The role of translational bioinformatics in drug discovery

Natalie S. Buchan[1], Deepak K. Rajpal[2], Yue Webster[3], Carlos Alatorre[3], Ranga Chandra Gudivada[3,1], Chengyi Zheng[3,2], Philippe Sanseau[1] and Jacob Koehler[3,3]

[1] GlaxoSmithKline, Computational Biology, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, UK
[2] GlaxoSmithKline, Computational Biology, 5 Moore Drive, 3.2084, Research Triangle Park, NC 27709, USA
[3] Eli Lilly and Company, Indianapolis, IN 46285, USA

The application of translational approaches (e.g. from bed to bench and back) is gaining momentum in the pharmaceutical industry. By utilizing the rapidly increasing volume of data at all phases of drug discovery, translational bioinformatics is poised to address some of the key challenges faced by the industry. Indeed, computational analysis of clinical data and patient records has informed decision-making in multiple aspects of drug discovery and development. Here, we review key examples of translational bioinformatics approaches to emphasize its potential to enhance the quality of drug discovery pipelines, reduce attrition rates and, ultimately, lead to more effective treatments.

Introduction

Translational research in drug discovery and development is broadly defined as the closer integration and use of discovery and preclinical activities with clinical applications [1]. However, clinical data can also feed back to drive research activities. Opportunities for success using translational research has been underpinned by the establishment of new organizations, journals and conferences based on the principle that closer interactions between discovery and development research will deliver better drugs that are produced over shorter timeframes [2]. This is an attractive proposition when, at best, a flat number of new drugs are approved every year, while R&D budgets have been growing steadily [3] and many key patents will be lost in the near future. [4]. It could take 13 years and a 'capitalized' cost of US$ 1778 million to discover a new drug [5]. In addition, costs of developing a new drug have been growing exponentially [6]. Reducing costs or time to launch in the different steps of drug discovery and development are therefore top priorities. It is essential to reduce the number of failures, also know as attrition, in the different steps of drug discovery and development. The three major reasons for drug failure are lack of clinical efficacy (~30%), toxicology (>20%) and commercial concerns (>20%) [7]. A major reason for the high attrition not attributed to commercial decisions is likely to be the use of animal models to predict efficacy and safety in humans. Indeed, the poor translation of preclinical findings from animals to patients is a major issue in therapeutic areas, such as neurosciences [8]. Therefore, the application of translational research has the potential to result in a fundamental paradigm shift by leveraging human-derived data to overcome some of the inherent limitations associated with the use of animal models. Figure 1 illustrates the limitations of using only animal models, the good situation of having both animal and human positive data, but also that opportunities could be missed by not translating interesting human-based efficacy and safety findings that are not observed in animals.

Another trend is the ever-increasing amount of diverse information generated during all phases involved in the discovery and development of new drugs [9]. Obtaining this information has a cost and it is essential that the data are used and analyzed as effectively as possible. Computational manipulation, integration

Corresponding author:. Sanseau, P. (philippe.x.sanseau@gsk.com)
[1] Current address: Corning Life Sciences, 271 County Route, 64 Big Flats, NY 14814, USA.
[2] Current address: Department of Research and Evaluation, KPSC, 100, S. Los Robles, 2nd floor, Pasadena, CA 91101.
[3] Current address: Dow AgroSciences, 9330 Zionsville Rd, Indianapolis, IN 46268, USA.

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2011.03.002

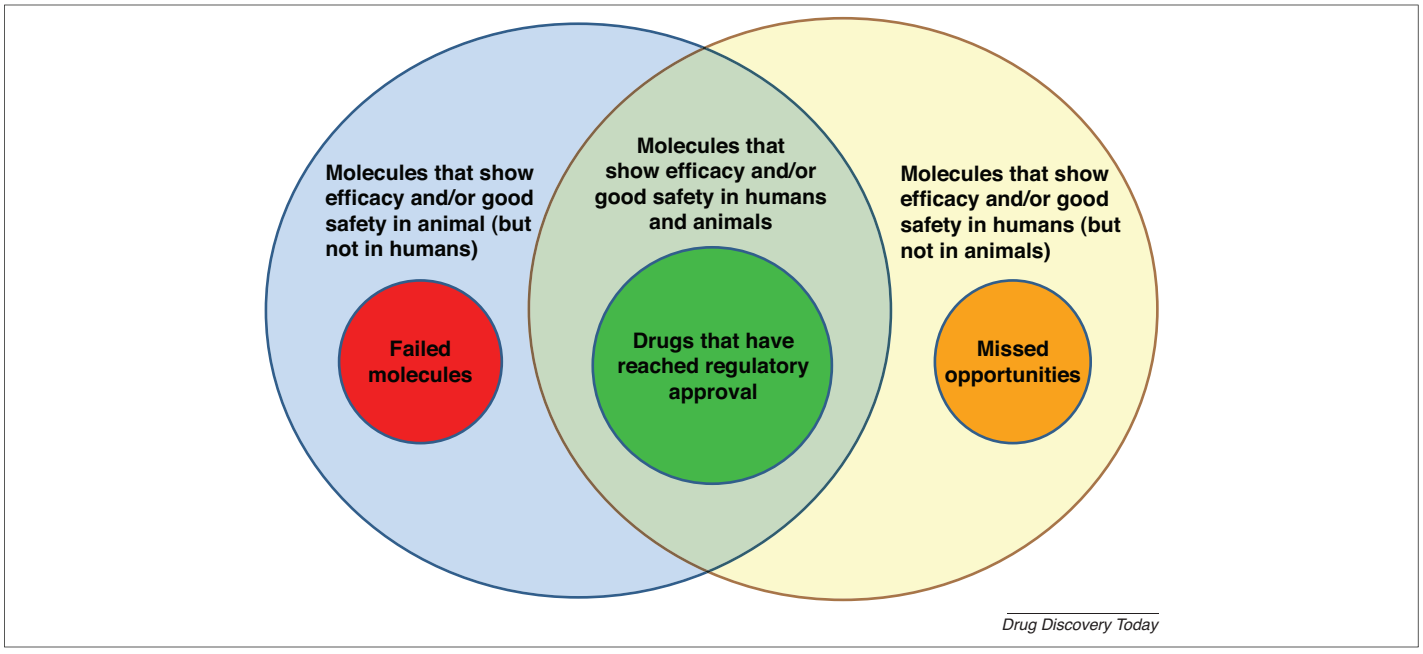

FIGURE 1

Challenges and opportunities in translational research. Regulatory processes require that drugs that reach approval (green circle) are successfully tested in both animals and humans for efficacy and safety. However, translating positive efficacy and/or safety findings from animal models to humans is a major cause of attrition (failed molecules highlighted by the red circle). Furthermore, the current drug development approach, which requires that efficacy is demonstrated in animal models, misses the opportunity to identify drugs that work only in humans, but not in animals (missed opportunities represented by the orange circle). Using data from human and animal studies in a translational fashion could, for example, rescue failed molecules or highlight missed opportunities.

and analysis of these data to support translational research, has the potential to generate greater value than restricting its use to only the original context in which it was generated. As such, the translational bioinformatics field was defined in 2006 by the American Medical Informatics Association (AMIA) as the development of storage, analytical and interpretive methods, to optimize the transformation of increasingly voluminous biomedical data, and genomic data in particular, into proactive, predictive, preventive and participatory health (http://www.amia.org/inside/stratplan). It has been growing steadily and includes areas such as not only molecular bioinformatics, but also clinical informatics or public health informatics [10–12]. There are many opportunities for the application of translational bioinformatics approaches across the pharmaceutical pipeline (Fig. 2) [13]. Here, we discuss how translational bioinformatics has impacted many drug discovery and development milestones, highlight important data resources and discuss the challenges and future directions.

Impact of translational bioinformatics on preclinical discovery research

Discovery biology

Platform technologies, such as genomics, proteomics and metabolomics (the study of the genes, proteins or metabolites of an organism as molecules of interest), have been widely used over the years to understand the inherent biology of normal versus diseased states. There are various recent examples of pharmacogenomics impacting clinical research (as reviewed by [14]). For example, characterization of a cancer genome (adenocarcinoma of the tongue) through high-throughput sequence analysis, suggested that the tumor cells are driven by the RET oncogene, implying that the tumor is sensitive to RET inhibitors. Subsequent administration of sunitinib, a RET inhibitor, was associated with a stable disease for four months [15]. In addition, polymorphism analysis of CYP2D6, which is required for the bioactivation of tamoxifen, resulted in an increased understanding of the clinical importance of having a functional CYP2D6 in the treatment of estrogen receptor-positive breast cancers. These studies showed an association of better clinical outcomes in the presence of two functional CYP2D6 alleles, when compared with either non-functional or reduced-function alleles [16]. Similar approaches open up the possibility of using complete genome characterization to offer opportunities in clinical decision making and in developing therapeutic protocols for rare cancers, especially where none might currently exist.

Following the seminal work of Golub *et al.* [17] (who proposed a novel leukemia sample classification based on global gene expression profiles), various approaches have been proposed to use transcriptomics studies in understanding disease biology and in drug discovery. Recently, methods that query public gene expression repositories using gene expression data, rather than textual searches, further opens up the possibility of discovering novel relationships between diseases, treatments and biological signals [18]. Further impacts of genomics platforms in translational bioinformatics will be realized through collaboration with clinical and healthcare disciplines [19,20].

An additional complementary source of information for translational bioinformatics is human phenotypic data, particularly those derived from patient clinical histories. These data were recently used to generate a phenotype disease network (PDN), which identifies comorbidity relationships between diseases, based upon the probability that individuals are affected by both diseases at a substantially higher rate than by chance alone [21].

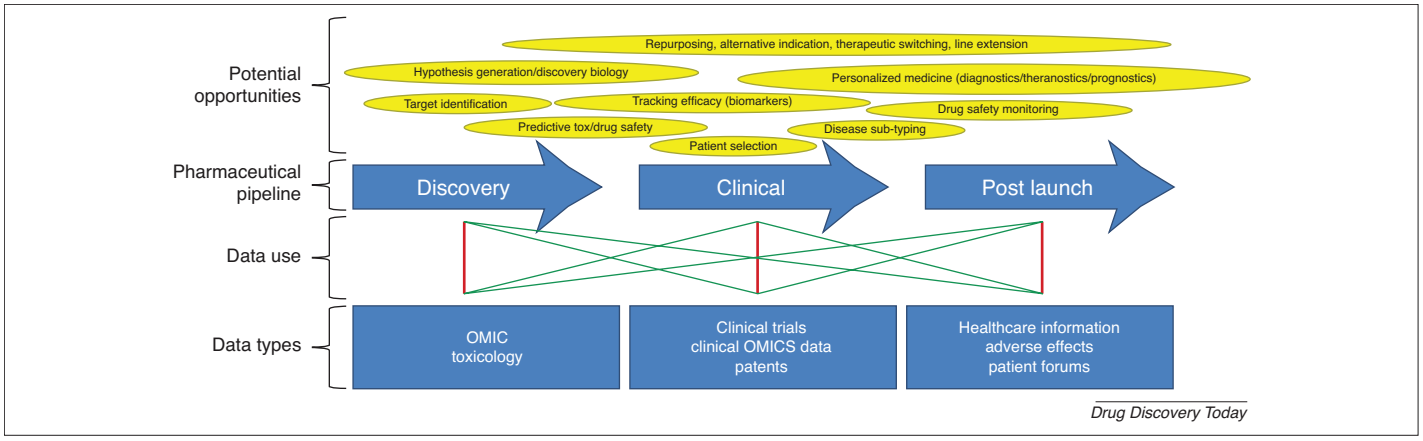

FIGURE 2

A schematic view of the pharmaceutical pipeline in the context of translational bioinformatics, defined by phases (discovery, clinical and post launch) with data type aligned to the phase in which they are collected. The term 'OMICs' detailed under data types covers platform technologies such as genomics, proteomics and metabolomics. Red lines linking data types back to the pipeline highlight phases that are impacted via traditional bioinformatics methods. The green diagonal lines reveal how data can be used in a translational manor and integrated with additional data types to affect alternative phases of the pipeline. The yellow bubbles represent potential opportunities through which integration and analysis of data types via translational bioinformatics could impact the pipeline; for example, electronic healthcare records and adverse event reporting are used traditionally to inform post-launch decisions. However, this healthcare information could be exploited to inform and support decisions in activities undertaken during the discovery phase, such as target identification or disease indication.

PDN nodes represented disease phenotypes and edges corresponded with connections between phenotypes displaying significant comorbidities from disease histories of more than 30 million patients. The PDN also highlighted differences in disease progression with regards to gender and race. Furthermore, patients with diseases that were highly connected nodes in the PDN had worse prognosis compared with patients with diseases represented by less connected nodes. Exploring these phenotypic 'maps' could be helpful in addressing disease progression and identifying populations of patients with differences in biological mechanisms, environmental factors and healthcare quality. Similar approaches, particularly when combined with additional genomics data sets, are likely to be useful for generating hypotheses on the mechanisms underpinning common diseases. Such studies could be exploited in the early-stage biology efforts of drug discovery for related diseases. Development of the methodology might also lead to the selection of appropriate patient populations for clinical research programs.

Drug repurposing

Drug repurposing is the process of finding new uses for existing drugs outside the scope of the original medical indication [22]. Pharmaceutical companies are increasingly under pressure to justify escalating costs of research in producing therapies [5]. Taking advantage of existing chemical molecules for new or alternative indications is therefore of great interest.

Various translational bioinformatics approaches have been used for repurposing drugs for new indications. Expression-based connectivity maps catalog gene expression signatures as biological responses to several perturbations [23]. Comparison of these signatures across different small molecule drugs could provide hypotheses on mechanisms of action and/or new indications. Using the connectivity map approach, the authors suggested that the compound sirolimus should be evaluated in patients with acute lymphoblastic leukemia and dexamethasone resistance. One limitation is the lack of diversity of the cell lines used. It will be necessary to increase the number of cell lines to realize fully the potential of this approach.

Expression-based disease networks incorporate large-scale disease–disease, drug–drug and disease–drug networks by directly matching gene expression profiles, such as those available at the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus [24]. Exploring disease–drug relationships in this manner can enhance understanding of disease mechanisms and identify alternative indications. A recent study using this approach integrated gene expression and protein interaction networks to identify 59 gene subclusters dysregulated in various diseases but enriched in drug targets, which could be investigated for novel therapies in various common diseases, such as cancer and psychiatric disease [25]. To further explore their proposal that similar molecular phenotypes are shared by similar diseases and that similar drugs could be therapeutically relevant in their treatment, the authors examined a list of drugs with their reported target genes as well their US Food and Drug Administration (FDA)-approved indications or off-label uses. The authors found that fluorouracil, which is an FDA-approved drug used for actinic keratosis, has shown positive indications for malignant colon tumor treatment. Similarly, doxorubicin is FDA approved as a treatment for both urothelial carcinoma and acute myeloid leukemia.

Literature mining techniques can identify trends around disease that can point to areas of emerging science and identify targets for pathway-guided therapeutic opportunities [26,27]. Similarity ensemble approaches to compare targets by the chemical similarity of their ligands, rather than their sequence or structural similarity, have resulted in various hypotheses for repurposing [28–30]. Two commercial compounds were recently proposed as potential ligands for protein farnesyltransferase, highlighting their possible use in treating several disease conditions [31]. Studying similarities between phenotypic adverse effects of drugs to determine the likelihood that they share a common target, can suggest new uses for marketed drugs [32]. In this case, one example is the potential

use of donepezil to treat depression. Adverse effect data can be used for new (off) target identification as well as drug repurposing. Although exploiting adverse event data to understand potential pathways associated with adverse events is reasonable, a similar approach might also lead to the identification of pathways modulated by compounds [33]. Drug–disease 'guilt by association', uses the principle that if two diseases share similar therapies it is logical to speculate that additional drugs currently used for only one might also prove to be therapeutic for the other [34]. The authors propose novel uses for several drugs. For example, although rituximab is approved for the treatment of non-Hodgkin's lymphoma and rheumatoid arthritis, this approach suggests that this drug could be used for cataracts, gastric ulcers and stomach cancers. None of these three diseases had been evaluated in clinical trials with rituximab at the time of the analysis. By using various preclinical methods to evaluate the hypotheses generated by drug-repurposing approaches, one can quickly take molecules into further *in vivo* validation strategies.

Drug safety

Adverse drug reactions (ADRs) account for 2.4–6.4% of all hospitalizations in the western world and ~100,000 deaths are attributed to serious ADRs (SADRs) in the USA [35,36]. It is crucial to predict and avoid ADRs as early as possible. Translational bioinformatics approaches integrating knowledge of genes and pathways associated with ADRs could impact early drug discovery by elucidating underlying pathways and mechanisms.

To offset the cost of extensive preclinical safety assessment on various targets, translational bioinformatics methods are beginning to be used in further ways to understand ADRs. Extending the 'systems chemical biology' idea [37], Scheiber [33] proposed a workflow to extract compounds sharing common ADRs, performed *in silico* target prediction and revealed pathways putatively linked to ADRs via these targets. Yang [38] used literature mining and a CitationRank, based on Google's PageRank algorithm, to create a database of gene–SADR relationships on six major SADRs: deafness, cholestasis, QT prolongation, muscle toxicity, Stevens–Johnson syndrome (SJS) and torsades de points. Pathway enrichment on each SADR core gene set revealed pathways that might underpin SJS. Exploring mechanisms shared by genes associated with SADRs could offer clues to understanding and reducing ADRs by either avoiding or carefully modulating them.

Impact of translational bioinformatics approaches on clinical research

Drug repurposing

Scientific data can be rapidly tested by patients to evaluate an existing compound for a new therapeutic indication. Amyotrophic lateral sclerosis (ALS) is a progressive, fatal, neurodegenerative disease, which usually leads to paralysis and death within five years of onset [39]. No cure has yet been found and, although the sole FDA-approved drug, riluzole, can prolong life by three to six months, it cannot change the course of the disease. Fornai *et al.* reported that lithium can slow down disease progression [40]. In response, an ALS patient-driven study, where members of the social-networking community PatientsLikeMe.com collaborated in a self-experiment with lithium, showed no correlation between lithium and a reduction in disease progression [41]. The first results were reported before the scientific community was able to initiate such a clinical study. The patient-led study was later corroborated by clinical trial NCT00818389 [42], and highlights the potential of using patient forums in clinical research, especially in orphan diseases with smaller patient populations. However, the potential lack of appropriate controls or appropriate study design, for example, could limit the applications of such an approach. These patient-initiated trials also have regulatory, ethical and legal implications.

Patient selection

It is important to select the most appropriate patients for each clinical trial. PAthway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM) stratifies patients into clinically relevant subtypes, by inferring signatures from transcriptomic and genetic copy number variation (CNV) data in patients and integrating protein interactions captured from the NCBI Pathway Interaction Database (http://pid.nci.nih.gov) [43]. Interpreting these signatures in the context of patient survival profiles enabled the separation of patients with glioblastoma into subgroups with significantly different survival outcomes. Combining patient data with pathways improves the stratification of patients and has the potential for selecting appropriate patients for clinical trials and highlighting therapeutic targets specific to individual patients.

Biomarkers

Biomarkers are used in clinical trials to study the direct interaction between a compound and its target and monitor consequences of these interactions. Translational bioinformatics approaches using genome-wide RNA data from human skeletal muscle in insulin-resistant patients, reviewed in the context of pathway knowledge, identified novel candidate biomarkers for use in clinical trials [44]. The study suggested that, whereas the transcriptome of patients with type 2 diabetes mellitus is indistinguishable from control subjects, one third of microRNAs (miRNAs) expressed in muscle were altered in these patients. Over 47,000 mRNAs and >500 human miRNAS were profiled in 118 subjects, 71 having diabetes. A tissue-specific gene-ranking system was generated to stratify miRNA target genes and a weighted scoring system was introduced to reflect the net impact of miRNA-induced up- and downregulated gene expression. Validation by protein profiling and enrichment analysis of the ranked target genes identified six canonical pathways with proven links to metabolic disease. The results suggest that insulin resistance in humans is associated with coordinated changes in miRNAs, which target specific metabolically aligned pathways and represent plausible biomarkers for muscle status. Muscle biopsy sampling in clinical trials, to monitor levels of miRNA biomarkers, could survey pharmacodynamics and early-stage efficacy of compounds.

Impact of translational bioinformatics approaches post launch

Translational bioinformatics approaches have the potential to impact the delivery of personalized medicine to the patient. Traditional diagnosis exploits data types including symptoms, medical and family history and laboratory data such as histopathological assessment. Advances in translational bioinformatics have

encouraged the integration of these data types with additional biological data, such as that generated via '-omic' technologies. Integration and analysis of this metadata will increase opportunities to identify biomarkers, which can then be used to monitor a patient's disease status and aid decision-making to tailor medical care to the individual. It is also important to be aware of the important challenges that remain, especially in making available, and using appropriately, the large amount of very diverse research data to clinical practice.

Diagnostic solutions

Histopathology currently remains a key component of patient diagnosis, but supplementary parallel molecular tests are becoming increasingly available. Gatza *et al.* [45] exemplified the ability to stratify heterogeneous collections of patient breast tumors into pathway-aligned subgroups. Translational bioinformatics using high-throughput expression data from patients and cell lines with histological data were examined in the context of biological pathway signatures aligned to drug targets. Classification of breast cancers into subgroups on the basis of homogeneous patterns of pathway activation has the potential to aid physicians in selecting the most rational drug combination. The study also established trends in genetic alterations associated with each subgroup by integrating a breast tumor study, which included the capture of copy number variation (CNV) data and expression data, to define subgroup-specific patterns of CNV localized to distinct regions of chromosomes that were not evident when looking at the heterogeneous data set.

This strategy provides an opportunity to evaluate the benefit to a patient of a particular treatment by stratifying patients in the context of biological pathways and genetic alterations that are aligned with specific drug combinations and their targets. Once validated in clinical trials, this approach could be a diagnostic tool with the potential to treat patients using the most effective pathway and/or mechanistically aligned therapies.

Theranostic solutions

A theranostic test is a diagnostic to establish whether a patient is most likely to be helped or harmed by a treatment. This type of test can highlight whether a patient is at a higher risk of experiencing an adverse event with a particular drug, indicating that selection of an alternative treatment is required. A recent translational bioinformatics approach exploited genetic data, the human interactome, DrugBank and Adverse Event Reporting System (AERS) data to provide biological rationale for a compound to potentially trigger an adverse event [46]. The method was applied to Long-QT syndrome (LQTS), a congenital disease that can be induced by drugs and has been reported as an ADR [47]. Thirteen LQTS disease susceptibility genes were used as seed nodes to define a subnetwork from the human interactome. A random walk-based algorithm was exploited to identify and rank nodes proximal to seed nodes. The resulting network, defined as the neighborhood, was found to be enriched with targets of FDA-approved drugs known to cause LQTS as an adverse event. Additional drugs that are likely to be associated with QT effects were predicted and evidence was sought for these predictions within AERS. Furthermore, the neighborhood was analyzed in the context of two genome-wide association studies (GWAS) to reveal previously unknown single-nucleotide polymorphisms that are likely to affect the QT interval. Mutations in candidate genes present in the neighborhood in patients with no known cause of LQTS were screened to identify causative mutations. Targets in the LQTS neighborhood could be categorized into positive and negative regulators of the QT interval. It is possible that the output from this type of analysis could be used to assess patient risk and guide the choice in selecting effective treatment with the lowest adverse event risk. Many of these network approaches will become more useful when more information on the interactions such as directionality (e.g. inhibition) or biological activity (e.g. phosphorylation) is integrated.

Prognostic solutions

Understanding the dynamic classification for stages of a disease will aid in the identification of novel biomarkers to help determine the prognosis and likelihood of therapeutic response. For cancer, prognostic makers could help to predict the likelihood of relapse following surgical removal of a tumor. Markers could also be used to help gauge the severity of treatment required, depending on the level of risk for the patient.

A translational bioinformatics analysis has been described by Li *et al.* [48]. The multiple survival screening algorithm was created to address the dilution of 'real' cancer gene expression signals in the high level of 'passenger signals' associated with tumor cells and genome instability induced by mutated tumor suppressor genes. Prognostic gene signatures for both estrogen receptor-negative and -positive subtypes in patients with breast cancer were generated and used to stratify patients into low-, intermediate- and high-risk groups. Signatures identified in the training set were validated and shown to be highly predictive in eight independent microarray testing sets covering 1375 samples. The signatures were validated further by network analysis using I2D protein interactions to highlight metastasis-aligned modules of direct interactions between signature genes and breast cancer driver-mutating genes.

Data sources for translational bioinformatics: opportunities and challenges

Access to relevant information is the foundation of translational bioinformatics. Key data resources are summarized in Table 1. The volume of molecular data in the form of DNA sequencing [49], gene expression microarrays [50] and proteomics [51] has grown exponentially in recent years. The adoption of next-generation sequencing technologies is expected to expand rapidly both the amount of sequencing information and the associated challenges in managing and storing vast amounts of data [52]. These platform data have been extensively used in drug discovery. However, the application of molecular data in drug development, such as for biomarker identification, has been more limited. As an illustration, it has been recently highlighted that more than 150,000 papers based on genomics technologies, such as microarrays, have claimed thousands of biomarkers. Despite this, fewer than 100 biomarkers have been validated to be used in a clinical setting [53]. The causes are likely to be diverse, and sometime case specific, but challenges in manipulating and analyzing research molecular data for clinical use or lack of standardization cannot be ignored. Cultural and training differences among scientists involved in early research or those in clinical activities might also play a role. In addition, organizational boundaries between discovery and

TABLE 1

Examples of data sources for applications in translational bioinformatics analyses

Name (sponsor)	URL	Description
Molecular resources		
GenBank Database (NCBI)	http://www.ncbi.nlm.nih.gov/genbank/	A repository of annotated, publicly available DNA sequences
Gene Expression Omnibus (GEO) (NCBI)	http://www.ncbi.nlm.nih.gov/geo/	A repository of publicly available gene expression profiles
Array Express Database (EBI)	http://www.ebi.ac.uk/microarray-as/ae/	A repository of publicly available gene expression profiles
PRoteomics IDEntifications (PRIDE) Database (EBI)	http://www.ebi.ac.uk/pride/	A repository of proteomics data
Health record resources		
Google Health (Google)	http://www.google.com/health/	A centralization service for personal health information, where users volunteer their health records
PatientsLikeMe	http://www.patientslikeme.com	A social-networking health site enabling members to share symptom and treatment information
UK Biobank (multiple)	http://www.ukbiobank.ac.uk	A medical research initiative aimed at creating a resource to study the health of 500,000 UK volunteers, currently aged 40–69 years
Clinical trial resources		
Clinical Trials.gov (NIH)	http://clinicaltrials.gov/	A registry of federally and privately supported clinical trials; provides details such as the purpose and summary results of a trial
ClinicalStudyResults.org (PhRMA)	http://www.clinicalstudyresults.org/	A repository for clinical study results for marketed pharmaceuticals
TrialTrove (Citeline)	http://www.citeline.com/products/trialtrove/overview/	A repository of clinical studies capturing trials from Phase I to Phase IV
EudraCT (muliple)	https://eudract.ema.europa.eu/eudract/index.do	A European clinical trials database
Safety data resources		
Adverse Event Reporting System (AERS) (FDA)	http://www.fda.gov/Drugs/GuidanceCompliance RegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm	A repository capturing information collected from the post-marketing safety surveillance program of the FDA for all approved drug and therapeutic biological products
Vaccine Adverse Event Reporting System (VAERS) (FDA/CDC)	http://vaers.hhs.gov/	A repository that captures the reporting of adverse events following immunization
VigiBase (WHO)	http://www.umc-products.com	A database of international drug safety data that provides access to information such as case reports and safety profile-related information
SIDER (EMBL)	http://sideeffects.embl.de	A repository of marketed medicines and their recorded adverse drug reactions extracted from public documents and package inserts
Drug Interaction DataBase (DIDB) (Washington University)	http://www.druginteractioninfo.org/	A repository of human drug interactions extracted form sources such as PubMed, New Drug Applications and FDA Prescribing Information
Phase Forward Lincoln Safety Group	http://www.phaseforward.com	A source of pharmacovigilance and risk management analysis on adverse event reporting, data mining, signal detection and signal management
SuperToxic (Charite University)	http://bioinformatics.charite.de/supertoxic	A database of toxic compounds extracted from literature and web sources that provides details of possible biological interactions
P450 drug interaction table (Indiana University)	http://medicine.iupui.edu/clinpharm/ddis/table.asp	A list of drugs that are metabolized by specific cytochrome P450 isoforms
Drug and drug target resources		
DrugBank (University of Alberta)	http://www.drugbank.ca	A repository that combines drug data, drug target and drug action information
Claim data resources		
The General Practice Research Database (GPRD) (MHRA)	http://www.gprd.com/home/	A repository of longitudinal medical records captured from primary care provided in the UK

development could make fluid exchange of information or approaches challenging. There are also many cases when use of information, especially clinical data, has to be restricted.

Medical records are another key data source for translational bioinformatics, especially in the computerized form of electronic health records (EHRs), such as the UK-based General Practice Research database (GPRD). As with paper-based methods, EHRs can capture diverse information, including text and images. In addition to the EHRs populated and maintained by hospitals and physicians, one should not underestimate the emergence of per-

sonal health initiatives, such as GoogleHealth, where anyone can organize and post their own medical records and share it securely with family members or doctors. Online health communities, such as PatientsLikeMe.com, aim to create repositories of real-world life-changing disease information. Although it helps patients to share experiences, such a resource could also provide useful information to investigators. The development of these web-based health-related communities is a fascinating new model of sharing medical information. With the challenges around standardization, it remains to be seen how much impact they will have directly on drug discovery and development.

A related consumer endeavor is 23andMe, where an individual could obtain genotypic information. The Electronic Medical Records and Genomics (eMERGE) project, funded by the National Institute of Health, is currently investigating whether EHRs can serve as a resource for GWAS. However, the use of EHR information presents several important challenges. There is a significant absence of consistency in how physicians report patient-associated data. Patient privacy and ethical considerations could appropriately limit access to data and the type of analysis possible, as well as reinforcing the need for robust anonymization and encryption solutions.

Availability of clinical trials data could also have a key role in translational bioinformatics. Requirements to post clinical trial information on the US Government website ClinicalTrials.gov have expanded [54]. In Europe, EudraCT has been established as a database of all clinical trials that started in the European Community from 1 May 2004 onwards. In addition, the pharmaceutical industry has also started to make clinical trial databases available. Some are company specific (e.g. GlaxoSmithKline Clinical Study Register or Roche Clinical Trial Protocol Registry and Results Database), whereas others are industry sponsored, such as PhRMA, the Pharmaceutical Research and Manufacturers of America with ClinicalStudyResults.org. Clinical trials information could also be accessed from literature or documents available from the FDA website. Despite the increasing amount of information related to clinical trials, not all data are made available. There are no requirements to post data on clinical trials completed before September 27, 2007 or to make Phase I safety data available. Increased pressure to expand the type of clinical trial data available [55] will positively impact translational bioinformatics activities. More generally, the disclosure of clinical trial results will also impact on the overall industry by providing earlier input to product differentiation, encourage competition between drugs with different profiles and, ultimately, help to obtain the best drugs in class more rapidly.

Safety-related data sources are also important for translational approaches. Human safety information could be used to inform the development of new classes of compound. Major resources include FDA AERS, the Vaccine Adverse Event Reporting System (VAERS) database and the VigiBase of the World Health Organisation (WHO), the latter offering more advanced queries, such as neural network analysis. The Phase Forward Lincoln Safety Group conducts a substantial data cleaning on the AERS, VAERS and VigiBase raw data to enable quantitative analyses.

Adverse effects data could also be used for translational bioinformatics. The Side Effect Resource (SIDER) database is an important resource as it compiles recorded adverse events from package inserts and public documents for marketed drugs and provides data on adverse effect frequency, classifications of drugs and adverse effects and links to drug–target interactions [56]. In addition, SuperToxic provides toxicity information and has links to targets and pathways [57].

Drug metabolism and drug–drug interaction databases are also attractive resources. For example, the Drug Interaction Database of Washington University contains *in vitro* and *in vivo* information on drug interactions in humans using data from publications, new drug applications and FDA prescribing information. Metabolism–drug interactions are also captured in resources, including the Indiana University P450 drug interaction table. Data providers, such as GeneGo (http://www.genego.com) and Ingenuity (http://www.ingenuity.com), also deliver curated information, mostly from literature, related to safety. General resources on drugs, such as DrugBank [58], which contains information about drug and drug targets, are also important. Additional important types of resource are biobanks, such as the UK Biobank, where these samples are combined with patient-derived information, such as disease, demographics, patient history or genomic information.

Semantic interoperability between data sets is essential to support translational bioinformatics analyses. As such, agreed ontologies and standards will be fundamental for powerful analyses across these multiple data sources. A prototype has recently been developed to annotate and index text data from multiple sources, such as microarray, clinical trial reports or descriptions of radiology images with unified medical language system (UMLS) concepts [59,60]. Natural language processing and machine learning approaches have also been applied to index biomedical literature automatically [61]. However, text mining still has limitations. For example, some issues remain with the indexing of medical subject headings (MeSH) around compliance with indexing policies or scalability [49].

The implementation of solid but flexible informatics infrastructures is essential. Several large initiatives are underway to build new computational frameworks to accelerate the translation of genomic findings into the clinic. These include the NIH-funded Informatics for Integrating Biology and the Bedside (http://www.i2b2.org) [62] and caGRID (http://cagrid.org) [63], a data-sharing framework applied to specific activities, including caTissue and cancer bench-to-bedside (caB2B). With many researchers still using traditional database systems or even files managed by individuals, infrastructure is essential and should provide the capability to search across large and diverse data sets.

Discussion and outlook

The combination of the challenges faced by the pharmaceutical industry and the increasing amount of data generated during the many phases of the drug discovery and development process makes translational bioinformatics approaches crucial. In addition to using and analyzing patient or preclinical data via traditional approaches, it is important to consider how they can be applied in a translational manner to as many phases of the drug discovery and development pipeline as possible. Indeed, innovative translational computational analysis approaches have already had an impact on the discovery, preclinical and clinical phases, as illustrated by the examples included in this review.

Where large amounts of diverse data are being manipulated and analyzed, there is a requirement for strong informatics capabilities. Interoperability of systems or databases, use of agreed information standards and integration of data to support computational analyses are still areas that need particular attention. Organizational boundaries between scientists could also hinder translational activities. However, one key challenge that should not be underestimated is the human one. The effective application of translational bioinformatics approaches requires the right competencies, skills and interest in human medicine. Traditionally, bioinformatics scientists focused on genomic analysis and infrastructure, whereas clinicians often lacked the computational skills necessary for managing and manipulating large and diverse data sets. Therefore, it is important to encourage cross-training such as for a clinician to learn about bioinformatics and for a computational biologist to be trained in human physiology [64]. As many competencies are already available, although sometimes in silos, it is even more essential to encourage direct interactions between researchers with a diverse background. In this context, meetings, such as the American Medical Informatics Association's Translational Bioinformatics Summits (http://summit2010.amia.org/), are key to help the cross-fertilization of ideas. Facilitation through the establishment of cross-disciplinary teams or co-location is also valuable in accelerating the cross-fertilization of ideas and projects between scientists involved in either discovery or development activities.

In the future, we expect to see an increasing amount of information to be made available for computational analysis and, therefore, a growing number of opportunities to influence all phases of drug discovery through translational bioinformatics. The increased external pressure on pharmaceutical companies will make clinical trials information more available [65]. Patient-based websites and communities are also likely to grow. However, using patient information for research activities might be limited owing to ethical or individual patient consent constraints. Overall, with the development of innovative computational analysis combined with the increasing amount of relevant biomedical information, it could be the 'perfect storm' for translational bioinformatics to address some of the challenges faced by the industry to discover and develop the safe and efficacious drugs of the future.

Acknowledgements

The authors thank Julie Huxley-Jones for editorial suggestions.

References

1 Woolf, S.H. (2008) The meaning of translational research and why it matters. *JAMA* 299, 211–213
2 Goldblatt, E.M. and Lee, W.H. (2010) From bench to bedside: the growing use of translational research in cancer medicine. *Am. J. Transl. Res.* 2, 1–18
3 Malik, N.N. (2009) Key issues in the pharmaceutical industry: consequences on R&D. *Expert Opin. Drug Discov.* 4, 15–19
4 Goodman, M. (2009) Pharmaceutical industry financial performance. *Nat. Rev. Drug Discov.* 8, 927–928
5 Paul, S.M. *et al.* (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* 9, 203–214
6 Munos, B. (2009) Lessons from 60 years of pharmaceutical innovation. *Nat. Rev. Drug Discov.* 8, 959–968
7 Kola, I. and Landis, J. (2004) Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 3, 711–715
8 Markou, A. *et al.* (2009) Removing obstacles in neuroscience drug discovery: the future path for animal models. *Neuropsychopharmacology* 34, 74–89
9 Loging, W. *et al.* (2007) High-throughput electronic biology: mining information for drug discovery. *Nat. Rev. Drug Discov.* 6, 220–230
10 Lussier, Y.A. *et al.* (2010) Current methodologies for translational bioinformatics. *J. Biomed. Inform.* 43, 355–357
11 Olaya, P. and Wasserman, M. (1991) Effect of calpain inhibitors on the invasion of human erythrocytes by the parasite *Plasmodium falciparum*. *Biochim. Biophys. Acta* 1096, 217–221
12 Kulikowski, C.A. and Kulikowski, C.W. (2009) Biomedical and health informatics in translational medicine. *Methods Inf. Med.* 48, 4–10
13 Payne, P.R. *et al.* (2009) Translational informatics: enabling high-throughput research paradigms. *Physiol. Genomics* 39, 131–140
14 Altman, R.B. *et al.* (2011) Pharmacogenomics: will the promise be fulfilled? *Nat. Rev. Genet.* 12, 69–73
15 Jones, S.J. *et al.* (2010) Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. *Genome Biol.* 11, R82
16 Schroth, W. *et al.* (2009) Association between CYP2D6 polymorphisms and outcomes among women with early stage breast cancer treated with tamoxifen. *JAMA* 302, 1429–1436
17 Golub, T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537
18 Engreitz, J.M. *et al.* (2010) Content-based microarray search using differential expression profiles. *BMC Bioinformat.* 11, 603
19 Vlaanderen, J. *et al.* (2010) Application of OMICS technologies in occupational and environmental health research; current status and projections. *Occup. Environ. Med.* 67, 136–143
20 Sethi, P. and Theodos, K. (2009) Translational bioinformatics and healthcare informatics: computational and ethical challenges. *Perspect. Health Inf. Manag.* 6, 1h
21 Hidalgo, C.A. *et al.* (2009) A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.* 5, e1000353
22 Ashburn, T.T. and Thor, K.B. (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* 3, 673–683
23 Lamb, J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935
24 Hu, G. and Agarwal, P. (2009) Human disease–drug network based on genomic expression profiles. *PLoS ONE* 4, e6536
25 Suthram, S. *et al.* (2010) Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput. Biol.* 6, e1000662
26 Agarwal, P. and Searls, D.B. (2009) Can literature analysis identify innovation drivers in drug discovery? *Nat. Rev. Drug Discov.* 8, 865–878
27 Li, Y. and Agarwal, P. (2009) A pathway-based view of human diseases and disease relationships. *PLoS ONE* 4, e4346
28 Keiser, M.J. *et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* 25, 197–206
29 Keiser, M.J. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature* 462, 175–181
30 Keiser, M.J. and Hert, J. (2009) Off-target networks derived from ligand set similarity. *Methods Mol. Biol.* 575, 195–205
31 DeGraw, A.J. *et al.* (2010) Prediction and evaluation of protein farnesyltransferase inhibition by commercial drugs. *J. Med. Chem.* 53, 2464–2471
32 Campillos, M. *et al.* (2008) Drug target identification using side-effect similarity. *Science* 321, 263–266
33 Scheiber, J. *et al.* (2009) Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. *J. Chem. Inf. Model.* 49, 308–317
34 Chiang, A.P. and Butte, A.J. (2009) Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. *Clin. Pharmacol. Ther.* 86, 507–510
35 van der Hooft, C.S. *et al.* (2008) Adverse drug reaction-related hospitalisations: a population-based cohort study. *Pharmacoepidemiol. Drug Saf.* 17, 365–371
36 Wilke, R.A. *et al.* (2007) Identifying genetic risk factors for serious adverse drug reactions: current progress and challenges. *Nat. Rev. Drug Discov.* 6, 904–916
37 Oprea, T.I. *et al.* (2007) Systems chemical biology. *Nat. Chem. Biol.* 3, 447–450
38 Yang, L. *et al.* (2009) A CitationRank algorithm inheriting Google technology designed to highlight genes responsible for serious adverse drug reaction. *Bioinformatics* 25, 2244–2250

39 Wijesekera, L.C. and Leigh, P.N. (2009) Amyotrophic lateral sclerosis. *Orphanet. J. Rare Dis.* 4, 3
40 Fornai, F. *et al.* (2008) Lithium delays progression of amyotrophic lateral sclerosis. *Proc. Natl. Acad. Sci. U. S. A.* 105, 2052–2057
41 Allison, M. (2009) Can Web 2.0 reboot clinical trials? *Nat. Biotechnol.* 27, 895–902
42 Aggarwal, S.P. *et al.* (2010) Safety and efficacy of lithium in combination with riluzole for treatment of amyotrophic lateral sclerosis: a randomised, double-blind, placebo-controlled trial. *Lancet Neurol.* 9, 481–488
43 Vaske, C.J. *et al.* (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 26, i237–i245
44 Gallagher, I.J. *et al.* (2010) Integration of microRNA changes *in vivo* identifies novel molecular features of muscle insulin resistance in type 2 diabetes. *Genome Med.* 2, 9
45 Gatza, M.L. *et al.* (2010) A pathway-based classification of human breast cancer. *Proc. Natl. Acad. Sci. U. S. A.* 107, 6994–6999
46 Berger, S.I. *et al.* (2010) Systems pharmacology of arrhythmias. *Sci. Signal.* 3, 1–15
47 Goldenberg, I. *et al.* (2008) Long QT syndrome. *Curr. Probl. Cardiol.* 33, 629–694
48 Li, J. *et al.* (2010) Identification of high-quality cancer prognostic markers and metastasis network modules. *Nat. Commun.* 1, 1–8
49 Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.* 26, 1135–1145
50 Schena, M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470
51 Mann, M. (1999) Quantitative proteomics? *Nat. Biotechnol.* 17, 954–955
52 Richter, B.G. and Sexton, D.P. (2009) Managing and analysing next-generation sequence data. *PLoS Comput. Biol.* 5, e1000369
53 Poste, G. (2011) Bring on the biomarkers. *Nature* 469, 156–157
54 Laine, C. *et al.* (2007) Clinical trial registration: looking back and moving ahead. *N. Engl. J. Med.* 356, 2734–2736
55 Wood, A.J. (2009) Progress and deficiencies in the registration of clinical trials. *N. Engl. J. Med.* 360, 824–830
56 Kuhn, M. *et al.* (2010) A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* 6, 343
57 Schmidt, U. *et al.* (2009) SuperToxic: a comprehensive database of toxic compounds. *Nucleic Acids Res.* 37, D295–D299
58 Wishart, D.S. *et al.* (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36, D901–D906
59 Lindberg, D.A. *et al.* (1993) The unified medical language system. *Methods Inf. Med.* 32, 281–291
60 Shah, N.H. *et al.* (2009) Ontology-driven indexing of public datasets for translational bioinformatics. *BMC Bioinformat.* 10 (Suppl. 2), S1
61 Neveol, A. *et al.* (2009) A recent advance in the automatic indexing of the biomedical literature. *J. Biomed. Inform.* 42, 814–823
62 Heinze, D.T. *et al.* (2008) Medical i2b2 NLP smoking challenge: the A-Life system architecture and methodology. *J. Am. Med. Inform. Assoc.* 15, 40–43
63 Oster, S. *et al.* (2008) caGrid 1.0: an enterprise grid infrastructure for biomedical research. *J. Am. Med. Inform. Assoc.* 15, 138–149
64 Butte, A.J. (2008) Translational bioinformatics: coming of age. *J. Am. Med. Inform. Assoc.* 15, 709–714
65 Smyth, R.L. (2009) Making information about clinical trials publicly available. *Br. Med. J.* 338, b2473