OXFORD

Briefings in Bioinformatics, 17(1), 2016, 33–42

doi: 10.1093/bib/bbv087
Advance Access Publication Date: 29 September 2015
Paper

Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery

Graciela H. Gonzalez, Tasnia Tahsin, Britton C. Goodale, Anna C. Greene and Casey S. Greene

Corresponding author. Casey S. Greene, Institute for Translational Medicine and Therapeutics, 10-131 Smilow Center for Translational Research, 3400 Civic Center Boulevard, Building 421, Philadelphia, PA 19104-5158, USA. Tel.: 215-573-2991; Fax: 215-573-9135; E-mail: csgreene@upenn.edu

Abstract

Precision medicine will revolutionize the way we treat and prevent disease. A major barrier to the implementation of precision medicine that clinicians and translational scientists face is understanding the underlying mechanisms of disease. We are starting to address this challenge through automatic approaches for information extraction, representation and analysis. Recent advances in text and data mining have been applied to a broad spectrum of key biomedical questions in genomics, pharmacogenomics and other fields. We present an overview of the fundamental methods for text and data mining, as well as recent advances and emerging applications toward precision medicine.

Key words: text mining; data mining; biomedical discovery; gene prioritization; pharmacogenomics; toxicology

Introduction

Technologies that resulted in the successful completion of the Human Genome project and those that have followed it afford an unprecedented breadth of data collection avenues (whole-genome expression data, chip-based comparative genomic hybridization and proteomics of signal transduction pathways, among many others) and have resulted in exceptional opportunities to advance the understanding of the genetic basis of human disease. However, high-throughput results are usually only the first step in a long discovery process, with subsequent and much more time-consuming experiments that, in the best of cases, culminate in the publication of results in journals and conference proceedings. Rather than stopping at the publication stage, the challenge for precision medicine is then to translate all of these research results into better treatments and improved health. To achieve this goal, a range of analytic methods and computational approaches have evolved from other domains and have been applied to an ever-growing set of specific problem areas. It would be impossible to enumerate the numerous biological questions targeted by computational approaches. We will focus here on an overview of text and data mining methods and their applications to discovery in a broad range of biomedical areas, including biological pathway extraction and reasoning, gene prioritization, precision medicine, pharmacogenomics and toxicology. The advances are plenty and the specific areas of application diverse, but the fundamental

Graciela H. Gonzalez is an Associate Professor in the Department of Biomedical Informatics at Arizona State University, Scottsdale, Arizona, United States.
Tasnia Tahsin is a PhD student in the Department of Biomedical Informatics at Arizona State University, Scottsdale, Arizona, United States.
Britton C. Goodale is a postdoctoral fellow in the Department of Microbiology and Immunology at the Geisel School of Medicine at Dartmouth College, Hanover, New Hampshire, United States.
Anna C. Greene is the Assistant Curriculum Director for the Graduate Program in Quantitative Biomedical Sciences at Dartmouth College, Hanover, New Hampshire, United States.
Casey S. Greene is an Assistant Professor in the Department of Systems Pharmacology and Translational Therapeutics in the Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, United States.
Submitted: 17 February 2015; **Received (in revised form):** 26 August 2015

motivation is to aid scientists in analyzing available data to suggest a road to discovery, to precise predictions that lead to better health.

Background

Data mining

Data mining is the act of computationally extracting new information from large amounts of data [1], and the biological sciences are generating enormous quantities of data, ushering in the era of 'big data'. Stephens *et al.* state that sequencing data alone constitutes ~35 petabases/year and will grow to ~1 zettabase/year by 2025 [2]. This creates a large opportunity for the development and deployment of novel mining algorithms, and two recent reviews on data and text mining in the era of big data are found in Che *et al.* [3] and Herland *et al.* [4]. A wide variety of methods for extracting value from different types and models of data fall under the umbrella of 'data mining'. Classification algorithms (decision trees, naïve Bayesian classification and other classifiers), frequent pattern algorithms (association rule mining, sequential pattern mining and others), clustering algorithms (including methods to cluster continuous and categorical data) and graph and network algorithms have all evolved to present a diverse landscape for research and an arsenal to deploy against the toughest data challenges. Most researchers consider some other areas, including text mining, as being under the data mining umbrella. For example, Piatetsky-Shapiro state: 'Data Mining in my opinion includes: text mining, image mining, web mining, predictive analytics, and much of the techniques we use for dealing with massive data sets, now known as Big Data' [5]. The methods applied to text mining, however, are specialized to such a degree that it is common to view it as a separate area of specialty. Data mining courses do not usually include any text mining material, but rather there are separate courses dedicated to it, and the same applies to textbooks.

A complete coverage of data mining techniques is beyond the scope of this article though we have included some important resources that cover this topic. *Kernel Methods in Computational Biology* by Schölkopf, Tsuda and Vert [6] covers methods specific to Computational Biology. *Introduction to Data Mining* [7] and *Data Mining: Concepts and Techniques, 3rd edn* [8] are two popular textbooks in data mining and give an excellent overview of the field. A more concise presentation can be found in the paper by Xindong Wu *et al.*, *Top 10 algorithms in data mining* [9], which were identified in December 2006 as C4.5, *k*-Means, SVM, Apriori, EM, PageRank, AdaBoost, *k*NN, Naive Bayes and CART, covering clustering, classification and association analysis, which are among the most important topics in data mining research:

- According to Jain *et al.* in 'Data clustering: a review', 'Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters)' [10].
- Classification is akin to clustering because it segments data into groups called classes, but unlike clustering, classification analyses require knowledge and specification of how classes are defined.
- Statistical learning theory seeks 'to provide a framework for studying the problem of inference that is, of gaining knowledge, making predictions, making decisions or constructing models from a set of data' states Bousquet *et al.* [11]. A textbook on statistical learning expands on these notions [12].
- Association analysis facilitates the unmasking of hidden relationships in large data sets. The discovered associations are then expressed as rules or sets of items that frequently occur together. Challenges to association analysis methods include that discovering such patterns can be computationally expensive given a large input data set and that there could potentially be many spurious associations 'discovered' that simply occur by chance. A well-known introduction to the topic is found in [13], and in particular, a seminal paper on mining association rules from clinical databases is found in Stilou *et al.* [14].
- Link analysis analyzes hyperlinks and the graph structure of the Web for the ranking of web search results. PageRank is perhaps the best-known algorithm for link analysis [15].

In a notable transition showing the power of new algorithms and data, data mining approaches are now being used to learn, not just the primary features but also context-specific features. For example, initial data mining approaches that constructed gene–gene networks built a single network [16]. In contrast, recent approaches learn multiple context-specific networks, allowing the construction of process-specific [17] and tissue-specific networks [18–20]. An individual is made up of a personalized combination of such context-specific networks, so we anticipate that continued advances in the context specificity of data mining approach will play an important role in the broad implementation of precision medicine.

Text mining

Text mining is a subfield of data mining that seeks to extract valuable new information from unstructured (or semi-structured) sources [21]. Text mining extracts information from within those documents and aggregates the extracted pieces over the entire collection of source documents to uncover or derive new information. This is the preferred view of the field that allows one to distinguish *text mining* from *natural language processing* (NLP) [22, 23]. Thus, given as input a set of documents, text mining methods seek to discover novel patterns, relationships and trends contained within the documents. Aiding the overall goal of discovering new information are NLP programs that go from the relatively simple text processing tasks at the lexical or grammatical levels (such as a tokenizing or a part-of-speech tagger), to relatively complex information extraction algorithms [like named entity recognition (NER) to find concepts such as genes or diseases, normalization to map them to their unique identifiers or relationship extraction and sentiment analysis systems, among others]. The greater the complexity of the task, the more likely it is to integrate methods from data mining (such as classification or statistical learning).

Although there is no current textbook that can be considered the definite guide on text mining as defined above, there are a couple of classic textbooks that cover fundamental NLP techniques and at least the first covers some of the analytics required to discover information: *Speech and Language Processing* by Jurafsky and Martin [24] and *Foundations of Statistical Natural Language Processing* by Manning and Schuetze [25]. The biomedical domain is one of the most interesting application areas for text mining, given both the potential impact of the information that can be discovered and the specific characteristics and volume of information available. The textbook *Text mining for biology and medicine* [26] offers an overview of the fundamental approaches to biomedical NLP, emphasizing different sub-areas in each chapter, although overall it does not totally adhere to the definition of text mining as a means for discovery given by

Hearst [23]. A good non-textbook review of the different subareas is the article 'Frontiers of biomedical text mining: current progress' [27]. For those just starting in the area, the article 'Getting Started in Text Mining' [28] is a good starting point. A more in-depth treatment of automated techniques applied to the biomedical literature and its contribution to innovative biomedical research can be found in 'Text-mining solutions for biomedical research: enabling integrative biology' [29].

Text mining sub-areas, briefly summarized, include:

- Information Retrieval deals with the problem of finding relevant documents in response to a specific information need (query). An overview of tools for information retrieval from the biomedical literature can be found in [30].
- NER is at the core of the automatic extraction of information from text and deals with the problem of finding references to entities (mentions) such as genes, drugs and diseases present in natural language text and tagging them with their location and type. NER is also referred to as 'entity tagging' or 'concept extraction'. This is a basic building block for almost all other extraction tasks. NER in the biomedical domain is generally considered to be more difficult than other domains, such as geography or news reports. This is owing to inconsistency in how known entities, such as symptoms or drugs, are named (e.g. nonstandard abbreviations and new ways of referring to them). An open-source NER engine, BANNER [31], with models to recognize genes and diseases mentioned in biomedical text, is currently available for gene and disease NER, and LINNAEUS is available for species [32]. Rebholz-Schuhmann *et al.* [33] present an overview of the NER solutions for the second CALBC task, including protein, disease, chemical (drug) and species entities. Campos *et al.* [34] discuss a recent survey of tools for biomedical NER. A system assigning text to a wide range of semantic classes using linguistic rules is presented in [35], illustrating a slightly different than standard NER because classes potentially overlap. Verspoor *et al.* [36] use the CRAFT corpus to improve the evaluation of gene NER (and some lower-level tasks like part-of-speech and sentence segmentation). Recent work in [37] presents an NER system for extracting gene and protein sequence variants from the biomedical literature. For locating chemical compounds, Krallinger *et al.* [38] summarize the task that was part of BioCreative IV and give a short overview of some of the techniques used.
- Named Entity Identification allows the linkage of objects of interest, such as genes, to information that is not detailed in a publication (such as their Entrez Gene identifier) [39]. Two open-source systems using largely dictionary-based approaches to normalize gene names appear in [39–41]. For normalizing disease names, [42] introduces DNorm, a new normalization framework using machine learning, with strong results.
- Association extraction is one of the higher-level tasks still considered purely an information extraction application. It uses the output from the prior subtasks to produce a list of (binary or higher) associations among the different entities of interest. Catalysts for advances in this area have been the Biocreative and BioNLP shared tasks, with excellent teams from around the world putting their systems to the test against carefully annotated data sets. A survey of submissions to Biocreative III [43] and BioNLP [44, 45] shows a good overview of approaches responsive to the respective shared tasks. Putting together associations into networks of molecular interactions that can explain complex biological processes is the next logical step, and one that still is considered the 'holy grail' of automatic biomolecular extraction. Ananiadou *et al.* [46] and Li *et al.* [47] discuss comprehensive surveys of methods for the extraction of network information from the scientific literature and the evaluation of extraction methods against reference corpora. Semantic-based approaches such as [48] will make their mark in the coming years.
- Event extraction is similar to association extraction but instead of separately extracting various relations between different entities in text, this task focuses on identifying specific events and the various players involved in it (arguments). For instance, the arguments of a transport event will include the molecule being transported, the cell to which it is being transported and the cell from which it is being transported. Event extraction was a key component of the BioNLP Shared Tasks in both 2011 [45] and 2013 [49], challenging the biomedical community to expand and cultivate their approaches in this area and leading to steadily improving results.
- Pathway extraction is a budding branch of biomedical text mining closely following the footsteps of event extraction. It involves the automated construction of biological pathways through the extraction and ordering of pathway-related events from text. Although, like [50] and [51], the majority of researchers in this domain have been focusing their efforts on supporting pathway curation through event extraction, rather than entirely automating the process. Tari *et al.* was able to achieve promising results for the automated synthesis of pharmacokinetic pathways by applying an automated reasoning-based approach for event ordering [52]. The first shared task on Pathway Curation was organized by BioNLP in 2013 [49] to establish the current state-of-the-art performance level for extracting pathway-relevant events such as phosphorylation and transport.

In the end, a set of the different subtask solutions are used in a pipeline that allows information to be integrated and analyzed toward knowledge discovery. However, this multiplies the effects of errors down the pipeline, leaving systems highly vulnerable.

An overarching challenge for biomedical text mining is to incorporate the many knowledge resources that are available to us into the NLP pipeline. In the biomedical domain, unlike the general text mining domain, we have access to large numbers of extensive, well-curated ontologies and knowledge bases. Biomedical ontologies provide an explicit characterization of a given domain of interest. The quality of data mining efforts would likely increase if existing ontologies (e.g. UMLS [53] and BioPortal [54]) were used as sources of terms in building lexicons, for figuring out what concept subsumes another, and as a way of normalizing alternative names to one identifier. For example, using ontologies as described enabled the use of unstructured clinical notes for generating practice-based evidence on the safety of a highly effective, generic drug for peripheral vascular disease [55].

Today, the data being generated is massive, complex and increasingly diverse owing to recent technological innovations. However, the impact of this data revolution on our lives is hampered by the limited amount of data that has been analyzed. This necessitates data mining tools and methods that can match the scale of the data and support timely decision-making through integration of multiple heterogeneous data sources.

Finally, another area in which the field has fallen short is that of making text mining applications that are easily adaptable by end users. Many researchers have developed systems that can be adapted by other text mining specialists,

but applications that can be tuned by bench scientists are mostly lacking.

Application areas

Pathway extraction and reasoning

Analyzing the intricate network of biological pathways is an essential precursor to understanding the molecular mechanisms of complex diseases affecting humans. Without acquiring a deeper insight into the underlying mechanisms behind such diseases, we cannot advance in our efforts to design effective solutions for preventing and treating them. However, given the vast amount of data currently available on biological pathways in biomedical publications and databases and the highly interconnected nature of these pathways, any attempt to manually reason over them will invariably prove to be largely ineffective and inefficient. As a result, there is a growing need for computational approaches to address this demanding task through automated pathway analysis. Pathway analysis can be either quantitative or qualitative and is a key focus of the growing field of Systems Biology. Quantitative pathway analysis uses dynamic mathematical models for simulating pathways and can be especially useful in drug discovery and the development of patient-specific dosage guidelines [56]. Some examples of techniques used in this form of analysis include ordinary differential equations [57], Petri Nets [58], and π-calculus [59]. Qualitative pathway analysis uses static, structural representations of pathways to answer qualitative questions about them; for instance it may be used to explain why a certain phenomenon occurs in the pathway based on existing pathway knowledge. Artificial intelligence paradigms, such as symbolic (i.e. explicit representations) or connectionist (i.e. massively parallelized) approaches, can greatly inform this type of pathway analysis [60]. Although some of the techniques principally addressing quantitative pathway analysis, such as Petri Nets and π-calculus, may also be used to perform qualitative pathway analysis, they typically tend to provide limited functionality [61]. Therefore, richer languages such as Maude [62], BioCham [63] and action languages [52, 64, 65] are more popular in this domain. In recent years, hybrid approaches have been applied for qualitative pathway reasoning. For instance, [66] presents a qualitative pathway reasoning system that uses Petri net semantics as the pathway specification language and action languages as the query language. Pathway reasoning, as a technique, relies on either humans defining the pathway information needed or the development of new algorithms to extract, represent and reason over biological pathways, which is an area of growing interest.

Gene prioritization and gene function prediction

Complex diseases present diverse symptoms because they are caused by multiple genes and environmental factors that differ for each individual and can diverge at different stages of the disease process. This complexity is reflective of epistatic effects where causative genes have an impact on the expression of many other genes. Because variant expression levels vary across the genome, it is difficult to determine true causative genes or distinguish key sets affected by the disease from high-throughput experiments. For example, the Affimetrix U133 Plus 2.0 microarray chip from the Repository of Molecular Brain Neoplasia Data shows >7500 2-fold differentially expressed genes in brain cancer tissue when compared with normal brain tissue [67]. The validation of a single causative gene is a long and expensive process [68], often taking up to a year and even longer, which necessitates using gene prioritization to pare down the list of potential gene targets to a manageable size. Gene prioritization methods that suggest the most significant prospects for further validation are critically needed, and method development in this area would greatly facilitate discovery.

Many gene prioritization algorithms have been developed to address this problem, such as GeneWanderer [69], GeneSeeker [70], GeneProspector [71], SUSPECTS [72], G2D [73] and Endeavour [74], among others [75, 76]. A comparative review of these methods can be found in Tranchevent *et al.* [77]. The general premise of these methods is to rank genes based on the similarity between a set of candidate genes compared with genes already known to be associated with the disease (usually called the training set). Similarity is established based on different parameters (depending on the specific method) and may include purely biological measures (such as cytogenetic location, expression patterns, patterns of pathogenic mutations or DNA sequence similarity), biological measures plus annotation of the genes using different protein databases (for example, UniProt [78] and InterPro [79]), or other vocabularies and ontologies (such as the Gene Ontology [80, 81], eVOC [82], MeSH [83] and term vectors from the literature). In these methods, the closer a gene in the candidate list coincides with the profile of the training genes, the higher it is ranked.

Gene prioritization includes the areas of gene function prediction. The Critical Assessment of protein Function Annotation experiment was the first large community-wide evaluation of 54 methods that were compared on a core set of annotations using evaluation metrics to ascertain the top methods [84]. Earlier computational methods for prioritization were compared through a large-scale biological assay of yeast mitochondrial phenotypes and found to be effective [85, 86]. A related but distinct gene prioritization problem is the identification of genes with tissue-specific expression patterns [87]. Existing webservers such as GeneMANIA [88, 89] and IMP [90] allow biologists to perform gene prioritization by network connectivity, and servers such as PILGRM allow for prioritization directly by gene expression [91]. Predicted functions, in addition to curated functions, have also shown promise for interpreting the results of genome-wide association studies, which aim to pair genetic variants with associated genes and pathways [92].

Precision medicine and drug repositioning

Precision medicine is determining prevention and treatment strategies based on an individual's predisposition in an effort to provide more targeted and therefore effective treatments [93]. This area is poised for intense growth based on the ease of obtaining patient data and the development of computational methods with which to analyze this personalized data. While precision medicine is a nascent field, there have been many advances in the personalized treatment of cancer. Some hospitals are already using genetic data to direct treatment options for cancer patients (e.g. BRCA1 and BRCA2 [94], BRAF [95] testing), though drugs targeted to specific mutations lag behind and is an area where computational drug repositioning will potentially have a strong impact [96].

On the clinical side of translational research, the demand for timely and accurate knowledge has the urgency of life itself. Emily Whitehead was the first child with acute lymphoblastic

leukemia to be treated and cured with an experimental T cell therapy called CAR T cell therapy at the Children's Hospital of Philadelphia [97]. The therapy enables the patient's T cells to recognize and attack malignant B cells, but this treatment can also trigger an intense immune reaction, which Emily experienced. She suffered from a high level of the interleukin 6 protein, and her doctors suggested trying tocilizumab (Actemra), a rheumatoid arthritis drug, to combat the extraneous protein production [97, 98]. This drug returned Emily's vital signs back to normal. In this case, rather than relying on the serendipity of a team member knowing about the right drug, specialized text mining could have been used to mine the literature for the relevant drugs. In such a scenario, either the literature would be mined in advance, stored in a database that extracts relationships between drugs and genes or proteins or it could be searched in real time. As an example of this, Essack *et al*. created a sickle cell disease knowledgebase by mining 419 612 PubMed abstracts related to red blood cells, anemia or this disease [99]. Some databases (such as PharmGKB) store such relationships, but are not the result of automatic extraction. Manual curation is still the current standard for such databases, with the value of text mining applications yet to be fully realized. Currently, despite notable advances in entity mention extraction and normalization, the use of text mining is mostly limited to aiding curators to speed up the process.

Data and text mining methods are useful for biomedical predictions and can be successfully extended to biomedical discoveries as well. Sirota *et al*. used publicly available gene expression data for both drugs and diseases to ascertain if Food and Drug Administration-approved drugs could be repositioned for use in new diseases [100]. They discovered and experimentally validated the use of cimetidine, generally used for heartburn and peptic ulcers, as a treatment option for lung adenocarcinoma illustrating the use of a computational approach as an efficient, yet powerful, approach to drug discovery [100, 101]. Frijters *et al*. successfully found links between genes, drugs, pathways and diseases through their tool CoPub Discovery that mines the biomedical literature for the elucidation of new relationships between these concepts [102]. Based on their predictions, they validated two different drugs' role in cell proliferation through a cell assay to illustrate the validity of their tool for finding novel associations. This tool may be useful in finding new connections between drugs and their targets, as well as the ability to repurpose drugs for disease treatment.

Data integration

Data integration represents a particularly important type of computational approach. Integrative analyses can identify patterns that are evident across many distinct experiments. Patterns from imperfectly matched experiments are likely to be general responses to a common environment as opposed to unique features of an experiment [103–106]. Integrative analyses, while they have substantial potential to identify general principles, also raise specific challenges, largely driven by potentially undesirable features of the data. For example, Huttenhower *et al*. [107] found that the mutual information between data sets was largely driven by the experimental platform and not relevant biological signals.

To address this challenge, many integrative methods use either carefully curated and selected data sets [100, 101, 108, 109] or supervised machine learning methods [19, 90, 110–118]. As an example of carefully selected data sets, Sirota *et al*. [100] used a labeled compendium of gene expression experiments of disease state and drug treatment to identify drugs that induced an expression profile that was anti-correlated with disease. In addition, gene expression values were analyzed using rank-based statistics, which may also mitigate platform-specific noise. Supervised analyses can mitigate the effects of technical artifacts by grading each data set by how much information each provides about different aspects of biology. Many methods have been successfully applied to this challenge including Bayesian [90] and ridge regression [118] approaches. For example, Greene *et al*. [19] used a Bayesian approach to weigh each of approximately 1000 data sets by how well they captured tissue-specific functional relationships. This approach produced tissue-specific networks for 144 human tissues, and networks generated by the tissue-specific Bayesian integration of the complete compendium outperformed an approach that integrated only tissue-specific data sets on both coverage of tissues and overall network metrics. To combat platform-specific signals, Greene *et al*. [19] calculated the mutual information across data sets for non-related pairs of genes to identify and down-weight data set similarity that was independent of biology.

In addition to approaches that rely on the curation of data sets or supervised methods, new techniques based on advances in deep learning are now also being applied to the challenge of data integration [119]. For example, Tan *et al*. [119] performed an analysis using denoising autoencoders of gene expression to extract features from a set of ~2000 breast cancer biopsies. In this approach, a neural network model is trained to reconstruct the observed data from data where noise has been added. The identified features corresponded to subtype, estrogen receptor status and other features that had a well-documented role in the biology of breast cancer. Of particular note, the features generalized to an independent data set generated on a distinct platform without a loss in accuracy, suggesting that the model had identified these biological features without overfitting to the platform. Unsupervised methods capable of identifying biological signals without confounding technical artifacts present substantial opportunities for new algorithms that integrate large-scale data compendia where the curated knowledge required by supervised algorithms is limited or unavailable.

Pharmacogenomics

The field of pharmacogenomics has benefitted significantly from recent progress in text and data mining for biomedical discoveries. Pharmacogenomics studies the genetic basis of individual drug responses by exploring the relationship between drugs, genes and diseases and analyzing pharmacokinetic and pharmacodynamic pathways. Pertinent pharmacogenomics-related information is typically extracted through the manual curation of data from pharmacogenomics literature and stored in the freely accessible PharmGKB database. However, the substantial level of advancement in the field of drug detection, gene detection and disease detection along with the increased efficacy of methods for the extraction of relations between drugs, genes and diseases has now made it possible to use automated systems to help with this curation process [120]. Two good reviews on pharmacogenetic text mining have been recently published by [121] and [122], respectively, while in 2014, Laiotaki *et al*. presented design specifications for building an integrated information system for offering personalized drug recommendations using genotype-to-phenotype knowledge on pharmacogenomics [123]. Every year the field of

pharmacogenomic text mining continues to expand in different novel directions, gradually turning the vision of personalized medicine into reality.

Toxicology

The field of toxicology has an increasing need for text and data mining approaches capable of predicting chemical–biological interactions of thousands of chemicals that humans are exposed to either intentionally (via pharmaceuticals, diet) or unintentionally (contaminated air, water and food). Substantial data are required to identify potential toxicological effects of each chemical, and for regulators, such as the US Environmental Protection Agency (EPA) and European Chemical Agency, to make decisions protective of human health. The EPA inventories chemicals in commerce under the Toxic Substances Control Act (TSCA) [124]. In a 2009 review, 'The toxicity data landscape for environmental chemicals', Judson *et al.* reported 75 000 chemicals in the TSCA database and identified 9912 chemicals under prioritization for testing by the EPA [125]. A lack of toxicity data limits the ability of regulators to make informed decisions and for health agencies to assess risk and respond in the case of exposure. Judson *et al.* further report that evaluation of almost 10 000 chemicals under the current testing paradigm would be both cost and time prohibitive, as *in vivo* studies require 2–3 years and millions of dollars per chemical [125]. The urgent need for more data has prompted development of *in vitro* high-content and High-Throughput Screening (HTS) methods to evaluate many biological endpoints relatively inexpensively. Multiple data mining approaches will be required to use these data to address the large knowledge gaps in toxicology. These include both broad analyses that leverage HTS and *in vivo* data across chemicals to predict biological effects of new compounds, as well as deeper analysis of genome-wide data sets at multiple levels of biological organization to predict how chemicals disrupt biological processes.

Regulatory agency research initiatives, combined with increasing use of HTS and high-content approaches by independent researchers, are rapidly expanding the universe of toxicological data available to the public. A vast array of data is currently being collected through Tox21, a multi-agency collaborative HTS effort to identify chemical–biological interactions and chemical concentrations that cause toxic effects [126]. The EPA ToxCast program is evaluating chemical toxicity with 700 biochemical and cell-based HTS assays and using this information to identify chemical signatures that predict potential toxicity and prioritize chemicals for further testing. In parallel with the expansion of pharmacogenomic approaches to pharmaceutical development, computational toxicology has used data mining to identify features of environmental chemicals that mediate activity leading to potential adverse effects. Ekins *et al.* reviewed quantitative structure activity relationship and machine learning models that have been developed to predict specific toxicity endpoints such as hepatotoxicity, cardiotoxicity and genotoxicity from HTS, molecule descriptor and literature data compilations [127]. Predictive power of models developed from the first ToxCast HTS data set (~300 chemicals) was limited, potentially because of a lack of redundant chemicals with positive signal to cover the array of mechanisms that lead to toxic effects *in vivo* or the large chemical space between training and test data sets [128, 129]. Prediction of whole animal toxicity encompassing diverse biological endpoints presents a particular challenge because a chemical can disrupt multiple molecular pathways and have different effects depending on the biological context. For example, 2,3,7,8-tetrachlorodibenzo-p-dioxin is an activator of the aryl hydrocarbon receptor pathway and tumor promoter [130]. Exposure during early development, however, leads to developmental abnormalities, including heart defects [131]. Accurate descriptors and classification of chemicals in training sets is essential, but depends on rich data sets as well as knowledge of biological pathways [132].

Transcriptomic, proteomic and metabolomics studies in the context of chemical exposures are beginning to provide biological pathway information that is critical to understanding mechanisms of chemical-induced toxicity. Several projects, lead by Tox21 collaborators and others, aim to identify and classify signals of chemical exposure from transcriptome data [133]. Gusenleitner *et al.* used the National Toxicology Program DrugMatrix and the TG-Gates (Toxico genomics project-Genomics Assisted Toxicity Evaluations) databases, which contain over 5000 arrays of rat tissues and primary rat hepatocytes exposed to therapeutic, industrial and environmental chemicals, to develop a predictive model of genotoxicity from *in vitro* data [133]. An analysis of the same data by Tawa *et al.* identified gene modules associated with liver toxicity [134]. The ability to associate gene modules discovered in other tissues in these data sets with toxicological endpoints is limited by the available clinical pathology and histology annotation. Context-specific algorithms and unsupervised methods therefore have the potential to make great contributions to the field of toxicogenomics.

In parallel with the expansion of pharmacogenomic approaches to improve the development of pharmaceuticals and personalized medicine, computational toxicology has used data mining to identify features of environmental chemicals that mediate activity and cause potential toxicity. Several efforts aimed at literature-based chemical annotation are underway, including the Comparative Toxicogenomic Database, which leverages text mining and manual curation to provide chemical–gene–disease interaction data [135]. Accurate classification of new chemicals depends on comprehensive annotation of previously studied chemicals with toxicity information. Data mining across biological contexts will identify the chemical–pathway interactions that increase sensitivity of certain individuals, such as the young or populations with particular genetic polymorphisms, to complex chemical/stressor exposures.

Engagement of the broader scientific community is important for addressing the challenges of computational toxicology. With the release of data from 1800 Toxcast chemicals in 2013, the EPA hosted a series of challenges focused on method development for chemical lowest effect level prediction from HTS data [136]. Data mining tools and methods that can integrate vast amounts of heterogeneous data will be needed to prioritize genes, pathways and chemicals for further investigation. A key component to the success of computational approaches in toxicology will be validation of model predictions by scientists at the bench. Centralized model repositories, databases such as the Comparative Toxicogenomics Database and the admetSAR structure activity database [137] and web-based analysis tools are essential to facilitate research community access and leverage existing data to inform future *in vitro* and *in vivo* toxicology research.

Conclusion

We have reviewed recent advances in text and data mining in the context of emerging application domains in the biomedical

sciences. Computational methods contribute to this field by bringing knowledge from literature, either extracted or curated, together with high-throughput data sets to identify both known and new relationships between genes, pathways, drugs, environmental contaminants and diseases. Different approaches are often used for mining unstructured text and structured biomedical data. For this reason, integrating across both unstructured and structured resources presents additional challenges, but combining these domains will also present new opportunities. Systems that can extract relationships from both literature and data simultaneously present the opportunity to identify meaningful patterns from data, identify literature support for those patterns, and where warranted, identify relationships that are highly consistent in large-scale throughput data sets but absent from literature. This presents the opportunity to develop computational algorithms that not only identify biological principles but also recognize when those principles may represent novel discoveries.

Key Points

1. The era of 'big data' presents biomedical researchers unprecedented challenges and opportunities for discovery.
2. Automatic methods for text and data mining are essential tools that need to be deployed to deal with large data sets of highly heterogeneous, but complimentary, data.
3. Key advances in data and text mining will empower bench scientists rather than replace them.
4. A major challenge in the big data era for text and data mining is the integration of different sources such as curated databases, biomedical literature and results from assays to answer questions or generate novel hypotheses.

Funding

This work was supported by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative [GBMF4552 to C.S.G.] and the National Institute of Environmental Health Science [Ruth L. Kirschstein Postdoctoral Fellowship Award F32ES025082 to B.C.G.] and the National Library of Medicine [R01LM011176 to G.H.G.]

References

1. Witten IH, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. San Francisco: Morgan Kaufmann Publishers Inc, 2005.
2. Stephens ZD, Lee SY, Faghri F, *et al*. Big data: astronomical or genomical? *PLOS Biol* 2015;**13**:e1002195.
3. Che D, Safran M, Peng Z. From big data to big data mining: challenges, issues, and opportunities. *Database Syst Adv Appl Lect Notes Comput Sci* 2013;**7827**:1–15.
4. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. *J Big Data* 2014;**1**:2.
5. Piatetsky-Shapiro G. The journey of knowledge discovery. In: M Gaber (ed). *Journeys to Data Mining*. Germany: Springer-Verlag Berlin Heidelberg, 2012, 173–96.
6. Schölkopf B, Tsuda K, Vert J-P. *Kernel Methods in Computational Biology*. Cambridge: The MIT Press, 2004.
7. Tan PN, Steinbach M, Kumar V. *Introduction to Data Mining*. Pearson Education, 2007.
8. Han J, Kamber M, Pei J. *Data Mining: Concepts and Techniques*. Waltham: Morgan Kaufmann, 2012.
9. Wu X, Kumar V, Ross Quinlan J, *et al*. Top 10 algorithms in data mining. *Knowl Inf Syst* 2007;**14**:1–37.
10. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. *ACM Comput Surv* 1999;**31**:264–323.
11. Bousquet O, Boucheron S, Lugosi G. Introduction to statistical learning. In: *Advanced Lectures on Machine Learning*. Germany: Springer, Berlin Heidelberg, 2004, 169–207.
12. Friedman J, Hastie T, Tibshirani R. *The Elements of Statistical Learning*. New York, NY: Springer-Verlag, 2001.
13. Tan PN, Steinbach M, Kumar V. Association analysis: basic concepts and algorithmstitle. In: *Introduction to Data Mining*. Pearson Education, 2007;327–414.
14. Stilou S, Bamidis PD, Maglaveras N, *et al*. Mining association rules from clinical databases: an intelligent diagnostic process in healthcare. *Stud Health Technol Inform* 2001;**84**:1399–403.
15. Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*. New York: Cambridge University Press, 2008.
16. Lee I, Date S V, Adai AT, *et al*. A probabilistic functional network of yeast genes. *Science* 2004;**306**:1555–8.
17. Myers CL, Troyanskaya OG. Context-sensitive data integration and prediction of biological networks. *Bioinformatics* 2007;**23**:2322–30.
18. Goya J, Wong AK, Yao V, *et al*. FNTM: a server for predicting functional networks of tissues in mouse. *Nucleic Acids Res* 2015;**43**:W182–7.
19. Greene CS, Krishnan A, Wong AK, *et al*. Understanding multicellular function and disease with human tissue-specific networks. *Nat Genet* 2015;**47**:569–76.
20. Chowdhary R, Tan SL, Zhang J, *et al*. Context-specific protein network miner—an online system for exploring context-specific protein interaction networks from the literature. *PLoS One* 2012;**7**:e34480.
21. Feldman R, Sanger J. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2007.
22. Hearst MA. Untangling text data mining. In: *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*, 1999;3–10.
23. Hearst M. What is text mining? http://people.ischool.berkeley.edu/~hearst/text-mining.html (26 July 2015, date last accessed).
24. Jurafsky D, Martin JH. *Speech and Language Processing*, 2nd edn. Upper Saddle River: Pearson, 2009.
25. Manning CD, Schütze H. *Foundations of Statistical Natural Language Processing*. *Cambridge: MIT Press*, 1999.
26. Ananiadou S, Mcnaught J. *Text Mining for Biology And Biomedicine*. Norwood: Artech House, Inc., 2005.
27. Zweigenbaum P, Demner-Fushman D, Yu H, *et al*. Frontiers of biomedical text mining: current progress. *Brief Bioinform* 2007;**8**:358–75.
28. Cohen KB, Hunter L. Getting started in text mining. *PLoS Comput Biol* 2008;**4**:e20.
29. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. *Nat Rev Genet* 2012;**13**:829–39.
30. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database (Oxford)* 2011;**2011**:baq036.
31. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. *Pac Symp Biocomput* 2008;652–63.

32. Gerner M, Nenadic G, Bergman CM. LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinformatics* 2010;**11**:85.
33. Rebholz-Schuhmann D, Jimeno Yepes A, Li C, *et al*. Assessment of NER solutions against the first and second CALBC silver standard corpus. *J Biomed Semantics* 2011;**2** (Suppl 5):S11.
34. Campos D, Matos S, Oliveira JL. Biomedical named entity recognition: a survey of machine-learning tools. In: Sakurai S (ed). *Theory and Applications for Advanced Text Mining*. InTech Open Access, 2012.
35. Cohen KB, Christiansen T, Baumgartner WA Jr, *et al*. Fast and simple semantic class assignment for biomedical text. In: *Proceedings of the 2011 Workshop on Biomedical Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics (ACL)-HLT 2011, 2011;38–45.
36. Verspoor K, Cohen KB, Lanfranchi A, *et al*. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics* 2012;**13**:207.
37. Wei C-H, Harris BR, Kao H-Y, *et al*. tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics* 2013;**29**:1433–9.
38. Krallinger M, Leitner F, Rabal O, *et al*. Overview of the chemical compound and drug name recognition (CHEMDNER) task. In: *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*. Bethesda, MD, USA, 2013;2:2–33.
39. Hakenberg J, Plake C, Leaman R, *et al*. Inter-species normalization of gene mentions with GNAT. *Bioinformatics* 2008;**24**:i126–32.
40. Hakenberg J, Gerner M, Haeussler M, *et al*. The GNAT library for local and remote gene mention normalization. *Bioinformatics* 2011;**27**:2769–71.
41. Wermter J, Tomanek K, Hahn U. High-performance gene name normalization with GENO. *Bioinformatics* 2009;**25**:815–21.
42. Leaman R, Islamaj Dogan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics* 2013;**29**:2909–17.
43. Krallinger M, Vazquez M, Leitner F, *et al*. The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC Bioinformatics* 2011;**12** (Suppl 8):S3.
44. Kim J-D, Ohta T, Pyysalo S, *et al*. Overview of BioNLP'09 shared task on event extraction. In: *Proceedings of the BioNLP Shared Task 2009 Workshop*. Madison, WI: Omnipress, Inc, 2009;1–9.
45. Kim J-D, Pyysalo S, Ohta T, *et al*. Overview of BioNLP shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*. Madison, WI: Omnipress, Inc, 2011;1–6.
46. Ananiadou S, Pyysalo S, Tsujii J, *et al*. Event extraction for systems biology by text mining the literature. *Trends Biotechnol* 2010;**28**:381–90.
47. Li C, Liakata M, Rebholz-Schuhmann D. Biological network extraction from scientific literature: state of the art and challenges. *Brief Bioinform* 2014;**15**:856–77.
48. Faiz R, Amami M, Elkhlifi A. Semantic event extraction from biological texts using a Kernel-based method. *Adv Knowl Discov Manag* 2014;**527**:77–94.
49. Nédellec C, Bossy R, Kim J-D, *et al*. Overview of BioNLP shared task 2013. In: *Proceedings of the BioNLP Shared Task 2013 Workshop*. Madison, WI: Omnipress, Inc, 2013;1–7.
50. Ohta T, Pyysalo S, Ananiadou S, *et al*. Pathway curation support as an information extraction task. In: *Proceedings of the Fourth International Symposium on Languages in Biology and Medicine (LBM 2011)*. Singapore: LBM, http://www.nactem.a-c.uk/papers/Ohta_LBM_2011.pdf.
51. Ohta T, Pyysalo S, Rak R, *et al*. Overview of the pathway curation (PC) task of BioNLP shared task 2013. In: *Proceedings of BioNLP Shared Task 2013 Workshop*. Madison, WI: Omnipress, Inc, 2013;67–75.
52. Tari L, Anwar S, Liang S, *et al*. Synthesis of pharmacokinetic pathways through knowledge acquisition and automated reasoning. *Pac Symp Biocomput* 2010;465–76.
53. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:267D–270.
54. Whetzel PL, Noy NF, Shah NH, *et al*. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res* 2011;**39**:W541–5.
55. Leeper NJ, Bauer-Mehren A, Iyer S V, *et al*. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. *PLoS One* 2013;**8**:e63499.
56. Materi W, Wishart DS. Computational systems biology in drug discovery and development: methods and applications. *Drug Discov Today* 2007;**12**:295–303.
57. Hoops S, Sahle S, Gauges R, *et al*. COPASI—a complex pathway simulator. *Bioinformatics* 2006;**22**:3067–74.
58. Pinney J, Westhead D, McConkey G. Petri Net representations in systems biology. *Biochem Soc Trans* 2003;1513–5.
59. Curti M, Degano P, Priami C, *et al*. Modelling biochemical pathways through enhanced π-calculus. *Theor Comput Sci* 2004;**325**:111–40.
60. Smolensky P. Connectionist AI, symbolic AI, and the brain. *Artif Intell Rev* 1987;**1**:95–109.
61. Salim Khan, Keith Decker, William Gillis CS. A multi-agent system-driven ai planning approach to biological pathway discovery. *ICAPS-03* 2003;246–55.
62. Eker S, Knapp M, Laderoute K, *et al*. Pathway logic: executable models of biological networks. *Electron Notes Theor Comput Sci* 2004;**71**:144–61.
63. Fages F. Modelling and querying interaction networks in the biochemical abstract machine BIOCHAM. *J Biol Phys Chem* 2002;**4**:64–73.
64. Grell S, Schaub T, Selbig J. Modelling biological networks by action languages via answer set programming. *Constraints* 2006;**4079**:285–99.
65. Tari L, Hakenberg J, Gonzalez G, *et al*. Querying parse tree database of Medline text to synthesize user-specific biomolecular networks. *Pac Symp Biocomput* 2009;87–98.
66. Anwar S, Baral C. Pathway specification and comparative queries: a high level language with petri net semantics. In: *Proceedings of Twenty-Eighth AAAI Conference on Artificial Intelligence*. QC, Canada: AAAI, 2014;981–8.
67. Madhavan S, Zenklusen J-C, Kotliarov Y, *et al*. Rembrandt: helping personalized medicine become a reality through integrative translational research. *Mol Cancer Res* 2009;**7**:157–67.
68. Page GP, George V, Go RC, *et al*. 'Are we there yet?' Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. *Am J Hum Genet* 2003;**73**:711–9.
69. Köhler S, Bauer S, Horn D, *et al*. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 2008;**82**:949–58.
70. Van Driel MA, Cuelenaere K, Kemmeren PPCW, *et al*. GeneSeeker: extraction and integration of human

disease-related information from web-based genetic databases. *Nucleic Acids Res* 2005;**33**:W758–61.
71. Yu W, Wulf A, Liu T, *et al*. Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. *BMC Bioinformatics* 2008;**9**:528.
72. Adie EA, Adams RR, Evans KL, *et al*. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics* 2005;**6**:55.
73. Perez-Iratxeta C, Bork P, Andrade-Navarro MA. Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Res* 2007;**35**:W212–6.
74. Tranchevent L-C, Barriot R, Yu S, *et al*. ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res* 2008;**36**:W377–84.
75. Tiffin N, Kelso JF, Powell AR, *et al*. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res* 2005;**33**:1544–52.
76. Lombard Z, Tiffin N, Hofmann O, *et al*. Computational selection and prioritization of candidate genes for fetal alcohol syndrome. *BMC Genomics* 2007;**8**:389.
77. Tranchevent L-C, Capdevila FB, Nitsch D, *et al*. A guide to web tools to prioritize candidate genes. *Brief Bioinform* 2011;**12**:22–32.
78. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res* 2014;**43**:D204–12.
79. Mitchell A, Chang H-Y, Daugherty L, *et al*. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res* 2014;**43**:D213–21.
80. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.
81. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res* 2014;**43**:D1049–56.
82. Kelso J, Visagie J, Theiler G, *et al*. eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res* 2003;**13**:1222–30.
83. Lipscomb CE. Medical Subject Headings (MeSH). *Bull Med Libr Assoc* 2000;**88**:265–6.
84. Radivojac P, Clark WT, Oron TR, *et al*. A large-scale evaluation of computational protein function prediction. *Nat Methods* 2013;**10**:221–7.
85. Hess DC, Myers CL, Huttenhower C, *et al*. Computationally driven, quantitative experiments discover genes required for mitochondrial biogenesis. *PLoS Genet* 2009;**5**:e1000407.
86. Hibbs MA, Myers CL, Huttenhower C, *et al*. Directing experimental biology: a case study in mitochondrial biogenesis. *PLoS Comput Biol* 2009;**5**:e1000322.
87. Ju W, Greene CS, Eichinger F, *et al*. Defining cell-type specificity at the transcriptional level in human disease. *Genome Res* 2013;**23**:1862–73.
88. Warde-Farley D, Donaldson SL, Comes O, *et al*. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* 2010;**38**:W214–20.
89. Zuberi K, Franz M, Rodriguez H, *et al*. GeneMANIA prediction server 2013 update. *Nucleic Acids Res* 2013;**41**:W115–22.
90. Wong AK, Park CY, Greene CS, *et al*. IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic Acids Res* 2012;**40**:W484–90.
91. Greene CS, Troyanskaya OG. PILGRM: an interactive data-driven discovery platform for expert biologists. *Nucleic Acids Res* 2011;**39**:W368–74.
92. Pers TH, Karjalainen JM, Chan Y, *et al*. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* 2015;**6**:5890.
93. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015;**372**:793–5.
94. Levy-Lahad E, Lahad A, King M-C. Precision medicine meets public health: population screening for BRCA1 and BRCA2. *J Natl Cancer Inst* 2015;**107**:420.
95. Gonzalez D, Fearfield L, Nathan P, *et al*. BRAF mutation testing algorithm for vemurafenib treatment in melanoma: recommendations from an expert panel. *Br J Dermatol* 2013;**168**:700–7.
96. American Cancer Society. *Personalized Cancer Care: Where it Stands Today*. http://www.cancer.org/research/acsresearchupdates/more/personalized-cancer-care-where-it-stands-today (22 July 2015, date last accessed).
97. The Children's Hospital of Philadelphia. *Relapsed Leukemia: Emily's Story*. http://www.chop.edu/stories/relapsed-leukemia-emilys-story#.Va_9QUV9Tlc (22 July 2015, date last accessed).
98. The New York Times. *A Breakthrough Against Leukemia Using Altered T-Cells*. http://www.nytimes.com/2012/12/10/health/a-breakthrough-against-leukemia-using-altered-t-cells.html?_r=0 (22 July 2015, date last accessed).
99. Essack M, Radovanovic A, Bajic VB. Information exploration system for sickle cell disease and repurposing of hydroxyfasudil. *PLoS One* 2013;**8**:e65190.
100. Sirota M, Dudley JT, Kim J, *et al*. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci Transl Med* 2011;**3**:96ra77.
101. Dudley JT, Sirota M, Shenoy M, *et al*. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci Transl Med* 2011;**3**:96ra76.
102. 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**:e1000943.
103. Farley JU, Lehmann DR, Ryan MJ. Generalizing from 'imperfect' replication. *J Bus* 1981;**54**:597–610.
104. Würbel H. Behaviour and the standardization fallacy. *Nat Genet* 2000;**26**:263.
105. Richter SH, Garner JP, Würbel H. Environmental standardization: cure or cause of poor reproducibility in animal experiments? *Nat Methods* 2009;**6**:257–61.
106. Richter SH, Garner JP, Zipser B, *et al*. Effect of population heterogenization on the reproducibility of mouse behavior: a multi-laboratory study. *PLoS One* 2011;**6**:e16461.
107. Huttenhower C, Haley EM, Hibbs MA, *et al*. Exploring the human genome with functional maps. *Genome Res* 2009;**19**:1093–106.
108. Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. *Nat Biotechnol* 2006;**24**:55–62.
109. Kong X, Mas V, Archer KJ. A non-parametric meta-analysis approach for combining independent microarray datasets: application using two microarray datasets pertaining to chronic allograft nephropathy. *BMC Genomics* 2008;**9**:98.
110. Myers CL, Robson D, Wible A, *et al*. Discovery of biological networks from diverse functional genomic data. *Genome Biol* 2005;**6**:R114.

111. English SB, Butte AJ. Evaluation and integration of 49 genome-wide experiments and the prediction of previously unknown obesity-related genes. *Bioinformatics* 2007;**23**:2910–7.
112. Parrish JR, Yu J, Liu G, *et al*. A proteome-wide protein interaction map for Campylobacter jejuni. *Genome Biol* 2007;**8**:R130.
113. Lee I, Li Z, Marcotte EM. An improved, bias-reduced probabilistic functional gene network of baker's yeast, Saccharomyces cerevisiae. *PLoS One* 2007;**2**:e988.
114. Guan Y, Myers CL, Lu R, *et al*. A genomewide functional network for the laboratory mouse. *PLoS Comput Biol* 2008;**4**:e1000165.
115. Lee I, Lehner B, Vavouri T, *et al*. Predicting genetic modifier loci using functional gene networks. *Genome Res* 2010;**20**:1143–53.
116. Park CY, Wong AK, Greene CS, *et al*. Functional knowledge transfer for high-accuracy prediction of understudied biological processes. *PLoS Comput Biol* 2013;**9**:e1002957.
117. Tan J, Grant GD, Whitfield ML, *et al*. Time-point specific weighting improves coexpression networks from time-course experiments. *Evol Comput Mach Learn Data Min Bioinform* 2013;**7833**:11–22.
118. Mostafavi S, Ray D, Warde-Farley D, *et al*. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol* 2008;**9** (Suppl 1):S4.
119. Tan J, Ung M, Cheng C, *et al*. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. *Pacific Symp Biocomput* 2015;132–43.
120. Sangkuhl K, Berlin DS, Altman RB, *et al*. PharmGKB: understanding the effects of individual genetic variants. *Drug Metab Rev* 2008;**40**:539–51.
121. Garten Y, Coulet A, Altman RB. Recent progress in automatically extracting information from the pharmacogenomic literature. *Pharmacogenomics* 2010;**11**:1467–89.
122. Hahn U, Cohen KB, Garten Y, *et al*. Mining the pharmacogenomics literature—a survey of the state of the art. *Brief Bioinform* 2012;**13**:460–94.
123. Lakiotaki K, Patrinos GP, Potamias G. Information technology meets pharmacogenomics: Design specifications of an integrated personalized pharmacogenomics information system. *In: IEEE-EMBS Internationnal Conference on Biomedical and Health Informatics,* 2014;13–16.
124. US EPA. *How to Access the Inventory, TSCA Chemical Substance Inventory*. http://www.epa.gov/opptintr/existingchemicals/pubs/tscainventory/howto.html. (23 July 2015, date last accessed).
125. Judson R, Richard A, Dix DJ, *et al*. The toxicity data landscape for environmental chemicals. *Environ Health Perspect* 2009;**117**:685–95.
126. Agents NRC (U. S. and C on TT and A of E). *Toxicity Testing in the 21st Century: A Vision and a Strategy*. Washington, DC: National Academies Press, 2007.
127. Ekins S. Progress in computational toxicology. *J Pharmacol Toxicol Methods* 2014;**69**:115–40.
128. Judson R, Elloumi F, Setzer RW, *et al*. A comparison of machine learning algorithms for chemical toxicity classification using a simulated multi-scale data model. *BMC Bioinformatics* 2008;**9**:241.
129. Thomas RS, Black MB, Li L, *et al*. A comprehensive statistical analysis of predicting in vivo hazard using high-throughput in vitro screening. *Toxicol Sci* 2012;**128**:398–417.
130. Budinsky RA, Schrenk D, Simon T, *et al*. Mode of action and dose-response framework analysis for receptor-mediated toxicity: The aryl hydrocarbon receptor as a case study. *Crit Rev Toxicol* 2014;**44**:83–119.
131. Kopf PG, Walker MK. Overview of developmental heart defects by dioxins, PCBs, and pesticides. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* 2009;**27**:276–85.
132. Dix DJ, Houck KA, Judson RS, *et al*. Incorporating biological, chemical, and toxicological knowledge into predictive models of toxicity. *Toxicol Sci* 2012;**130**:440–1; author reply 442–3.
133. Gusenleitner D, Auerbach SS, Melia T, *et al*. Genomic models of short-term exposure accurately predict long-term chemical carcinogenicity and identify putative mechanisms of action. *PLoS One* 2014;**9**:e102579.
134. Tawa GJ, AbdulHameed MDM, Yu X, *et al*. Characterization of chemically induced liver injuries using gene co-expression modules. *PLoS One* 2014;**9**:e107230.
135. Davis AP, Grondin CJ, Lennon-Hopkins K, *et al*. The comparative toxicogenomics database's 10th year anniversary: update 2015. *Nucleic Acids Res* 2015;**43**:D914–20.
136. Toxicology UE-NC for C. *Computational Toxicology Research Program (CompTox)*. Factsheets. http://www.epa.gov/ncct/challenges.html (15 January 2015, date last accessed).
137. Cheng F, Li W, Zhou Y, *et al*. admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. *J Chem Inf Model* 2012;**52**:3099–105.