BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 4. 357–368
Advance Access published on 28 June 2011 doi:10.1093/bib/bbr005

Literature mining, ontologies and information visualization for drug repurposing

Christos Andronis, Anuj Sharma, Vassilis Virvilis, Spyros Deftereos and Aris Persidis

Submitted: 17th November 2010; Received (in revised form): 2nd February 2011

Abstract

The immense growth of MEDLINE coupled with the realization that a vast amount of biomedical knowledge is recorded in free-text format, has led to the appearance of a large number of literature mining techniques aiming to extract biomedical terms and their inter-relations from the scientific literature. Ontologies have been extensively utilized in the biomedical domain either as controlled vocabularies or to provide the framework for mapping relations between concepts in biology and medicine. Literature-based approaches and ontologies have been used in the past for the purpose of hypothesis generation in connection with drug discovery. Here, we review the application of literature mining and ontology modeling and traversal to the area of drug repurposing (DR). In recent years, DR has emerged as a noteworthy alternative to the traditional drug development process, in response to the decreased productivity of the biopharmaceutical industry. Thus, systematic approaches to DR have been developed, involving a variety of *in silico*, genomic and high-throughput screening technologies. Attempts to integrate literature mining with other types of data arising from the use of these technologies as well as visualization tools assisting in the discovery of novel associations between existing drugs and new indications will also be presented.

Keywords: *literature mining; ontologies; information visualization; drug repurposing*

INTRODUCTION

The enormous increases in research and development (R&D) spending, the dearth of approvals of New Chemical Entities (NCEs) and the competition from generic drugs has driven biopharmaceutical companies to evaluate new business paradigms, one of them being Drug repurposing (DR). DR has traditionally been part of the drug development process as a strategy to preserve and extend the value of patents through reformulation strategies [1]. However, the productivity challenges of traditional drug discovery together with the exemplary success of sildenafil, duloxetine and thalidomide [2] as repositioned drugs, has sparked a renewed interest to DR, this time as an alternative approach to traditional drug discovery [3]. Several directed approaches to DR have been appearing in the scientific literature in past last 2 years [4–13], both from academia and industry, many of them based on computational techniques. These *in silico* efforts have often been driven by the massive amounts of data generated by high-throughput experiments and many times stored in bioinformatics and cheminformatics databases.

Corresponding author. Aris Persidis, Biovista Inc., 2421 Ivy Road, Charlottesville, VA 22903, USA. Tel: +1-434-971-1141/434-971-1143; Fax: +1-434-971-1144; E-mail: arisp@biovista.com

Christos Andronis is Vice President for Research and Development at Biovista. He has a PhD in Biochemistry from Imperial College, London, UK. He has been involved in the design and implementation of Information Extraction systems and other Computational applications relevant to Drug Discovery.

Anuj Sharma is currently a staff scientist and developer at Biovista with expertise in Ontologies and Literature mining research.

Vassilis Virvilis is Head of IT at Biovista. He has a PhD in Artificial Intelligence. He has a long standing interest in Machine Learning Algorithms and Optimization.

Spyros Deftereos is Vice President for Drug Discovery at Biovista. He is a Medical Doctor with a PhD in Biomedical Informatics. He has a long standing interest in movement disorders, neuromuscular diseases and clinical neurophysiology.

Aris Persidis is co-founder and president of Biovista. He has a PhD in Biochemistry from Cambridge University. Dr. Persidis has authored the Industry-Trends series of papers in Nature Biotechnology (1997–2000).

Much of the knowledge covering modern biology and medicine is often buried in various forms of free-text documents. Literature mining techniques are increasingly being developed to harness the information stored in scientific articles and to infer relationships between biomedical concepts, even if they are not mentioned in the same abstract. Advances in literature mining have enabled the creation of networks of associations among many types of biomedical entities, including genes, drugs and diseases [14–20].

Bringing together information gathered from literature mining or from other high-throughput technologies, such as microarrays and proteomics, has been a driving force behind the recent interest in building and using biomedical ontologies [21, 22]. At the basic level, ontologies are created with the purpose of describing a scientific domain; in other words, they provide a means of describing concepts and their inter-relations in the context of a knowledge domain of interest. Distant relations between islands of knowledge that are seemingly unrelated to each other, such as the association of an existing drug to a novel medical condition [23] can be identified by traversing an ontology.

Graphical visualization of biomedical networks can be of great use when we are in a discovery mode trying to reveal connections between distant entities. Graphical layouts such as heat maps also help us visualize patterns and trends hidden in data produced by high-throughput experiments of e.g. drug interventions to gene expression or to disease progression.

In this article, we will review various literature mining approaches to drug discovery and drug repurposing. Ontological resources as well as information visualization methodologies suited to drug repurposing will also be described.

TECHNOLOGICAL OVERVIEW

Literature mining

PubMed, the most widely used repository of biomedical articles contains over 20 million abstracts and is growing with a rate of over 850 000 abstracts per year: the number of articles added each year to PubMed has almost tripled in the last decade (Figure 1). As research on a single topic may span across many scientific disciplines and biomedical journals, it is increasingly difficult for scientists to follow all advances in their field of interest.

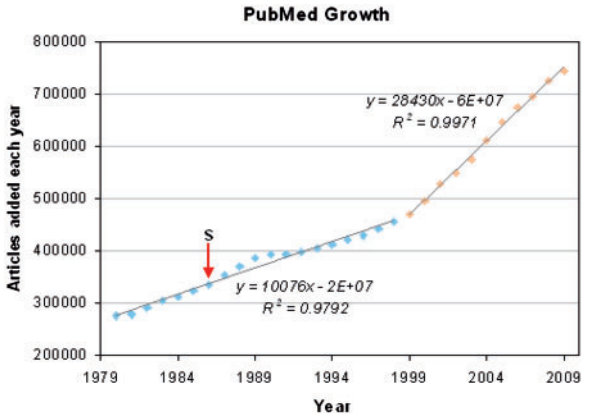

Figure 1: PubMed has almost tripled its growth rate in the past 10 years. The number of articles added to PubMed each year is plotted against the year they were added. The trend lines represent the rate PubMed is growing over the years starting from 1980. The arrow ('S') points to the year Swanson published his fish oil—Raynaud's disease paper (1986).

The diffusion of knowledge to many different journals and scientific disciplines has started gradually creating 'islands of knowledge' and led to the development of literature mining methodologies aiming to link concepts and arguments that are not mentioned in the same article. The process of inferring implicit knowledge from seemingly unrelated concepts has been called literature-based discovery (LBD).

The foundations for this view of the biomedical literature and its potential to be used for drug discovery were originally laid by Don R. Swanson in the mid-1980s [24]. Using an approach to literature mining, referred to as the ABC model, Swanson made several scientific hypotheses, including the beneficial effects of fish oil to patients with Raynaud's disease [25], and the potential of magnesium to treat migraines [26]. The fish oil connection to Raynaud's disease was later validated in clinical trials [27], whereas the prophylactic role of magnesium in migraines is now well established in the clinical practice [28].

The premise of the ABC model (Figure 2A) is that there are two concepts or bodies of knowledge that do not communicate explicitly with each other. However, part of the knowledge of one such domain may complement the knowledge of the other one. Suppose that one scientific community knows that B is one of the characteristics of disease C. Another scientific discipline has found that

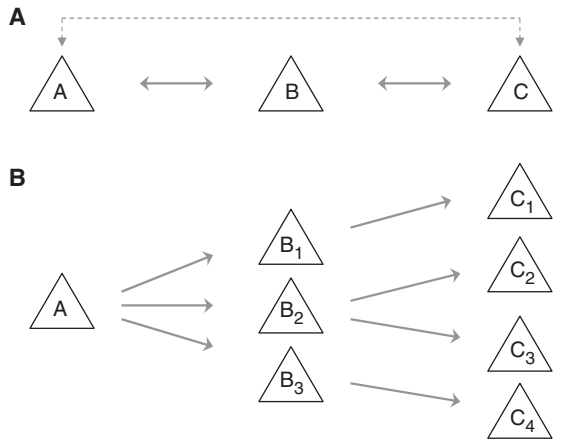

Figure 2: (A) Graphical representation of Swanson's ABC model. Concepts A and C may have an implicit connection to each other, if they share an explicit connection with concept B. This is also called the closed discovery model. **(B)** The open discovery model [29]. Concept A is linked to one or more concepts C through intermediate concepts B.

substance A affects B. Discovery in this case is making the implicit link AC through the B-connection.

Swanson's analogical reasoning approach was later formalized by Weeber and coworkers, who proposed a two step process for discovery based on the ABC model. In the 'closed discovery' process, where the two starting concepts A and C are already known, the purpose of the discovery process is to interpret the implicit relationship between A and C (Figure 2A). This is the original conceptual model of Swanson. The open discovery process (Figure 2B) begins with a concept A, for example, a disease and looks for arguments B related to concept A, for example, the mechanism of the disease. At the second step, concepts C are sought that share the B arguments; for example drugs, the effects of which are mediated by this mechanism [29, 30].

Both open and closed LBD modes have inspired literature mining researchers to create tools attempting to (semi) automate the discovery process, taking advantage of information extraction (IE) techniques to efficiently extract relevant information from the scientific literature and integrating additional data types and sources to the discovery process on top of the scientific literature. Almost all applications of LBD described in the literature attempt to discover treatments for diseases using known drugs and nutritional supplements as their basis, demonstrating the applicability and relevance of LBD to drug repurposing [13, 17–19, 35, 40, 42–46, 75, 82]. It is also of interest that relationships between concepts have almost exclusively been based on co-occurrence of terms and facts in the same abstract.

Smalheiser and Swanson proceeded with the creation of a web tool called Arrowsmith, which builds on the closed discovery mode of the original ABC model, adding a level of automation to the interpretation of the implicit connection between the A and C concepts [31]. Gordon and Lindsay [32] employed the open discovery mode and used lexical statistics over titles and abstracts to recreate Swnason's discoveries. Weeber *et al*. [29] pursued the same goal using the Unified Medical Language System (UMLS) [33] and lexical tools to map natural language text to UMLS concepts. The tool they built, called DAD was based on the MetaMap program [34] to map words in the abstract to UMLS concepts. Weeber *et al*. [35], following their elaboration of the ABC model, made four novel therapeutic applications for thalidomide: myasthenia gravis, chronic hepatitis C, *Helicobacter pylori*-induced gastritis and acute pancreatitis. LitLinker [36] and Telemakus [37] are two other systems taking advantage of the MetaMap program and UMLS. LitLinker also used MeSH [38] for term selection and reduction steps. Telemakus also enables the visualization of the networks it creates, in a conceptual graph. Van der Eijk *et al*. [39] extracted information from abstracts and then built a co-occurrence based Associative Concept Space (ACS) algorithm to place all concepts in an *n*-dimensional space. Concepts in close proximity are considered to be related and if these concepts do not have a direct relationship, a potentially novel relationship has been discovered. Wren *et al*. [40] constructed a co-occurrence based network of biomedical concepts extracted from MEDLINE and then used the strength of the associations to infer novel relationship between cardiac hypertrophy and Chlorpromazine. The authors then went ahead and validated this relationship in animal models. Narayanasamy *et al*. [41] also created an association graph by mining MEDLINE for biomedical concepts and then used this network to find transitive associations in a tool called TransMiner. In a study pursuing the closed discovery mode of LBD, Ahlers *et al*. [42] extracted semantic predications from MEDLINE and then used this information to identify proteins that potentially provide a link between cancer and anti-psychotic agents. Petric *et al*. [43] put

an emphasis on rare terms in the discovery process, linking autism to calcineurin. Hristovski *et al*. [44] produced a tool called BITOLA that computes association rules between concepts extracted from MEDLINE. These rules are used to represent the known relations between the concepts. Results are ranked according to parameters that measure association strength. In another study demonstrating the discovery potential of the closed discovery mode, Baker *et al*. [45] used MeSH terms to find drugs and their effects in MEDLINE abstracts and connected those to diseases using proteins as the intermediary B concepts. Ha *et al*. [13] have built a program employing cheminformatics and text mining techniques to predict novel relations between existing drugs and drug targets. Finally, Zhu *et al*. [18] created a data mining tool using aggregate web services to characterize compounds and find non-obvious relationships between them and genes, diseases and biomedical articles.

The LBD approach may be pursued without resorting to the use of free text for the extraction of starting or ending concepts. Srinivasan and Libus [46] employed MeSH terms and developed an algorithm that used a weighting system to select B (intermediate) concepts finally producing ranked lists of C concepts to explore the therapeutic potential of turmeric/curcumin, a dietary supplement.

Information Extraction

Swanson's original discoveries were based on manual scanning of the biomedical literature. There were approximately 7 million papers in PubMed in 1986, the year Swanson published his first discovery (Figure 1), whereas in 2009 there were over 19 million abstracts stored in this repository. Most practitioners of the LBD approach in the last decade had to resort to some form of text mining in order to identify and extract all relevant facts from free text. In some cases, MeSH headings were used as a substitute of text mining. Although, MeSH is a remarkable resource for annotations, a significant amount of information is still present only in the abstract and is missing from the MeSH headings [47].

An important first step toward the discovery of novel links between seemingly unrelated concepts is the extraction of these concepts and their relationships from a single article or abstract, in a process called Information Extraction (IE). IE usually begins with Named Entity Recognition (NER), which deals with the correct identification of biomedical terms in free text. Terms might be identified using controlled vocabularies, such as UMLS, MeSH for diseases, Uniprot [48] and NCBI Entrez Gene [49] for genes and Reactome [50] for pathways, or through Natural Language Processing (NLP) and machine learning techniques [51, 52]. The NER step is a prerequisite for any IE project; however, it is a quite difficult task due to the lack of standardization of names and the issues of synonymy and polysemy [16, 53]. These problems are most evident with genes/proteins. Genes/proteins are described with a variety of descriptors, such the gene symbol, the gene name, the gene product name and various synonyms. However, in most cases there are more than one gene symbols per gene and scientists tend to refer to a gene with different names in the literature, many times not using the 'official' symbol. A good example is the p38 MAP kinase (Entrez Gene ID: 1432). The official gene symbol for that gene is MAPK14 and the official full name is mitogen-activated protein kinase 14. However, very few scientists refer to that gene with the name MAPK14—at least in the abstract of the article. To alleviate this issue, the Entrez Gene database contains a variety of synonyms for MAPK14, including p38. However, the issue of synonymy and lack of standardization are not the only challenges for determining the identity of a gene in free text. p38 is a synonym for over 20 different genes from a variety of organisms, including humans, flies and viruses. This phenomenon is called polysemy and refers to the capacity of a name to have multiple meanings (i.e. functions). Recent methods, mostly based on machine learning techniques have tackled the issue of gene disambiguation with promising results. Most of these approaches are based on the context or other characteristic features surrounding the genes [54–58].

At its simplest form, IE only attempts to find co-occurring concepts within the same abstract or sentence. Co-occurrence is based on the notion that if two concepts are mentioned in the same body of text, they are possibly related to each other. Although, this method of association seems very limited in terms of its granularity, it seems to be well suited for situations where one is in exploratory mode by way of not filtering out ambiguous relationships and by providing associations of almost any type. Actually, co-occurrence has been the dominant method of relationship employed in the LBD articles that are described in the previous section. The disadvantages of co-occurrence, i.e. its

inability to provide any information regarding the nature of the relationship and its potential high rate of false positives have been the driving force behind the development of NLP techniques which attempt to identify concepts and their relationships in a unified way. In the typical NLP setup, text is first tokenized to identify word, and sometimes sentence boundaries followed by tagging of the type of the word (e.g. a noun) by specialized systems called part-of-speech (POS) taggers. Entities (words) extracted are then semantically mapped to a biomedical category (e.g. a gene or a disease) and a syntactic tree is constructed representing the structure of the whole sentence or phrase. NLP methodologies have been improving in recent years [52, 59] and by providing directionality in concept relationship might soon be in position to offer better guidance for new discoveries than co-occurrence.

Semantic web technologies and ontologies

Biology differs from other scientific disciplines, such as Physics in that biological/biochemical processes have not been completely modeled mathematically. Relationships between domain entities must therefore be recorded in more subjective ways, as opposed to mathematical formulas, while retaining the ability to apply computational techniques. Ontological representation has provided an answer to the issues presented by the incomplete understanding of biological processes [60]. At the same time, ontologies have been providing a means of standardization unifying facts from various domains of knowledge that might have originally been recorded in different formats. Ontologies can be used to answer questions that may be inferential in nature. It is this characteristic that makes the use of Ontological resources significant in drug repurposing. Despite some shortcomings of popular semantic web languages to comprehensively define the often incomplete nature of biomedical data [61, 62], these languages have been used extensively for representing relationships in the biomedical field.

Semantic web [63] technologies are widely used as a means for formal description of concepts and their relationships in any given knowledge domain and they have been applied in the field of bioinformatics as discussed above. Web Ontology Language (OWL) [64], built on top of Resource Description Framework (RDF) [65], is the standard currently adopted by the W3C working group for authoring ontologies. A large number of tools have been developed that allow easy manipulation and generation of ontologies using OWL including PROTEGE [66], FACT^{++} [67] and HermiT [68].

Owing to the presence of a large number of small yet useful ontological resources contributed by different domains of bioinformatics, in 2003 an initiative called OBO was launched by Ashburn *et al.* [69]. The objective of the initiative was to define a standard around which biological ontologies could be built so that they could be integrated or used with computational approaches seamlessly. The initiative later diversified into the OBO Foundry, a more elaborate attempt at establishing principles for ontological development for representing biomedical data. The availability of powerful tools for querying an ontology in OWL format and the acceptance of OWL as the *de facto* standard language has brought about several efforts for converting data from OBO to OWL format [70]. An increasing number of biomedical ontological resources such as Gene Ontology (GO) are now available in OWL format in addition to the original OBO format. Finally, with the conversion of several ontologies to OWL and RDF formats, similarity metrics have been developed to combine biomedical ontologies for the purpose of inferring unknown relationships. A detailed treatment and broad classification of these techniques can be found in [71].

Among the multitude of ontologies capturing biomedical information GO and UMLS are among the most referenced ontologies in drug repurposing applications [29, 37]. Originally started as collaboration between three model organism databases in 1998, GO [72] has grown to include several major repositories from different organisms. GO defines terms and relations for three domains: cellular components, biological processes and molecular function. The ontology can be downloaded in several formats including OBO, RDF, OWL and MySQL. While not a unification standard for biological information, GO is one of the most significant steps in the direction of providing a common terminology for describing genes and their relationships. Unified Medical Language System (UMLS) (http://www.nlm.nih.gov/research/umls/) is a compendium of data gathered from over 100 knowledge sources, including some of the most important thesauri and vocabularies such as MeSH, MedDRA, OMIM and GO. At the heart of UMLS lies the meta-thesaurus which contains over 1 million concepts and over

4.0 million concept names and describes relationships between these concepts based on information extracted from several controlled vocabularies (English language sources only). UMLS provides a single standardized format for accessing information which otherwise is represented in different controlled vocabularies in varying formats.

Finding inferences through visualization techniques

Ontological approaches to finding novel connections are commonly associated with visualization technologies aiming to help scientists recognize non-obvious patterns and connections, e.g. for gene–gene relations in the context of a signaling pathway. For this reason, most semantic tools (that manipulate ontologies in OWL/RDF formats) also include comprehensive visualization options to enable knowledge discovery. Graphical representations enable the human brain to take advantage of spatial relationships that might not be immediately evident and to form hypotheses by recognizing patterns and other visual clues [73]. Three types of visualization techniques are commonly found in drug repurposing articles: network graph representations [74], heat-maps connecting sets of experiments with panels of drugs, genes, etc. [75] and tools enabling the visualization of docking experiments [76]. Luckily there is a multitude of graphics languages (SBML, GML, SIF) and tools (Cytoscape, Graphviz, etc.) that can be used to visualize complex data with minimum effort.

Many other types of structured data visualization techniques have been reviewed by Katifori *et al*. [77]. 3D visualizations such as hyperbolic trees may be an even better fit for the currently advancing 3D hardware accelerated landscape. The main problem of any 2D visualization is the cluttered image produced when massive information is passed to the visualization system. Since the primary aim of visualization is to provide overview and allow exploration in the future a lot of work will be concentrated in how we can use more effectively 3D technologies.

DRUG REPURPOSING APPLICATIONS

In the past 2 years directed efforts toward finding new uses for existing drugs have emerged, both from academia and companies. These attempts have used text mining with or without other sources of data (e.g. protein interaction, genetic data and cheminformatics) to perform novel discoveries. Frijters *et al*. [17] constructed a literature mining tool called CoPub Discovery with the purpose of finding novel connections between drugs, genes and diseases. CoPub Discovery is based on Swanson's ABC model. Gene names and other biomedical concepts were extracted from Medline abstracts and related to each other using co-occurrence. The authors used a mutual information-based metric to assess the strength of co-citations and presented a series of case studies with novel open and closed model discoveries, including disease–gene, drug–disease, drug–biological process and biological process–drug relationships. The latter scenario led to the identification of dephostatin, a tyrosine phosphatase inhibitor and damnacanthal, a tyrosine kinase inhibitor, being relevant to cell proliferation. *In vitro* cell proliferation experiments validated the influence of these two compounds in the process of cell proliferation at low micro molar concentrations.

Li *et al*. [75] integrated protein interaction data with literature mining to devise a computational framework that can be used to build disease-specific drug–protein connectivity maps. The authors used Alzheimer's disease (AD) as a case study. They started with a list of disease-related proteins taken from OMIM [78], which contains manually curated disease gene lists, and amended that with interacting proteins from OPHID [79] and their own nearest-neighbor protein interaction expansion method. Using this list of AD-relevant proteins they retrieved a set of abstracts from PubMed and then extracted all known drugs from these articles with the purpose of identifying new drugs associated with AD. Using this approach, they selected diltiazem, an antihypertensive agent and quinidine, an antiarrhythmia agent as candidate treatments for AD.

The intimate relationship between text mining tasks such as information retrieval and information extraction with ontologies has been described previously [80, 81]. These examples demonstrate the use of ontologies both as controlled vocabularies and as a means of storing the extracted information with the purpose of augmenting the knowledge captured by the ontology. Finally, they also provide evidence for the application of semantic technologies for higher order inference from ontologies.

A good example of the utility of semantic queries being successfully applied to drug–disease relation identification is the work done by Qu *et al*. [82].

A semantic web of ontologies called the disease drug correlation ontology (DDCO) was constructed by combining ontologies taken from DrugBank, Entrez Gene, GO, OMIM, KEGG, BioCarta, Reactome, UMLS and GEO (the Gene Expression Omnibus). An association network was constructed using the combination of information captured by the semantic web and SPARQL, the query language for ontologies was then employed to deduce drug–disease associations and verified by finding drugs for the disease Systemic Lupus Erythematosus. Cockell *et al.* [7] also present a graph based approach for inferring novel drug–disease relationship. A graph of 120 000 concepts with 570 000 relationships between drugs, proteins and diseases is constructed by combining data from DrugBank, UniProt, HRPD, KEGG, Pfam, SynAtlas, G-sesame, OpenBabel and BLAST. The Ondex platform [83] was applied for collating data from these diverse sources into one large data set and the visualization of the data set. Inference of drug repurposing possibilities is then performed using the interconnection graph and semantic web query techniques.

Choi *et al.* [84] described a methodology for constructing a Small Molecule Ontology (SMO), combining data freely available from three sources—DrugBank, PubChem and UniProt—using the Protege tool. The ontology is developed on OWL-DL using RDF triples. The finished ontology is queried using SPARQL in order to demonstrate the applicability of semantic web technologies to inferring novel relations for drug repurposing. The ontology schema and the specific instance created for the work performed are publicly available.

Cure and Giroud [85] present a workflow for improving the quality of data stored in drug databases. The first step is the conversion of existing terminologies into semantic web compatible format. This is followed by manual curation and refining of the concepts and associations of the resulting ontologies. New concepts are associated with the ontologies by applying inductive reasoning to the drug database. The ontologies can then be used to find and repair data quality violations in drug databases. These checks for violations may then also be performed each time the database updates, ensuring the data remain consistent.

The visualization techniques in articles relevant to drug repurposing are either used to conceptualize an automatic algorithm that seeks associations of interest or simply act as an exploration tool to manually perform the detection.

The application of graph-theoretic analysis for the discovery of novel gene–disease associations is exemplified by the work of Cerami *et al.* [86].The authors used the Newman–Girvan module detection algorithm to find candidate 'drivers', i.e. genes responsible for Glioblastoma, such as the AGAP2 gene and three signaling modules, including one involved in microtubule organization. Keiser *et al.* [74] demonstrated the ability to predict new targets for existing drugs by comparing the similarity of the ligands binding to these targets. In a similar fashion, Campillos *et al.* [87] predicted novel targets for existing drugs by looking at the side effect similarity of otherwise unrelated drugs. A more comprehensive effort was done by Hu and Agarwal [9] by combining multiple interaction networks.

Pujol *et al.* [8] argues that diseases are complex and involve multiple pathways, more than the current, high target specificity drugs, can address. To that end, they suggest employing network statistics and topological techniques such as node centrality, modularity, between-ness, shortest path, clustering and more to decipher the complex biological networks.

In this context and in an attempt to build an all inclusive global database, Paolini *et al.* [88] merged structure–activity relationships (SAR) from several sources in one system. In order to visualize the associations between various classes of drug targets, they used networks to display the connectivity of the targets according to the drugs binding to them, followed by heat maps to compare the promiscuity among all known drug target families. The above methods require prior knowledge of the drugs or the targets involved and are already stored in a database from where the knowledge can be mined. Iorio *et al.* [4] suggested a similar approach based on gene signatures from transcriptional data generated by high-throughput experiments so no prior knowledge is required for the drug under consideration. The drug–drug network that they built is based on the drug distance which is a metric of the drug profile similarity. Literature mining data are also used in TransMiner by Narayanasamy *et al.* [41] in order to build a network and find transitive connections.

Li *et al.* [75] used a heat map to compare similarity of protein–compound association profiles, as extracted from the literature, in combination with

hierarchical clustering. A more ambitious approach using heat maps is the attempt by Korbel *et al*. [89] to link phenotypes with genomic data through literature and comparative genome analysis. One more encompassing attempt to grasp and visualize the target drug from conception to production, was done by Campbell *et al*. [73] also by using heat maps in the initial stages of drug development.

A number of companies providing software and services in the field of Literature mining have also engaged in computational drug repurposing providing services and using their technology to pursue their own discovery pipelines. Ariadne Genomics (Rockville, MD, USA) have recently published the repurposing of Fulvestrant, an estrogen receptor antagonist originally used in breast cancer, to glioblastoma using publicly available microarray data combined with their own suite of pathway analysis tools [90]. GeneGo (St Joseph, MI, USA) combines chemical structural analysis tools with molecular interaction and pathway analysis data to produce a list of putative new indications.

Biovista (Charlottesville, VA, USA) also makes use of literature mining techniques to infer new uses for existing drugs. The workflow for discovery is based on systems literature analysis (SLA) [91] which treats large sets of scientific literature as a vast system of interconnections between research parameters. SLA-based discovery combines IE tools, a database of relations among biomedical entities and inferential algorithms rooted in the LBD approach, to arrive at previously unknown relationships between drugs, genes, diseases and adverse drug reactions. The overall platform, called Clinical Outcome Search Space™ (COSS), also incorporates other data types, such as cheminformatic and pathway analysis data and also data derived from clinical pharmacology sources (e.g. PK/PD data). Central to the COSS platform is a proprietary storage and retrieval engine which utilizes custom technological solutions to achieve high-performance access to the underlying data. Fast retrieval is a very important component of the workflow as it allows multiple iterations and experimentation cycles needed in a typical discovery scenario, at a production environment.

This technology has enabled Biovista to repurpose Dimebon, an anti-histamine drug and Pirlindol, an anti-depressant drug, to Multiple Sclerosis [92, 93]. In addition, Biovista's technology is being used by the FDA's Office of Clinical Pharmacology (OCP) in its assessment of the adverse event profiles of major drug classes and drugs being evaluated by FDA. Figure 3 depicts two ways of visualizing literature-based relationships between genes and diseases, used at Biovista.

CONCLUDING REMARKS

Drug repurposing relies heavily on prior knowledge surrounding a drug and a putative novel indication. The more one knows about the molecular basis of a disease or the mechanism of action of a drug (its target and signaling pathways affected), the more 'educated' a guess will be concerning the selection of the right indication and avoiding signaling pathways that may lead to adverse drug reactions. Some of the known facts related to drugs, genes/proteins and diseases are stored in various bioinformatic and clinical databases. However, most of these facts are still recorded in free-text form in bibliographic repositories. As evidenced from the various examples given in this review, biomedical literature mining, especially the combination of efficient IE with LBD seems to be well suited as a strategy to generate scientific hypotheses related to finding new uses for existing drugs. A novel hypothesis can be made more solid by the structure conferred to knowledge by ontologies. In addition, the ability to gain an overview of the existing knowledge by Information Visualization techniques, combined with the various clustering algorithms that bring similar concepts close in the space of a graph, may provide additional guidance to drug repurposing exercises.

Tools assisting in efficient scanning of the biomedical literature are increasingly becoming more accessible to scientists engaged in drug discovery. PubMed now offers powerful Boolean-type search capabilities and online tools assisting in information extraction and retrieval are increasingly becoming more accessible. It is obvious that a wider deployment of online hypothesis generation tools would be equally valuable; however, any such tool should incorporate a solid IE step that would maximize the number of concepts (and their relationships) being utilized and also be updated on a regular basis. At the same time, it is clear that the value of these tools will increase as other data types get integrated into the hypothesis generation workflow.

An important point regarding any computationally based drug repurposing program is that all

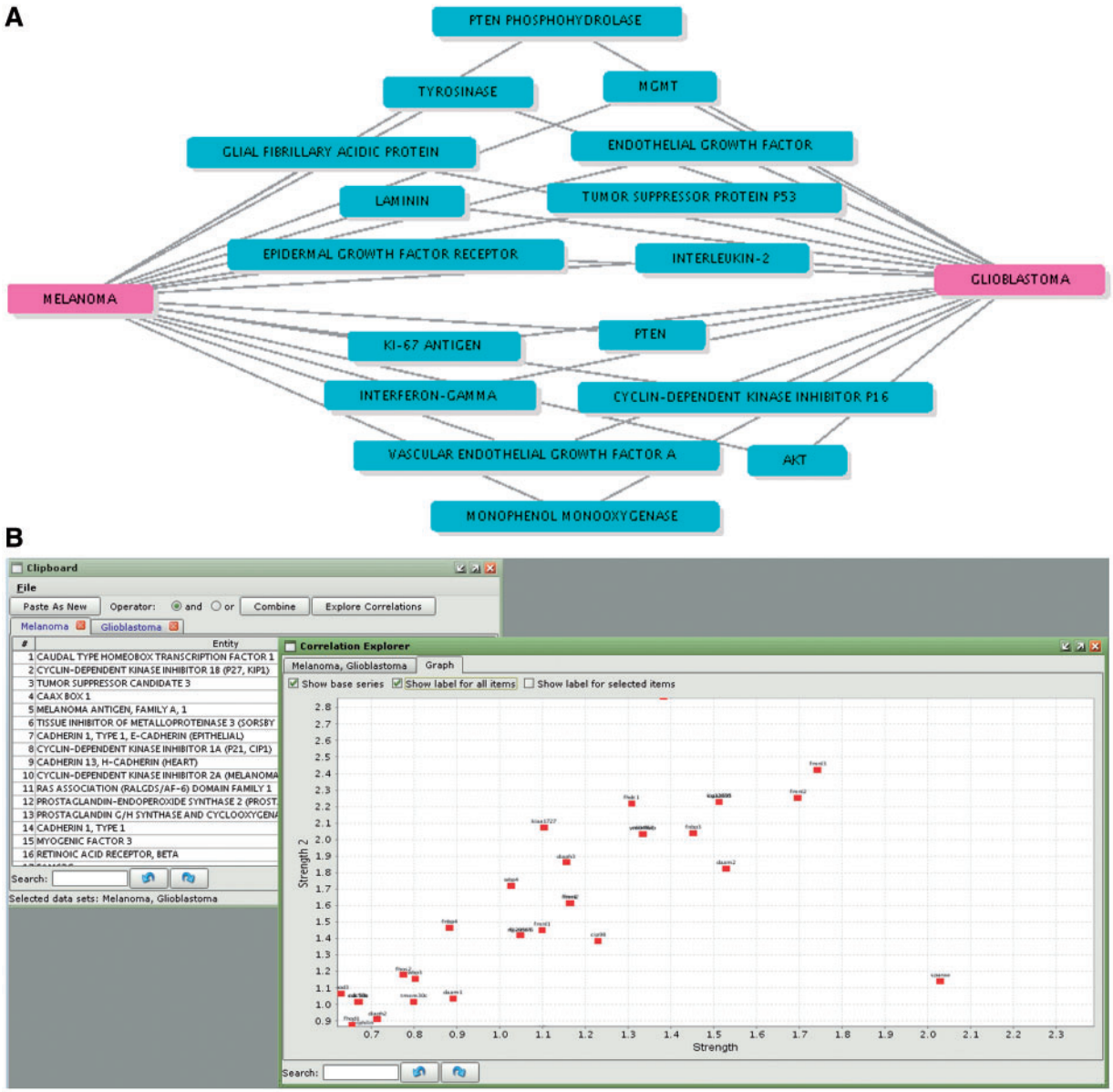

Figure 3: Screenshots from the analysis modules of Biovista's COSS platform. First (**A**) and second (**B**) order correlations between genes related to Melanoma and Glioblastoma Multiforme. The screenshot focuses on the strongest correlations, as determined by literature mining alone.

hypotheses formed should always be followed by manual curation by the respective domain experts. Manual curation increases the value of a hypothesis by removing uncertainties and taking into account other parameters related to the novel association, such as pharmacokinetic/pharmacodynamic data, etc.

A final point regarding computational drug repurposing is the aspect of adverse drug reactions associated with the novel use of a drug. Although, the premise of drug repurposing is based on the utilization of existing drugs, the potential of a drug to generate an adverse reaction should not be assumed non-existent when given to a patient under conditions that are different from its original use, e.g. for chronic use or in a different formulation or dosage. This is especially true for drugs without extended post-market experience, such as compounds that have been discontinued during the late clinical phases before entering the market. It would therefore be desirable for any drug repurposing system to also be able to predict adverse drug reactions alongside the novel use of the drug in a different indication.

Key Points

- The productivity challenges of traditional drug discovery, has sparked a renewed interest to Drug Repurposing, as an alternative approach to traditional drug discovery.
- As research on a single topic may be spanning across many scientific disciplines and biomedical journals, it is increasingly difficult for scientists to follow all advances in their field of interest. Advances in literature mining have made it possible to infer relationships between biomedical concepts, even if they are not mentioned in the same abstract.
- Biomedical literature mining, especially the combination of efficient IE with LBD seems to be well suited as a strategy to generate scientific hypotheses related to finding new uses for existing drugs.
- Ontologies capture domain knowledge, concepts and their relationships, and have been used to infer unknown relationships, through automated reasoning, making them indispensable in drug repurposing.
- Visualization techniques in articles relevant to drug repurposing are either used to conceptualize an automatic algorithm for detection of an association of interest or simply act as an an exploration tool to manually perform the detection.

References

1. Fleming E, Ma P. Drug life-cycle technologies. *Nat Rev Drug Discov* 2002;**1**(10):751–2.
2. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* 2004;**3**(8):673–83.
3. Campas C. Drug repositioning summit: finding new routes to success. *Drug News Perspect* 2009;**22**(2):126–8.
4. Iorio F, Bosotti R, Scacheri E, *et al*. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci USA* 2010;**107**(33):14621–6.
5. Dudley JT, Schadt E, Sirota M, *et al*. Drug discovery in a multidimensional world: systems, patterns, and networks. *J Cardiovasc Transl Res* 2010;**3**(5):438–47.
6. Kotelnikova E, Yuryev A, Mazo I, *et al*. Computational approaches for drug repositioning and combination therapy design. *J Bioinform Comput Biol* 2010;**8**(3):593–606.
7. Cockell SJ, Weile J, Lord P, *et al*. An integrated dataset for *in silico* drug discovery. *J Integr Bioinform* 2010;**7**(3):116.
8. Pujol A, Mosca R, Farrés J, *et al*. Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol Sci* 2010;**31**(3):115–23.
9. Hu G, Agarwal P. Human disease-drug network based on genomic expression profiles. *PLoS One* 2009;**4**(8):e6536.
10. Kinnings SL, Liu N, Buchmeier N, *et al*. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. *PLoS Comput Biol* 2009;**5**(7):e1000423.
11. Chiang AP, Butte AJ. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clin Pharmacol Ther* 2009;**86**(5):507–10.
12. Scheiber J, Chen B, Milik M, *et al*. Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. *J Chem Inf Model* 2009;**49**:308–17.
13. Ha S, Seo YJ, Kwon MS, *et al*. IDMap: facilitating the detection of potential leads with therapeutic targets. *Bioinformatics* 2008;**24**(11):1413–5.
14. Jenssen TK, Laegreid A, Komorowski J, *et al*. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* 2001;**28**(1):21–8.
15. Chun HW, Tsuruoka Y, Kim JD, *et al*. Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. *Pac Symp Biocomput* 2006; 4–15.
16. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet* 2006;**7**(2):119–29.
17. Frijters R, van Vugt M, Smeets R, *et al*. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput Biol* 2010;**6**(9). pii: e1000943.
18. Zhu Q, Lajiness MS, Ding Y, *et al*. WENDI: a tool for finding non-obvious relationships between compounds and biological properties, genes, diseases and scholarly publications. *J Cheminform* 2010;**2**:6.
19. Baker NC, Hemminger BM. Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. *J Biomed Inform* 2010;**43**(4):510–9.
20. Agarwal P, Searls DB. Literature mining in support of drug discovery. *Brief Bioinform* 2008;**9**(6):479–92.
21. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. *Brief Bioinform* 2006;**7**(3):256–74.
22. Bard JB, Rhee SY. Ontologies in biology: design, applications and future challenges. *Nat Rev Genet* 2004;**5**(3): 213–22.
23. Qu XA, Gudivada RC, Jegga AG, *et al*. Inferring novel disease indications for known drugs by semantically linking drug action and disease mechanism relationships. *BMC Bioinform* 2009;**10**(Suppl 5):S4.
24. Swanson DR. Medical literature as a potential source of new knowledge. *Bull Med Libr Assoc* 1990;**78**(1):29–37.
25. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med* 1986;**30**(1):7–18.
26. Swanson DR. Migraine and magnesium: eleven neglected connections. *Perspect Biol Med* 1988;**31**(4):526–57.
27. DiGiacomo RA, Kremer JM, Shah DM. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. *Am J Med* 1989;**86**(2):158–64.
28. Schiapparelli P, Allais G, Castagnoli Gabellari I, *et al*. Non-pharmacological approach to migraine prophylaxis: part II. *Neurol Sci* 2010;**31**(Suppl 1):S137–9.
29. Weeber M, Klein H, de Jong-van den Berg LT, *et al*. Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *J Am Soc Inform Sci Tech* 2001;**52**:548–57.
30. Weeber M, Kors JA, Mons B. Online tools to support literature-based discovery in the life sciences. *Brief Bioinform* 2005;**6**(3):277–86.
31. Smalheiser NR, Swanson DR. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput Methods Programs Biomed* 1998; **57**(3):149–53.
32. Lindsay RK, Gordon MD. Literature-based discovery by lexical statistics. *J Am Soc Inform Sci* 1999;**50**:574–87.

33. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**(Database issue):D267–70.
34. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001;17–21.
35. Weeber M, Vos R, Klein H, *et al*. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *J Am Med Inform Assoc* 2003;**10**(3):252–9.
36. Yetisgen-Yildiz M, Pratt W. Using statistical and knowledge-based approaches for literature-based discovery. *J Biomed Inform* 2006;**39**(6):600–11.
37. Fuller SS, Revere D, Bugni PF, *et al*. A knowledgebase system to enhance scientific discovery: Telemakus. *Biomed Digit Libr* 2004;**1**(1):2.
38. Medical Subject headings factsheet. 2010. http://www.nlm.nih.gov/pubs/factsheets/mesh.html (10 January 2011, date last accessed).
39. van der Eijk CC, van Mulligen EM, Kors JA, *et al*. Constructing an associative concept space for literature-based discovery. *J Am Soc Inform Sci Technol* 2004;**55**(5):436–44.
40. Wren JD, Bekeredjian R, Stewart JA, *et al*. Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics* 2004;**20**(3):389–98.
41. Narayanasamy V, Mukhopadhyay S, Palakal M, *et al*. TransMiner: mining transitive associations among biological objects from text. *J Biomed Sci* 2004;**11**(6):864–73.
42. Ahlers CB, Hristovski D, Kilicoglu H, *et al*. Using the literature-based discovery paradigm to investigate drug mechanisms. *AMIA Annu Symp Proc* 2007;6–10.
43. Petric I, Urbancic T, Cestnik B, *et al*. Literature mining method RaJoLink for uncovering relations between biomedical concepts. *J Biomed Inform* 2009;**42**(2):219–27.
44. Hristovski D, Peterlin B, Mitchell JA, *et al*. Using literature-based discovery to identify disease candidate genes. *Int J Med Inform* 2005;**74**(2–4):289–98.
45. Baker NC, Hemminger BM. Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. *J Biomed Inform* 2010;**43**(4):510–9.
46. Srinivasan P, Libbus B. Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics* 2004;**20**(Suppl 1):i290–6.
47. Kostoff RN, Block JA, Stump JA, *et al*. Information content in Medline record fields. *Int J Med Inform* 2004;**73**(6):515–27.
48. UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* 2010;**38**(Database issue):D142–8.
49. NCBI Entrez Gene. http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene (10 January 2011, date last accessed).
50. Matthews L, Gopinath G, Gillespie M, *et al*. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 2009;**37**(Database issue):D619–22.
51. Zweigenbaum P, Demner-Fushman D, Yu H, *et al*. Frontiers of biomedical text mining: current progress. *Brief Bioinform* 2007;**8**(5):358–75.
52. Cohen KB, Hunter L. Getting started in text mining. *PLoS Comput Biol* 2008;**4**(1):e20.
53. Cohen AM, Hersh WR. A survey of current work in biomedical text mining. *Brief Bioinform* 2005;**6**(1):57–71.
54. Hatzivassiloglou V, Duboué PA, Rzhetsky A. Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics* 2001;**17**(Suppl 1):S97–106.
55. Tveit A, Saetre R, Laegreid A, *et al*. ProtChew: Automatic extraction of protein names from biomedical literature. In *Proceedings of the 21st international Conference on Data Engineering Workshops* 2005;1161.
56. Schuemie MJ, Kors JA, Mons B. Word sense disambiguation in the biomedical domain: an overview. *J Comput Biol* 2005;**12**(5):554–65.
57. Alexopoulou D, Andreopoulos B, Dietze H, *et al*. Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. *BMC Bioinformatics* 2009;**10**:28.
58. Xu H, Markatou M, Dimova R, *et al*. Machine learning and word sense disambiguation in the biomedical domain: design and evaluation issues. *BMC Bioinformatics* 2006;**7**:334.
59. Chen ES, Hripcsak G, Xu H, *et al*. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. *J Am Med Inform Assoc* 2008;**15**(1):87–98.
60. Ontologies in Bioinformatics *EMBER Project Workbook on Bioinformatics*. (T. Atttwood, ed.) Vary, J. P. (Ed.), 2000.
61. Stevens R, Aranguren ME, Wolstencroft K, *et al*. Using OWL to model biological knowledge. *Int J Hum-Comput Stud* 2007;**65**:7583–94.
62. Schulz S, Stenzhorn H, Boeker M, *et al*. Strengths and limitations of formal ontologies in the biomedical domain. *RECIIS Rev Electron Comun Inf Inov Saude* 2009;**3**(1):31–45.
63. http://www.w3.org/2001/sw/ (6 February 2011, date last accessed).
64. Web Ontology Language. http://www.w3.org/TR/2004/REC-owl-guide-20040210/ (10 January 2011, date last accessed).
65. Resource Description Framework. http://www.w3.org/TR/PR-rdf-syntax/ (10 January 2011, date last accessed).
66. PROTEGE. http://protege.stanford.edu/ (10 January 2011, date last accessed).
67. FACT++. http://owl.man.ac.uk/factplusplus/.
68. HermiT. http://hermit-reasoner.com/ (10 January 2011, date last accessed).
69. Ashburner M, Mungall CJ, Lewis SE. Ontologies for biologists: a community model for the annotation of genomic data. *Cold Spring Harb Symp Quant Biol* 2003;**68**:227–35.
70. Hoehndorf R, Oellrich A, Dumontier M, *et al*. Relations as patterns: bridging the gap between OBO and OWL. *BMC Bioinformatics* 2010;**11**:441.
71. Pesquita C, Faria D, Falcão AO, *et al*. Semantic similarity in biomedical ontologies. *PLoS Comput Biol* 2009;**5**(7):e1000443.
72. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**(1):25–9.
73. Campbell SJ, Gaulton A, Marshall J, *et al*. Visualizing the drug target landscape. *Drug Discov Today* 2010;**15**(1–2):3–15.
74. Keiser MJ, Setola V, Irwin JJ, *et al*. Predicting new molecular targets for known drugs. *Nature* 2009;**462**(7270):175–181.
75. Li J, Zhu X, Chen JY. Building disease-specific drug-protein connectivity maps from molecular interaction

networks and PubMed abstracts. *PLoS Comput Biol* 2009;**5**(7): e1000450.

76. Li YY, An J, Jones SJ. A large-scale computational approach to drug repositioning. *Genome Inform* 2006;**17**(2):239–47.
77. Katifori A, Halatsis C, Lepouras G, *et al*. Ontology Visualization Methods-A Survey. *ACM Comput Surveys* 2007;**39**(4). Article 10.
78. Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). http://www.ncbi.nlm.nih.gov/omim/ (10 January 2011, date last accessed).
79. Brown KR, Jurisica I. Online predicted human interaction database. *Bioinformatics* 2005;**21**(9):2076–82.
80. Jimeno-Yepes A, Berlanga-Llavori R, Rebholz-Schuhmann D. Exploitation of ontological resources for scientific literature analysis: searching genes and related diseases. *Conf Proc IEEE Eng Med Biol Soc* 2009;**2009**:7073–8.
81. Spasic I, Ananiadou S, McNaught J, *et al*. Text mining and ontologies in biomedicine: making sense of raw text. *Brief Bioinform* 2005;**6**(3):239–51.
82. Qu XA, Gudivada RC, Jegga AG, *et al*. Inferring novel disease indications for known drugs by semantically linking drug action and disease mechanism relationships. *BMC Bioinformatics* 2009;**10**(Suppl 5):S4.
83. Köhler J, Baumbach J, Taubert J, *et al*. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics* 2006;**22**(11):1383–90.
84. Choi J, Davis MJ, Newman AF, *et al*. A semantic web ontology for small molecules and their biological targets. *J Chem Inf Model* 2010;**50**(5):732–41.
85. Cure O, Giroud J. Ontology-based Data Quality enhancement for Drug Databases. WWW2007, May 8–12, 2007, Banff, Canada.
86. Cerami E, Demir E, Schultz N, *et al*. Automated network analysis identifies core pathways in glioblastoma. *PLoS One* 2010;**5**(2):e8918.
87. Campillos M, Kuhn M, Gavin AC, *et al*. Drug target identification using side-effect similarity. *Science* 2008;**321**(5886): 263–6.
88. Paolini GV, Shapland RH, van Hoorn WP, *et al*. Global mapping of pharmacological space. *Nat Biotechnol* 2006; **24**(7):805–15.
89. Korbel JO, Doerks T, Jensen LJ, *et al*. Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol* 2005;**3**(5):e134.
90. Kotelnikova E, Yuryev A, Mazo I, *et al*. Computational approaches for drug repositioning and combination therapy design. *J Bioinform Comput Biol* 2010;**8**(3):593–606.
91. Persidis A, Deftereos S, Persidis A. Systems literature analysis. *Pharmacogenomics* 2004;**5**(7):943–7.
92. Deftereos SN, Andronis C, Virvillis V, *et al*. Dimebon Ameliorates Disease Severity in the MOG-Induced Experimental Allergic Encephalomyelitis Animal Model of Progressive Multiple Sclerosis. American Neurological Association 135th Annual Meeting, September 12–15 2010, San Francisco, CA, USA.
93. Deftereos SN, Andronis C, Sharma A, *et al*. Systematic Drug Repurposing for CNS Indications: Account of Two Successful Case Studies. *A*. American Neurological Association 135th Annual Meeting, September 12–15 2010, San Francisco, CA, USA.