] Sayers E, Wheeler D. Building Customized Data Pipelines Using the Entrez Programming Utilities (eUtils). NCBI Short Courses U.S. National Center for Biotechnology Information 2004.

[221] Gyorffy B, Lánczky A, Szállási Z. Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients. Endocr-Relat Cancer 2012; 19: 197-208.

[222] Cooper BC, Sood AK, Davis CS, *et al*. Preoperative CA 125 levels: an independent prognostic factor for epithelial ovarian cancer. Obstet Gynecol 2002; 100: 59-64.

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

[224] Gadducci A, Zola P, Landoni F, *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 1995; 58: 42-7.

[225] Riedinger JM, Wafflart J, Ricolleau G, *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 2006; 17: 1234-8.

[226] Katsaros D, Cho W, Singal R, *et al*. Methylation of tumor suppressor gene p16 and prognosis of epithelial ovarian cancer. Gynecol Oncol 2004; 94: 685-92.

[227] Kommoss S, du Bois A, Ridder R, *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 2007; 96: 306-13.

[228] Sui L, Dong Y, Ohno M, Watanabe Y, Sugimoto K, Tokuda M. Survivin expression and its correlation with cell proliferation and prognosis in epithelial ovarian tumors. Int J Oncol 2002; 21: 315-20.

[229] Gadducci A, Ferdeghini M, Cosio S, Fanucchi A, Cristofani R, Genazzani AR. The clinical relevance of serum CYFRA 21-1 assay in patients with ovarian cancer. Int J Gynecol Cancer 2001; 11: 277-82.

[230] Tempfer C, Hefler L, Heinzl H, et al. CYFRA 21-1 serum levels in women with adnexal masses and inflammatory diseases. Br J Cancer 1998; 78: 1108-12.

[231] Bali A, O'Brien PM, Edwards LS, *et al*. Cyclin D1, p53, and p21Waf1/Cip1 expression is predictive of poor clinical outcome in serous epithelial ovarian cancer. Clin Cancer Res 2004; 10: 5168-77.

[232] Ferrandina G, Stoler A, Fagotti A, *et al*. p21WAF1/CIP1 protein expression in primary ovarian cancer. Int J Oncol 2000; 17: 1231-5.

[233] Plisiecka-Halasa J, Karpinska G, Szymanska T, *et al*. P21WAF1, P27KIP1, TP53 and C-MYC analysis in 204 ovarian carcinomas treated with platinum-based regimens. Ann Oncol 2003; 14: 1078-85.

[234] Brustmann H. Immunohistochemical detection of human telomerase reverse transcriptase (hTERT) and c-kit in serous ovarian carcinoma: a clinicopathologic study. Gynecol Oncol 2005; 98: 396-402.

[235] Diamandis EP, Scorilas A, Fracchioli S, *et al*. Human kallikrein 6 (hK6): a new potential serum biomarker for diagnosis and prognosis of ovarian carcinoma. J Clin Oncol 2003; 21: 1035-43.

[236] Korkolopoulou P, Vassilopoulos I, Konstantinidou AE, *et al*. The combined evaluation of p27Kip1 and Ki-67 expression provides independent information on overall survival of ovarian carcinoma patients. Gynecol Oncol 2002; 85: 404-14.

[237] Masciullo V, Ferrandina G, Pucci B, *et al*. p27Kip1 expression is associated with clinical outcome in advanced epithelial ovarian cancer: multivariate analysis. Clin Cancer Res 2000; 6: 4816-22.

[238] Newcomb EW, Sosnow M, Demopoulos RI, Zeleniuch-Jacquotte A, Sorich J, Speyer JL. Expression of the cell cycle inhibitor p27KIP1 is a new prognostic marker associated with survival in epithelial ovarian tumors. Am J Pathol 1999; 154: 119-25.

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

[240] Skirnisdottir I, Seidal T, Sorbe B. 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 2004; 14: 259-70.

[241] Psyrri A, Kassar M, Yu Z, *et al*. Effect of epidermal growth factor receptor expression level on survival in patients with epithelial ovarian cancer. Clin Cancer Res 2005; 11: 8637-43.

[242] Luo LY, Katsaros D, Scorilas A, *et al*. Prognostic value of human kallikrein 10 expression in epithelial ovarian carcinoma. Clin Cancer Res 2001; 7: 2372-9.

[243] Dong Y, Walsh MD, McGuckin MA, *et al*. Reduced expression of retinoblastoma gene product (pRB) and high expression of p53 are associated with poor prognosis in ovarian cancer. Int J Cancer 1997; 74: 407-15.

[244] Konstantinidou AE, Korkolopoulou P, Vassilopoulos I, *et al*. Reduced retinoblastoma gene protein to Ki-67 ratio is an adverse prognostic indicator for ovarian adenocarcinoma patients. Gynecol Oncol 2003; 88: 369-78.

[245] Lassus H, Leminen A, Vayrynen A, *et al*. ERBB2 amplification is superior to protein expression status in predicting patient outcome in serous ovarian carcinoma. Gynecol Oncol 2004; 92: 31-9.

[246] Scambia G, Testa U, Benedetti Panici P, *et al*. Prognostic significance of interleukin 6 serum levels in patients with ovarian cancer. Br J Cancer 1995; 71: 354-6.

[247] Suh DS, Yoon MS, Choi KU, Kim JY. Significance of E2F-1 overexpression in epithelial ovarian cancer. Int J Gynecol Cancer 2008; 18: 492-8.

[248] Sawada K, Radjabi AR, Shinomiya N, *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 2007; 67: 1670-9.

[249] Lambeck AJ, Crijns AP, Leffers N, *et al*. Serum cytokine profiling as a diagnostic and prognostic tool in ovarian cancer: a potential role for interleukin 7. Clin Cancer Res 2007; 13: 2385-91.

[250] Reimer D, Sadr S, Wiedemair A, *et al*. Clinical relevance of E2F family members in ovarian cancer--an evaluation in a training set of 77 patients. Clin Cancer Res 2007; 13: 144-51.

[251] Torng PL, Mao TL, Chan WY, Huang SC, Lin CT. Prognostic significance of stromal metalloproteinase-2 in ovarian adenocarcinoma and its relation to carcinoma progression. Gynecol Oncol 2004; 92: 559-67.

[252] Marth C, Fiegl H, Zeimet AG, *et al*. Interferon-gamma expression is an independent prognostic factor in ovarian cancer. Am J Obstet Gynecol 2004; 191: 1598-605.

[253] Sillanpaa S, Anttila M, Voutilainen K, *et al*. Prognostic significance of matrix metalloproteinase-9 (MMP-9) in epithelial ovarian cancer. Gynecol Oncol 2007; 104: 296-303.

[254] Hefler L, Mayerhofer K, Nardi A, Reinthaller A, Kainz C, Tempfer C. Serum soluble Fas levels in ovarian cancer. Obstet Gynecol 2000; 96: 65-9.

[255] Konno R, Takano T, Sato S, Yajima A. Serum soluble fas level as a prognostic factor in patients with gynecological malignancies. Clin Cancer Res 2000; 6: 3576-80.

[256] Buttitta F, Marchetti A, Gadducci A, *et al*. p53 alterations are predictive of chemoresistance and aggressiveness in ovarian carcinomas: a molecular and immunohistochemical study. Br J Cancer 1997; 75: 230-5.

[257] Reles A, Wen WH, Schmider A, *et al*. Correlation of p53 mutations with resistance to platinum-based chemotherapy and shortened survival in ovarian cancer. Clin Cancer Res 2001; 7: 2984-97.

[258] Kamat AA, Fletcher M, Gruman LM, *et al*. The clinical relevance of stromal matrix metalloproteinase expression in ovarian cancer. Clin Cancer Res 2006; 12: 1707-14.

[259] Hefler LA, Zeillinger R, Grimm C, *et al*. Preoperative serum vascular endothelial growth factor as a prognostic parameter in ovarian cancer. Gynecol Oncol 2006; 103: 512-7.

[260] Becker K, Pancoska P, Concin N, *et al*. Patterns of p73 N-terminal isoform expression and p53 status have prognostic value in gynecological cancers. Int J Oncol 2006; 29: 889-902.

[261] Huhtinen K, Suvitie P, Hiissa J, *et al*. Serum HE4 concentration differentiates malignant ovarian tumours from ovarian endometriotic cysts. Br J Cancer 2009; 100: 1315-9.

[262] Moore RG, Jabre-Raughley M, Brown AK, *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 2010; 203: 228 e1-6.

[263] Moore RG, McMeekin DS, Brown AK, *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 2009; 112: 40-6.

[264] Barbieri F, Lorenzi P, Ragni N, *et al*. Overexpression of cyclin D1 is associated with poor survival in epithelial ovarian cancer. Oncology 2004; 66: 310-5.

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

[266] Tai YT, Lee S, Niloff E, Weisman C, Strobel T, Cannistra SA. BAX protein expression and clinical outcome in epithelial ovarian cancer. J Clin Oncol 1998; 16: 2583-90.

[267] Secord AA, Lee PS, Darcy KM, *et al*. Maspin expression in epithelial ovarian cancer and associations with poor prognosis: a Gynecologic Oncology Group study. Gynecol Oncol 2006; 101: 390-7.

[268] Levidou G, Korkolopoulou P, Thymara I, *et al*. Expression and prognostic significance of cyclin D3 in ovarian adenocarcinomas. Int J Gynecol Pathol 2007; 26: 410-7.

[269] Materna V, Surowiak P, Markwitz E, *et al*. Expression of factors involved in regulation of DNA mismatch repair- and apoptosis pathways in ovarian cancer patients. Oncol Rep 2007; 17: 505-16.

[270] Thrall M, Gallion HH, Kryscio R, Kapali M, Armstrong DK, DeLoia JA. BRCA1 expression in a large series of sporadic ovarian carcinomas: a Gynecologic Oncology Group study. Int J Gynecol Cancer 2006; 16 (Suppl 1): 166-71.

[271] Bedrosian I, Lee C, Tucker SL, Palla SL, Lu K, Keyomarsi K. Cyclin E-associated kinase activity predicts response to platinum-based chemotherapy. Clin Cancer Res 2007; 13: 4800-6.

[272] Farley J, Smith LM, Darcy KM, *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 2003; 63: 1235-41.

[273] Rosen DG, Yang G, Deavers MT, *et al*. Cyclin E expression is correlated with tumor progression and predicts a poor prognosis in patients with ovarian carcinoma. Cancer 2006; 106: 1925-32.

[274] Sui L, Dong Y, Ohno M, *et al*. Implication of malignancy and prognosis of p27(kip1), Cyclin E, and Cdk2 expression in epithelial ovarian tumors. Gynecol Oncol 2001; 83: 56-63.

[275] Psyrri A, Yu Z, Bamias A, *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 2006; 15: 1179-83.

[276] Darcy KM, Tian C, Reed E. 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 2007; 67: 4474-81.

[277] Kudoh K, Ichikawa Y, Yoshida S, *et al*. Inactivation of p16/CDKN2 and p15/MTS2 is associated with prognosis and response to chemotherapy in ovarian cancer. Int J Cancer 2002; 99: 579-82.

[278] Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res 2009; 37: W305-11.

[279] Tranchevent LC, Barriot R, Yu S, *et al*. ENDEAVOUR update: a web resource for gene prioritization in multiple species. Nucleic Acids Res 2008; 36: W377-84.

[280] Szklarczyk D, Franceschini A, Kuhn M, *et al*. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 2011; 39: 561-8.

[281] Li W, Xu S, Lin S, Zhao W. Overexpression of runt-related transcription factor-2 is associated with advanced tumor progression and poor prognosis in epithelial ovarian cancer. BioMed Res Int 2012; 2012.

[282] Li W, Liu Z, Chen L, Zhou L, Yao Y. MicroRNA-23b is an independent prognostic marker and suppresses ovarian cancer progression by targeting runt-related transcription factor-2. FEBS Lett 2014; 588: 1608-15.

[283] Wang YQ, Xu MD, Weng WW, Wei P, Yang YS, Du X. BCL6 is a negative prognostic factor and exhibits pro-oncogenic activity in ovarian cancer. Am J Cancer Res 2015; 5: 255.

[284] Shan W, Li J, Bai Y, Lu X. miR-339-5p inhibits migration and invasion in ovarian cancer cell lines by targeting NACC1 and BCL6. Tumor Biol 2016; 37: 5203-5211.

[285] Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 2005; 33: D514-7.