Target discovery from data mining approaches☆

Yongliang Yang, S. James Adelstein and Amin I. Kassis

Harvard Medical School, Harvard University, Department of Radiology, Armenise Building, Room D2-137, 200 Longwood Avenue, Boston, MA 02115, USA

Data mining of available biomedical data and information has greatly boosted target discovery in the 'omics' era. Target discovery is the key step in the biomarker and drug discovery pipeline to diagnose and fight human diseases. In biomedical science, the 'target' is a broad concept ranging from molecular entities (such as genes, proteins and miRNAs) to biological phenomena (such as molecular functions, pathways and phenotypes). Within the context of biomedical science, data mining refers to a bioinformatics approach that combines biological concepts with computer tools or statistical methods that are mainly used to discover, select and prioritize targets. In response to the huge demand of data mining for target discovery in the 'omics' era, this review explicates various data mining approaches and their applications to target discovery with emphasis on text and microarray data analysis. Two emerging data mining approaches, chemogenomic data mining and proteomic data mining, are briefly introduced. Also discussed are the limitations of various data mining approaches found in the level of database integration, the quality of data annotation, sample heterogeneity and the performance of analytical and mining tools. Tentative strategies of integrating different data sources for target discovery, such as integrated text mining with high-throughput data analysis and integrated mining with pathway databases, are introduced.

Introduction

Target discovery is the most crucial step in a modern drug discovery campaign. Past records have indicated that the high failure rate of drug development can be largely attributed to improper target selection [1–3]. A target in the drug discovery process can be from a broad spectrum of moieties, such as molecular entities (genes/proteins/miRNA), disease biomarkers, biological pathways and crucial 'nodes' on a regulatory network, as long as it is relevant to a specific disease and its progression [4]. Target discovery can be grouped into two categories, a system approach and a molecular approach [1]. The system approach is a strategy that selects targets through the study of diseases in whole organisms using information derived from clinical trials and *in vivo* animal studies. The molecular approach, the mainstream of current target discovery strategies [3,5], is geared towards the identification of 'druggable' targets where activities can be modulated through interactions with small molecules or proteins and/or antibodies. Presently, the majority of 'druggable' targets are G-protein-coupled receptors (GPCRs) and protein kinases. Because the biological mechanisms of human diseases are rather complex, the most crucial task in target discovery is not only to identify, prioritize and select reliable 'druggable' targets but also to understand the cellular interactions underlying disease phenotypes, to provide predictive models and to construct biological networks for human diseases [1]. This requires extensive gathering and filtering of a multitude of available heterogeneous data and information.

We are embracing an unprecedented omics era with the explosion of biological data and information. For instance, the most popular biomedical literature database, MEDLINE/PubMed, currently contains more than 18 million literature abstracts, and more

☆ This article is a reprint of a previously published article. For citation purposes, please use the original publication details; Drug Discovery Today 14/3–4(2009), pp. 147–154.

DOI of original article: 10.1016/j.drudis.2008.12.005

Corresponding authors: Kassis, Yang, Y. (everbright99@gmail.com), A.I. (amin_kassis@hms.harvard.edu)

1359-6446/06/$ - see front matter. Published by Elsevier Ltd. doi:10.1016/j.drudis.2011.12.006

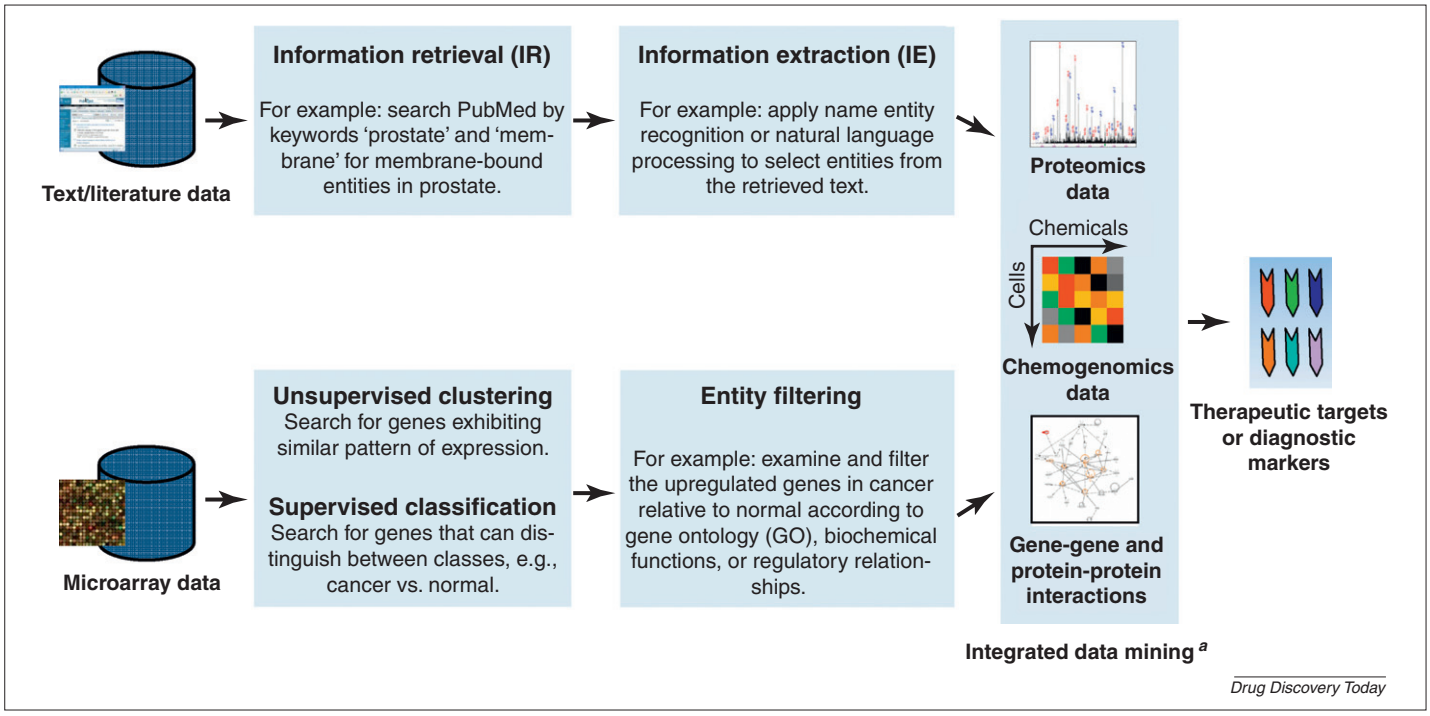

FIGURE 1

Workflow of text mining and microarray data mining integrated with other high-throughput data and interaction data for discovery of therapeutic targets or diagnostic markers.

[a] Text and microarray data can be combined with proteomics data or chemogenomics data to discover targets; different sources of data can be 'mapped' or 'visualized' based on gene–gene or protein–protein interaction pathways generated by high-throughput experiments to discover valuable targets in a systematic fashion.

than 60,000 new abstracts are added monthly. Analogously, the number of databases warehousing chemical, genomic, proteomic and metabolic data is rapidly growing with their size estimated to double every two years. This wealth of biological data and information presents immense new opportunities for target discovery in support of the drug discovery pipeline [3]. In pace with the growth of biological databases, the flourishing of bioinformatics, especially data mining approaches, to extract or filter valuable targets by combining biological ideas with computer tools or statistical methods has changed the way target discovery is conducted. Currently, text mining of literature databases and microarray data mining are the two prevailing approaches to target discovery [5]. With the recent development of high-throughput proteomics and chemical genomics, another two data mining approaches, proteomic data mining and chemogenomic data mining, have surfaced (Figure 1). To keep up with new scientific discoveries, there is a clear need to develop efficient data mining methods to fuel target discovery in the post-genomics era.

Text mining

Overview of text mining

Text mining (TM) can be defined as the computational discovery of new, previously unknown information, by automatically extracting information from different written resources [6]. Generally, TM consists of two major steps [7–9], information retrieval (IR) and information extraction (IE). First, IR finds literature or abstracts related to a particular topic with the aid of general search engines or specifically designed IR searching tools (Box 1). There are two very common searching approaches in IR: (i) rule-based or knowledge-based; and (ii) statistical or machine-learning [7]. The first approach uses patterns that rely on basic biological insights, for example '<prostate> and <membrane>' (Figure 1), to find the literature or abstracts of interest. The second approach uses syntactic parse trees (which can also be rule-based) or classifiers to classify the related biomedical literature. Named entity recognition (NER), a prerequisite for IE, relies on tools or methods for automatic term recognition to extract entities such as genes, proteins, drugs or other molecules. iHOP (information hyperlinked over proteins; also see Box 1) is an excellent example that browses sentences from Medline abstracts on the basis of the entities that appear. IE [6] is then used to identify or tabulate the relevant entities or facts from the retrieved documents. IE can roughly be divided into two approaches. The first and simplest approach to IE is co-occurrence, which identifies entities that co-occur within the text. Furthermore, co-occurrence could be used to extract relationships of a certain type, for example physical protein–protein interactions [6]. The second approach is to extract relations such as gene–gene or protein–protein interactions and biological pathways, which progress beyond the simple recognition of terms. Natural language processing (NLP) [10], a technology that combines syntax and semantics, has been widely applied in the second approach.

Applications to target discovery

Identification of disease-associated entities

Text mining has been broadly applied to identify disease-associated entities (genes/proteins) and to understand their roles in diseases. Very recently, Ozgur *et al.* [11] described a novel entity recognition method to retrieve and prioritize candidate genes associated with prostate cancer. First, an initial list of 15 genes

BOX 1

Websites of some popular text, microarray, pathway databases and associated mining tools

Resources	Descriptions	Web links
Text/structural databases		
PubMed Central	Full-text	http://www.pubmedcentral.nih.gov/
HighWire Press	Full-text	http://highwire.stanford.edu/
E-Biosci	Full-text	http://www.e-biosci.org/
PubMed	Abstracts	http://www.ncbi.nlm.nih.gov/pubmed/
UniProt	Information for proteins	http://www.uniprot.org/
InterPro	Protein domains	http://www.ebi.ac.uk/interpro/
Text mining tools		
Google Scholar	Search engine	http://scholar.google.com/
GoPubMed	PubMed engine	http://www.gopubmed.org/
Textpresso	Full-text search	http://www.textpresso.org/
BioRAT	Full-text search	http://bioinf.cs.ucl.ac.uk/biorat/
ABNER	Entity taggers	http://pages.cs.wisc.edu/~bsettles/abner/
iHOP	Entity recognition	http://www.ihop-net.org/UniPub/iHOP/
GeneWays	Pathway extraction	http://geneways.genomecenter.columbia.edu/
Microarray databases		
SMD	Raw datasets	http://genome-www5.stanford.edu/
Gene Expression Omnibus	Raw datasets	http://www.ncbi.nlm.nih.gov/geo/
Oncomine	Cancer datasets	http://www.oncomine.org/
CGAP database	Cancer datasets	http://cgap.nci.nih.gov/
caArray	Cancer datasets	http://array.nci.nih.gov/caarray/
Gene Expression Atlas	Human Tissues	http://symatlas.gnf.org
Clustering platform		
GenePattern		http://www.broad.mit.edu/cancer/software/genepattern/
GeneCluster 2		http://www.broad.mit.edu/cancer/software/genecluster2/gc2.html
ArrayMiner		http://www.optimaldesign.com/ArrayMiner/ArrayMiner.htm
Supervised analysis platform		
SAM		http://www-stat.stanford.edu/~tibs/SAM/
Pathway and interactome databases		
KEGG		http://www.genome.jp/kegg/
UniHI		http://theoderich.fb3.mdc-berlin.de:8080/unihi/home
PathwayExplorer		http://pathwayexplorer.genome.tugraz.at/
GenMAPP		http://www.genmapp.org/
Pathguide		http://www.pathguide.org/ *(A complete list of pathway databases)

(seed genes) that are well known to be related to prostate cancer was collected from a curated database, Online Mendelian Inheritance in Man (OMIM; also see Box 1). The list of seed genes was then used to construct a disease-specific gene-interaction network mined from the full text articles stored in PubMed Central (PMC), based on the dependency parsing and support vector machines (SVM) method. The extended list of genes in the gene-interaction network was then ranked and prioritized according to the closeness centrality in the literature-mined network. Remarkably, a total of 95% of the top 20 genes ranked by this method were previously confirmed to be associated with prostate cancer. Similarly, our group [12] has employed a combined textual-structural mining approach to retrieve potential enzyme targets in the extracellular space of cancerous cells for six common and lethal human tumors, by searching PubMed abstracts, universal gene/protein database – UniProt, conserved protein domains database – InterPro and NCBI Entrez. First, a literature mining tool LSGraph program was used to extract entities from the curated database mentioned above based on keywords and gene ontology (GO) terms. These entities were then enlarged by related functional annotations and clustered further based on cellular locations and biochemical functions within Ingenuity knowledgebase. Finally, this method has led to the identification of a list of cancer-related hydrolases for each tumor type, among which prostatic acid phosphatase (ACPP), prostate-specific antigen (PSA) and sulfatase 1 (SULF1) have been selected as suitable targets for our in-house enzyme mediated cancer imaging and therapy [13].

Identification of disease-associated networks

One elegant example of applying text/literature mining to identify disease-related networks is by Krauthammer *et al.* [14]. They have created a mining tool called GeneWays (Box 1), which automatically examines a large number of full-text research articles to predict the physical interactions (edges) among candidate disease genes (seed nodes) hidden in literature. First, mining in 25 scientific journals by GeneWays has led to a literature-derived interaction network that describes the direct relationship between entities (such as binding and phosphorylation). Then, a list of 60 AD (Alzheimer's disease) candidate genes manually prepared by an expert in the field was used as a set of seeds to search subnetworks that might harbor entities related to AD. This method performed well in predicting network nodes that match AD

BOX 2

Frequently used mining methodologies to analyze high-throughput data (for each methodology, three commonly used algorithms are given as examples)

I. Normalization

A transformation method applied to observational high-throughput data that adjusts the individual profiles to balance them appropriately so that meaningful biological comparisons can be made. For example, (i) linear regression analysis, (ii) non-linear regression analysis and (iii) lowest normalization.

II. Unsupervised clustering

A clustering approach in which the observational data are analyzed to determine whether the samples exhibit a similar pattern of expression without constraint on samples. For example, (i) hierarchical pairwise clustering, (ii) principal component analysis and (iii) self-organizing maps.

III. Supervised classification

An approach that builds a model to classify known samples (e.g. cancer vs. normal); it requires a training set and a test set to validate the classifiers. For example, (i) linear discriminant analysis, (ii) K-nearest neighborhood prediction and (iii) trained neural network.

candidate genes; the results were confirmed by experts in the field. Recently, considerable efforts have been made to develop mining tools for extracting interaction networks related to human diseases from the literature. For example, PolySearch [15] is a recently developed web-based tool to identify biomedical associations and networks from published abstracts and many well-annotated databases. Similarly, GenCLip [16] is a literature mining tool developed to discover gene clusters and networks related to disease pathogenesis.

Limitation and challenges

Although text mining is very useful to derive biological entities and insights from an astronomically large number of research articles, several problems still persist [6,7,17]. The first problem is with the term variation and term ambiguity of biomedical entities [18]. Term variation occurs when a biomedical concept can be denoted by various realizations. For example, <prostate> and <prostatic> can be used as keywords to search entities or networks related to prostate diseases. Vice versa, term ambiguity arises when the same term may refer to many biomedical concepts. For example, the string 'fat' can be referred to as both the symbol of Entrez Gene entry 2195, a cadherin associated with tumor suppression, and the symbol of Entrez Gene entry 948, which is a thrombospondin receptor associated with atherosclerosis and platelet glycoprotein deficiency [10]. These ambiguities can lead to erroneous relations between molecular biology and human diseases. To overcome this problem, methods for rapid development of controlled vocabularies in text mining have been proposed. For instance, the use of GO terms [19] (also known as 'controlled gene ontology vocabulary') designating subcellular location, molecular function and biological process has allowed more appropriate annotation for entities and enhanced retrieval. A second limitation is restricted access to the full text of papers and to citation information; more comprehensive, specific and detailed information is hidden in full-text articles than in abstracts. Thus, the number of entities identified from text mining can be greatly underestimated due to the condensed nature of literature abstracts. Finally, it is important for the researchers to know the levels of reliability and accuracy of various mining methods and their associated tools. Bridging the gaps between biologists and computational scientists is another challenging task. Therefore, while biologists should be made aware of the novelty of text mining for biomedical target discovery, computational researchers should be encouraged to develop more user-friendly methods and tools for biologists.

Microarray data mining

Overview of microarray data mining

Microarray data mining refers to applying bioinformatics approaches in microarray data analysis to discover entities and biological pathways that define a phenotype, such as a human disease [18,20,21]. Two basic approaches that are broadly applied in microarray data mining are: unsupervised clustering and supervised classification [22] (also see Box 2). In the former approach a group of genes that share coherent expression across a subset of conditions is determined using clustering methods such as hierarchical clustering, principal component analysis (PCA) and self-organizing maps (SOM) [22]. For instance, the SOM method finds an optimal set of 'centroids' around which the gene expression data appear to aggregate. Then, cell or tissue samples can be partitioned into groups with each centroid defining a cluster based on similarity measures for the data points such as Euclidean distance and the Pearson correlation coefficient [23]. By contrast, a supervised analysis approach searches for genes that can distinguish between known samples and conditions. In a typical supervised analysis, the global gene expression profiles of disease tissues/fluids will be compared to those in normal tissues/fluids (e.g. cancer vs. healthy tissues/fluids) from which a list of target genes or biological pathways that are important in diseases will be identified. Supervised classification methods such as linear discriminant analysis, nearest neighborhood search and genetic algorithms have been used in this approach [22]. Driven by the exponential growth of microarray data over the past few years, considerable effort has been made to develop microarray databases with timely public accessibility in a manner that facilitates the target discovery (Box 1). Accordingly, meta-analysis of multiple microarray datasets that addresses similar biological hypotheses has been proposed [24]. The merit of meta-analysis methods is that statistical measures across different studies could be compared and all positive results could be assessed simultaneously. In addition, other gene discovery strategies such as massively parallel signature sequencing (MPSS), serial analysis of gene expression (SAGE) and expressed sequence tags (EST) have also proved to be fruitful in identifying targets and markers [25].

Applications to target discovery

Identification of therapeutic targets

Microarray data mining has proved a fruitful approach to discovering target genes associated with human diseases. For example, IGFBP3 has been identified as a hypermethylation target of

prostate cancer from a data mining approach [26]. Briefly, significantly downregulated genes in prostate cancer, as compared with the normal prostate, were identified from Gene Expression Atlas database (Box 1). The retrieved genes were then organized by putative function using GeneCards (http://www.genecards.org/). Among the list of 631 retrieved genes, 16 of them were commonly identified as downregulated by other studies and, finally, IGFBP3 was selected and verified as a hypermethylation target of prostate cancer. As another elegant example, Ryu *et al.* [27] have recently strived to identify novel molecular signatures as therapeutic targets for aggressive melanoma, a cancer with one of the highest increasing rates in the USA. First, they have compared and analyzed a large amount of gene expression profiles from a series of melanoma cell lines representing discrete stages of malignant progression and primary human melanocytes through unsupervised hierarchical clustering methods implemented in GeneCluster (Box 1). This clustering analysis has enabled them to identify two distinct groups of cell lines, one primary melanoma group and one aggressive melanoma group. Further, a supervised microarray data mining platform, significance analysis of microarrays (SAM; also see Box 1), was employed together with functional annotation analysis to identify a panel of highly upregulated invasion-specific genes in aggressive melanoma, among which NF-κB, CXCL1, CXCL2, IL-8, MMP1 and IGFBP3 have been previously implicated in the promotion of tumor-associated angiogenesis, a crucial feature of tumor aggressiveness.

Identification of diagnostic or prognostic markers

Biomarkers are molecules that are indicators of the physiologic state and hallmarks of changes in a tissue or a bodily fluid during a disease process [28,29]. With today's growing needs for biomarker discovery, microarray data mining has been increasingly used to detect diagnostic or prognostic marker genes [29]. For instance, Kim *et al.* [30] have reported the mining of public gene expression data from the CGAP database and GEO database (Box 1) to identify candidate markers for lung cancer. First, several hundreds of differentially expressed genes in lung cancer were retrieved through meta-analysis of these two databases using Fisher's exact test method. Further, a systematic examination based on the annotated properties of the genes and a statistical *P* value cut-off led to 20 candidate genes that were subjected to experimental validations. Finally, seven candidate genes that are highly over-produced were selected as potential diagnostic markers for lung cancer. Similarly, our group [31] has successfully identified lists of blood-borne biomarkers for six common human cancer types through a combined mining strategy in the Oncomine microarray platform and a curated pathway knowledgebase. First, all of the significantly upregulated genes with defined GO cellular locations in cancer were collected with a false discovery rate cut-off. These retrieved genes were then subjected to pathway analysis and only those encoding secreted proteins in blood/serum/plasma as putative markers were kept in the list. Further, a comparison study of the retrieved marker genes across different tumor types has led to the identification of common and unique markers in six tumors, among which ErbB2, BRCA1/BRCA2, PSA, HABP2 and IGF-II have also been selected by other studies as candidate tumor markers or are already being used clinically. Remarkably, after manually consulting the iHOP database (Box 1) and other curated databases, 13 markers out of the common 35 markers (~1/3) across prostate, breast and lung tumors have been literature-confirmed to serve as prognostic markers for the progression and invasiveness of human tumors. In addition, MMP1, CD44, CP and NOTCH4 were selected and prioritized as promising blood-based markers according to the normalized fold change-abs[*t*] value.

Limitation and challenges

Although powerful, there are also a number of limitations and challenges for microarray data mining in target discovery [20,22,32]. First, data mining a list of target genes is not the end of the genomic analysis and, because gene expression levels do not always correlate with protein levels, follow-up experiments are required to validate the protein expression levels and protein functions [18,21]. Therefore, techniques such as quantitative RT-PCR, immunohistochemistry (IHC) or *in situ* hybridization (ISH) need to be applied to aid in the target discovery. Second, microarray data exist on a variety of scales depending on the specific technological platform as well as the individual experimental procedures. Therefore, microarray data from different labs are not always directly comparable [24]. Third, data availability and data integration can be a challenge for the microarray data mining approach. In the post-genomic era, the explosion of gene expression data requires timely data storage and update of gene databases. Moreover, the different formats of data storage across databases have posed a great challenge for data mining and analysis. Data integration is needed to combine data residing at different sources and databases into a uniform view or format. GO has provided such a solution for data integration by using a controlled vocabulary (GO terms) to describe genes and gene products in any organism [19]. Finally, computational and statistical expertise required for genomic data mining remains a great challenge for biologists to meet.

Emerging data mining approaches

Proteomic data mining

With the arrival of the post-genomic era of proteomics, a new technology based on high-throughput mass spectrometry (MS) analysis has emerged [33,34]. Accordingly, proteomic data mining is needed to analyze and extract useful information from MS data points. Since as many as 1–2 million data points may be included per sample in a high resolution MS instrument [35], proteomic data mining is challenging because of the size and dimension of the massive datasets. For example, a typical high-resolution MS-based analysis of a patient blood sample could result in the generation of 350,000–400,000 points, with the mass-to-charge (m/z) ratio and amplitude of the ion(s) being measured. It is impractical to analyze these many datasets with traditional plotting tools and spreadsheet methods. Therefore, there is a need to develop novel mining tools and methods to hurdle the target discovery from a proteomic approach. Recently, Open Proteomic Database (OPD) [36] and EMBL Proteomic Database (PRIDE) [37] became available to the public, and mining methods such as Bayesian analysis, rule-based analysis and likehood scoring have been proposed to discover patterns of diagnostic signatures [33]. Advanced computational methods are, however, still needed for integration, mining, comparative analysis and functional interpretation of high-throughput proteomic data.

Chemogenomic data mining

Another emerging data mining approach, chemogenomic data mining, interprets the data from chemical genomics, a new technology examining the phenotypes of interest (such as viability, cell morphology, behavior and gene expression profiles) in a parallel fashion by applying small molecules from chemical libraries to a library of cells [38,39]. In the 2D matrix resulting from chemogenomics screening, one dimension is the chemical library and the other dimension is the library of different cell types (Figure 1). This can create new ways to identify cellular drug targets and to discover disease pathways. The interpretation and filtering of multi-dimensional chemogenomic data is a difficult task. The challenge associated with chemogenomic data mining has initiated the development of mining tools and methods to profile and analyze data in a systematic way [40]. Notably, a number of supervised or unsupervised clustering algorithms have been proposed to obtain a subset of genes with significant functions from the overall pattern, such as hierarchical clustering, *k*-means, self-organizing maps, bioclustering and matrix operations [39].

Integrated data mining

Target discovery is an arduous task owing to the complexity of human diseases and the heterogeneity of various biological data. No single data mining approach is sufficient for understanding the cellular mechanisms and reconstructing the biological networks [1–2,4]. To retrieve and prioritize biologically meaningful targets, we need to integrate and analyze a wealth of data across many different disciplines [41–46]. Bioinformatics approaches that integrate different sources of data, taking merits and drawbacks of each into consideration, would significantly enhance the discovery of valuable targets [30]. Particularly, the combination or integration of text mining with high-throughput data analysis (such as genomic, proteomic or chemogenomic data) has been increasingly used to search disease markers and drug targets (Figure 1). By contrast, with the emergence of system biology, the continued growth of gene–gene and protein–protein interaction data has enabled scientists to analyze and visualize a variety of datasets in the context of biological networks or pathways, mainly with a manually curated knowledgebase such as KEGG (Kyoto Encyclopedia of Genes and Genomes; see Box 1) and experimental interactome databases such as UniHI (Unified Human Interactome; see Box 1). For instance, PathwayExplorer (Box 1) is a tool that mines high-throughput expression data based on curated pathway knowledgebases such as KEGG and GenMAPP (Box 1). In addition, it allows the mapping of expression profiles of genes or proteins simultaneously onto major regulatory, metabolic and cellular pathways. Below we have listed a few more concrete examples to demonstrate the usefulness of such integrated mining approaches.

Integrated text mining with high-throughput data analysis

Recently, Natarajan *et al.* [47] have successfully combined the mining of full-text articles with genomic data analysis to reveal the effect of sphingosine I-phosphate (SIP), a lysophospholipid stimulus involved in cell apoptosis, proliferation and migration, in invasive human glioblastoma and its downstream cascading events. First, they identified a set of 72 differentially expressed genes from microarray data analysis as a unique response to SIP, comparing them with the expression profiles under the influence of epidermal growth factor (EGF). This set of genes was then used to infer gene–gene interaction networks extracted by mining full articles from 20 popular scientific journals in the cancer research field over a five-year period (1999–2003), based on natural language processing (NLP) methods. Among the derived gene–gene interaction networks, they have mapped a particular interesting network triggered by SIP, in which matrix metalloproteinases-9 (MMP-9) was identified as a key player in invasive glioblastomas. Similarly, Li *et al.* [48] have applied combined literature mining and microarray analysis (LMMA) approach to construct a target network for the angiogenesis, a process of generating new capillary blood vessels and a fundamental step in the transition of tumors from a dormant state to a malignant state. This approach is particularly interesting because it has summarized and integrated large amounts of related literature and microarray data in a systematic fashion. First, they have collected all the related PubMed abstracts using 'angiogenesis' as a keyword, from which 1929 genes with HUGO symbols and 9514 co-citations were retrieved to construct a co-occurrence angiogenesis network. Next, the angiogenesis-related gene expression profiles of endothelial cells (EC) and solid tumors (ST) were collected from the Stanford Microarray Database (SMD). Further, the literature-based angiogenesis network was refined using the retrieved gene expression profiles through a multivariate selection procedure, based on the hypothesis that literature-co-cited gene pairs will indeed interact with each other if they are co-upregulated or co-downregulated. Finally, a refined angiogenesis network was derived in which numerous hub genes could be used as targets to inhibit tumor angiogenesis, such as tumor necrosis factor (TNF)-alpha, interleukin (IL)-1, -6 and vascular endothelial growth factor (VEGF).

Integrated mining with pathway databases

As an excellent example, Chassey *et al.* [49] have successfully built a Hepatitis C virus (HCV) infection protein target network by integrating yeast two-hybrid screening and literature mining with eight curated interaction knowledgebases, including BIND, BioGRID, DIP, GeneRIF, HPRD, IntAct, MINT and Reactome (see 'Pathguide' in Box 1). First, 314 protein–protein interactions between HCV and human proteins were identified by yeast two-hybrid experiments and 170 by text mining. For the text mining approach, all abstracts related to the keywords 'HCV' and 'protein interactions' were retrieved and subjected to gene name recognition and human expert curation. These derived protein–protein interactions were used as seeds and integrated into the eight curated knowledgebases to reconstruct a HCV–human interaction network, among which CORE protein, NS3 protein and NS5A protein were identified as major targets to develop anti-viral molecules. In addition to curated knowledgebases, experimental interactome databases have also been intensively used to identify potential targets. Very recently, Yue *et al.* [50] combined the proteomic data analysis with mining the UniHI database (Box 1), an experimental protein–protein interaction database, to construct a target network for the anticancer drug ganoderic acid D (GAD), a major component in a traditional Chinese herbal medicine. Briefly, 21 differentially expressed proteins were first identified as cellular targets of GAD through proteomic data analysis. These 21 proteins were then used as seeds to fish partner interacting proteins in the UniHI database. The iterative searching of such partner proteins has led to an expanded network

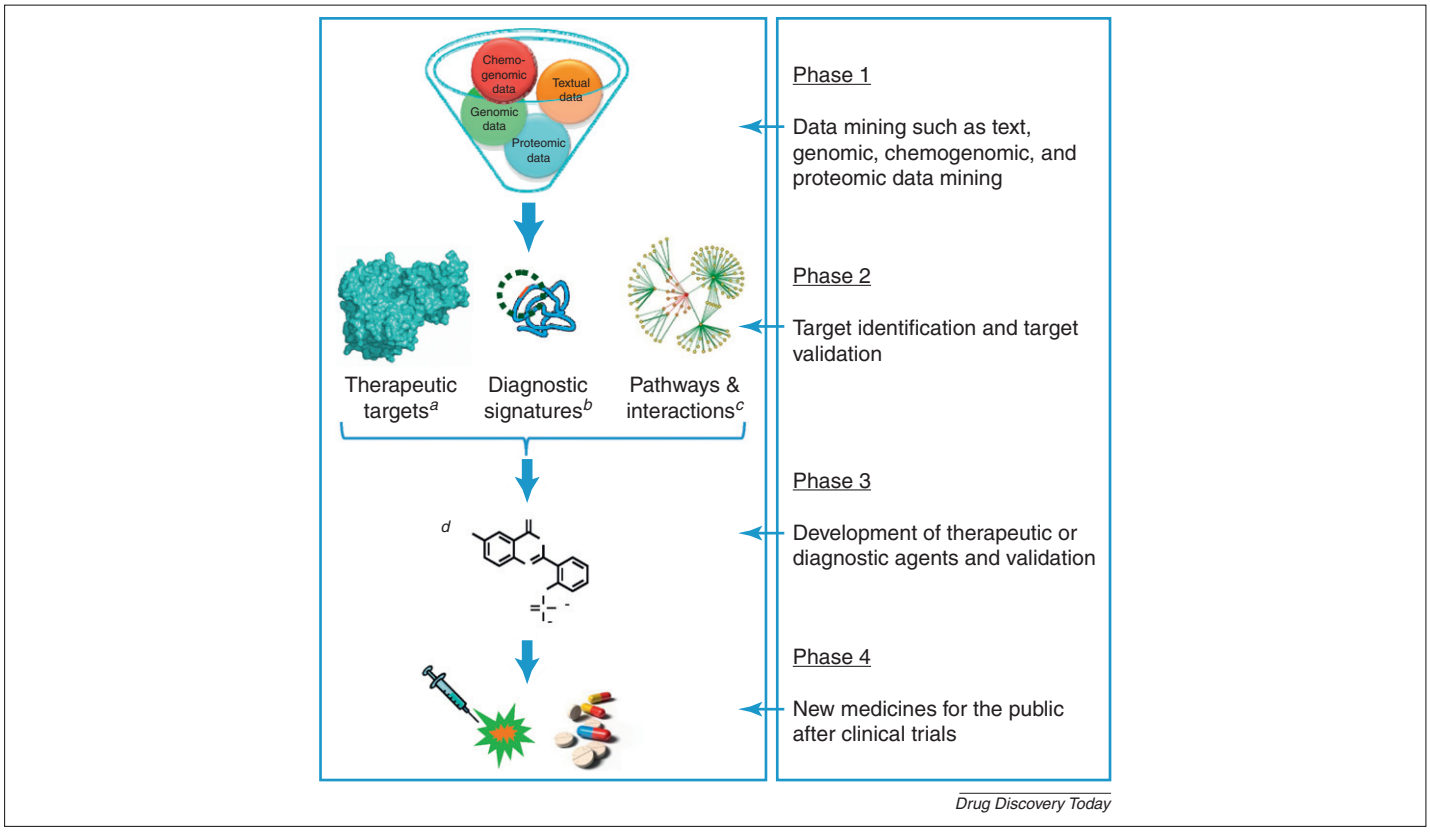

FIGURE 2

Scheme of drug discovery pipeline in the 'omics' era.
[a] Cartoon picture of human placental alkaline phosphatase as 'druggable' target. [b] Diagnostic signatures shed into human blood. [c] A cell growth and proliferation pathway for human prostate tumor. [d] An enzyme-activated prodrug structure of 2-(2′-phosphoryloxyphenyl)-6-iodo-4-(3H)-quinazolinone ($IQ_{2\text{-}P}$); *I: ^{127}I (chemotherapy), ^{131}I (radiotherapy), and ^{123}I (radioimaging) of ligand for tumor targets.

including all 21 experimentally derived proteins. Finally, they have identified the 14-3-3 protein family as a major player in the cytotoxicity mechanisms of GAD through the derived protein–protein interaction network.

Concluding remarks and future prospects

With the rising flood of biomedical data and information generated from a variety of innovative technologies, we are on the verge of an exciting omics drug discovery era. Inevitably, data mining approaches will become the first phase of future drug discovery pipelines by helping to select proper targets and better understand the cellular mechanisms or phenotypes of human diseases (Figure 2). Indeed, data mining has already been widely applied to identify targets for therapeutic invention and early diagnosis. Approaches consist of text mining, microarray data mining and another two emerging mining approaches: proteomic data mining and chemogenomic data mining. Fortunately, a large number of databases warehousing a variety of data, reliable mining tools and methods are under active development. Owing to the inherent limitations of various data mining approaches, however, we suggest that a combination or integration of different mining approaches should be applied to overcome the drawbacks of a single mining method. Consequently, future work should be directed towards the development of integrated databases in uniformed formats, and biologist-friendly software or tools for routine use to accelerate target discovery. This is challenging because human diseases are highly complex processes and biomedical data are largely heterogeneous and poorly defined. Nonetheless, data mining should help researchers to make earlier and crucial decisions in the drug discovery process. We have every reason to believe that data mining will play an increasingly significant part in future biomarker and drug discovery campaigns.

Acknowledgements

We wish to thank Dr Pavel Pospisil and Dr Lakshmanan K. Iyer for their helpful discussions in this project.

References

1 Lindsay, M.A. (2003) Target discovery. *Nat. Rev. Drug Discov.* 2, 831–838
2 Sams-Dodd, F. (2005) Target-based drug discovery: is something wrong? *Drug Discov. Today* 10, 139–147
3 Butcher, S.P. (2003) Target discovery and validation in the post-genomic era. *Neurochem. Res.* 28, 367–371
4 Chen, Y.P. and Chen, F. (2008) Identifying targets for drug discovery using bioinformatics. *Expert Opin. Ther. Targets* 12, 383–389
5 Sakharkar, M.K. and Sakharkar, K.R. (2007) Targetability of human disease genes. *Curr. Drug Discov. Technol.* 4, 48–58

6 Jensen, L.J. *et al.* (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.* 7, 119–129
7 Hirschman, L. *et al.* (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics* 18, 1553–1561
8 Rebholz-Schuhmann, D. *et al.* (2005) Facts from text – is text mining ready to deliver? *PLoS Biol.* 3, e65
9 Ananiadou, S. *et al.* (2006) Text mining and its potential applications in systems biology. *Trends Biotechnol.* 24, 571–579
10 Cohen, K.B. and Hunter, L. (2008) Getting started in text mining. *PLoS Comput. Biol.* 4, e20
11 Ozgur, A. *et al.* (2008) Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics* 24, i277–i285
12 Pospisil, P. *et al.* (2006) A combined approach to data mining of textual and structured data to identify cancer-related targets. *BMC Bioinform.* 7, 354
13 Pospisil, P. *et al.* (2007) Computational modeling and experimental evaluation of a novel prodrug for targeting the extracellular space of prostate tumors. *Cancer Res.* 67, 2197–2205
14 Krauthammer, M. *et al.* (2004) Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.* 101, 15148–15153
15 Cheng, D. *et al.* (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.* 36 (Web Server issue), W399–W405
16 Huang, Z.X. *et al.* (2008) GenCLiP: a software program for clustering gene lists by literature profiling and constructing gene co-occurrence networks related to custom keywords. *BMC Bioinform.* 9, 308
17 Oda, K. *et al.* (2008) New challenges for text mining: mapping between text and manually curated pathways. *BMC Bioinform.* 9 (Suppl. 3), S5
18 Desany, B. and Zhang, Z. (2004) Bioinformatics and cancer target discovery. *Drug Discov. Today* 9, 795–802
19 The Gene Ontology AmiGO (http://amigo.geneontology.org/)
20 Butte, A. (2002) The use and analysis of microarray data. *Nat. Rev. Drug Discov.* 1, 951–960
21 Ricke, D.O. *et al.* (2006) Genomic approaches to drug discovery. *Curr. Opin. Chem. Biol.* 10, 303–308
22 Mount, D.W. and Pandey, R. (2005) Using bioinformatics and genome analysis for new therapeutic interventions. *Mol. Cancer Ther.* 4, 1636–1643
23 D'haeseleer P, (2005) How does gene expression clustering work? *Nat. Biotechnol.* 23, 1499–1501
24 Rhodes, D.R. and Chinnaiyan, A.M. (2004) Bioinformatics strategies for translating genome-wide expression analyses into clinically useful cancer markers. *Ann. N. Y. Acad. Sci.* 1020, 32–40
25 Narayanan, R. (2007) Bioinformatics approaches to cancer gene discovery. *Methods Mol. Biol.* 360, 13–31
26 Perry, A.S. *et al.* (2007) In silico mining identifies IGFBP3 as a novel target of methylation in prostate cancer. *Br. J. Cancer* 96, 1587–1594
27 Byungwoo Ryu, *et al.* (2007) Comprehensive expression profiling of tumor cell lines identifies molecular signatures of melanoma progression. *PLoS ONE* 2, e594
28 Campagne, F. and Skrabanek, L. (2006) Mining expressed sequence tags identifies cancer markers of clinical interest. *BMC Bioinform.* 7, 481
29 He, Y.D. (2006) Genomic approach to biomarker identification and its recent applications. *Cancer Biomark.* 2, 103–133
30 Kim, B. *et al.* (2007) Clinical validity of the lung cancer biomarkers identified by bioinformatics analysis of public expression data. *Cancer Res.* 67, 7431–7438
31 Yang, Y. *et al.* (2008) Integrative genomic data mining for discovery of potential blood-borne biomarkers for early diagnosis of cancer. *PLoS ONE* 3, e3661
32 Li, S. *et al.* (2004) Microarray data mining using gene ontology. *Stud. Health Technol. Inform.* 107 (Pt 2), 778–782
33 Siepen, J.A. *et al.* (2008) PepSeeker: Mining Information from Proteomic Data. *Methods Mol. Biol.* 484, 319–332
34 Hanash, S.M. *et al.* (2008) Mining the plasma proteome for cancer biomarkers. *Nature* 452, 571–579
35 Gerling, I.C. *et al.* (2006) New data analysis and mining approaches identify unique proteome and transcriptome markers of susceptibility to autoimmune diabetes. *Mol. Cell Proteom.* 5, 293–305
36 Open Proteomic Database (OPD) (http://bioinformatics.icmb.utexas.edu/OPD/)
37 EMBL Proteomics Identifications Database (PRIDE) (www.ebi.ac.uk/pride/)
38 Bredel, M. and Jacoby, E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* 5, 262–275
39 Wuster, A. and Madan Babu, M. (2008) Chemogenomics and biotechnology. *Trends Biotechnol.* 26, 252–258
40 Kwon, H.J. (2006) Discovery of new small molecules and targets towards angiogenesis via chemical genomics approach. *Curr. Drug Targets* 7, 397–405
41 Hu, Y. *et al.* (2003) Analysis of genomic and proteomic data using advanced literature mining. *J. Proteome Res.* 2, 405–412
42 Troyanskaya, O.G. (2005) Putting microarrays in a context: integrated analysis of diverse biological data. *Brief Bioinform.* 6, 34–43
43 Li, S. *et al.* (2006) Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. *Bioinformatics* 22, 2143–2150
44 Gajendran, V.K. *et al.* (2007) An application of bioinformatics and text mining to the discovery of novel genes related to bone biology. *Bone* 40, 1378–1388
45 Jelier, R. *et al.* (2007) Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation. *BMC Bioinform.* 8, 14
46 Louie, B. *et al.* (2007) Data integration and genomic medicine. *J. Biomed. Inform.* 40, 5–16
47 Natarajan, J. *et al.* (2006) Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. *BMC Bioinform.* 7, 373
48 Li, S. *et al.* (2006) Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. *Bioinformatics.* 22, 2143–2150
49 De Chassey, B. *et al.* (2008) Hepatitis C virus infection protein network. *Mol. Syst. Biol.* 4, 230
50 Yue, Q.X. *et al.* (2008) Proteomics characterization of the cytotoxicity mechanism of ganoderic acid D and computer-automated estimation of the possible drug target network. *Mol. Cell Proteom.* 7, 949–961