BRIEFINGS IN BIOINFORMATICS. VOL 13. NO 1. 61–82
Advance Access published on 15 June 2011
doi:10.1093/bib/bbr018

Combining literature text mining with microarray data: advances for system biology modeling

Alberto Faro, Daniela Giordano and Concetto Spampinato

Abstract

A huge amount of important biomedical information is hidden in the bulk of research articles in biomedical fields. At the same time, the publication of databases of biological information and of experimental datasets generated by high-throughput methods is in great expansion, and a wealth of annotated gene databases, chemical, genomic (including microarray datasets), clinical and other types of data repositories are now available on the Web. Thus a current challenge of bioinformatics is to develop targeted methods and tools that integrate scientific literature, biological databases and experimental data for reducing the time of database curation and for accessing evidence, either in the literature or in the datasets, useful for the analysis at hand. Under this scenario, this article reviews the knowledge discovery systems that fuse information from the literature, gathered by text mining, with microarray data for enriching the lists of down and upregulated genes with elements for biological understanding and for generating and validating new biological hypothesis. Finally, an easy to use and freely accessible tool, GeneWizard, that exploits text mining and microarray data fusion for supporting researchers in discovering gene–disease relationships is described.

Keywords: *literature text mining; microarray data; biological databases; knowledge discovery*

INTRODUCTION

A huge amount of biomedical information is hidden in millions of research articles published in the last 20 years and this quantity is bound to increase exponentially [1]. Similarly, the publication biological databases is in great expansion, and a wealth of annotated gene databases, chemical, genomic, clinical and other types of data repositories, including drugs and microarray experiments are available on the Web. Thus a topical challenge of bioinformatics is to leverage on the combination of multi-type information sources, for a more effective system biology modeling and knowledge discovery [2, 3]. A first important step towards the organization and integration of multi-type biomedical information is the National Center for Biotechnology Information's (NCBI) Entrez Cross-Database [4] that interconnects PubMed abstracts with NCBI's databases on DNA sequence and chemical structure, thus speeding up the research of data related to a given disease. However, this system does not include disease–gene or disease–protein compendia and it is not

Corresponding author. Concetto Spampinato, Department of Informatics and Telecommunication Engineering – University of Catania, Viale Andrea Doria, 6 – 95127 – Catania, Italy. Tel: +39 (0) 95 7382372; Fax: +39 (0) 95 7382397; E-mail: cspampin@diit.unict.it

Alberto Faro is Full Professor of Artificial Intelligence at the Engineering Faculty of the University of Catania, Italy, where he is also the Dean of the Computer Engineering degree. His current research interests include: dynamic systems theory of cognition, intelligent learning environments, computer vision and mobile computing.

Daniela Giordano holds the Laurea degree in Electronic Engineering, grade 110/110 cum laude, from the University of Catania, Italy (1990), and a PhD in Educational Technology from Concordia University, Montreal (1998). Since 2001 she is Associate Professor of Information Systems of the DIEEI Department, Engineering Faculty of the University of Catania, where she teaches the graduate level course 'Cognitive Systems and Human-Computer Interaction'. Her research activity has developed along the following tracks: (i) Knowledge Management; (ii) Data Mining, information retrieval and visualization; (iii) Image and signal processing with soft-computing techniques; and (iv) Advanced learning technologies.

Concetto Spampinato received the Laurea in Computer Engineering in 2004, grade 110/110 cum laude, and the PhD in 2008 from the University of Catania, where he is currently Research Assistant. His research interests include image and signal processing for environmental applications, biomedical image processing, multimedia retrieval and bioinformatics.

capable to link its results to external databases, e.g. drugs databases. This linkage is essential to support the cross-linking of textual information with the relevant biological databases and to reinforce the connection between annotations in biological databases. Moreover, given the fast development of high-throughput methods with consequent release of experimental data, the need of developing targeted bioinformatics tools and methods that combine literature, biological databases and experimental data for reducing the time of database curation and for knowledge discovery in literature is heavily demanded [5]. Under this scenario, a first aim of this article is reviewing the knowledge discovery systems that integrate literature information, gathered by text mining, with microarray data. Usually, this is performed either with the goal of enriching the lists of down and up-regulated genes with elements for biological understanding or for generating and validating new biological hypothesis. To illustrate how text mining and microarray data integration may be achieved, we first provide, in the next section, an overview of the methods and tools that are used to perform text mining, i.e. information retrieval (IR), named entity recognition (NER), information extraction (IE) and knowledge discovery (KD), using as a starting point previous reviews [6–9], and focusing on the aspects of the integration between biological data and text data that have been recently investigated. This overview allows us, in 'Combining text and microarray data' section, to point out the differences among the current attempts at integrating experimental data in the mining loop. Finally, in 'GeneWizard' section we illustrate a new tool, GeneWizard that uses microarray data to evaluate and validate biological hypothesis mined from text. GeneWizard, based on the methods proposed by Faro *et al.* [10, 11], proposes novel relationships between genes and diseases by integrating literature discoveries and gene sets gathered from microarray data analysis. In detail, starting from a gene–disease relationship, it extracts a set of genes (related to the gene of the derived association) involved in the disease. Biological functions, namely, biological processes, cellular components and molecular functions, are then associated to the validated set of genes by using Gene Ontology (http://www.geneontology.org/) (GO). Finally, in the conclusions we outline the key challenges to advancing this type of integrated knowledge discovery systems.

TOOLS AND METHODS FOR LITERATURE TEXT MINING

The primary goal of literature text mining [12] is to distill knowledge that is hidden in text of published papers and to present it to the users in a coherent and concise form. More formally, the ultimate goal of text mining concerns the discovery of new, previously unknown information, by automatic text resources processing. Generally, systems for literature text mining include four main modules [13]: (i) IR to gather relevant text by querying databases of biomedical papers; (ii) NER to find the biological entities (e.g. genes, proteins) within text; (iii) IE to identify predefined relationships among biological entities from explicit statements in text; and (iv) KD to elicit relationships hidden in the information derived by the previous module. Recent text-mining systems have started taking into consideration the integration between literature and biological, chemical, medical and drugs databases. In the next sections each module of a text-mining system is reviewed, focusing on how the integration aspect is taken into account.

IR

Information retrieval is the first step of any literature text-mining system and aims at finding documents related to the user's query [9] or at identifying the text segments (articles, abstracts, etc.) pertaining to a specific topic. The most famous IR tool for biomedical papers is PubMed, that is mainly based on two search models: (i) a model that uses Boolean operators to retrieve documents by performing queries in the form of <DiseaseX> and <GeneY> and (ii) a vector space model [14] that represents each document by a vector of index terms, in which each term is characterized by a value according to a frequency-based weighting system. The space vector model is used to train machine learning methods for discriminating relevant papers and irrelevant papers with respect to the queries issued by the user. However, in order to exploit the full potential of IR systems for making scientific knowledge more accessible and enabling automatic knowledge discovery, some systems have extended text-based searching to operate on other sources of data (biological, chemical, medical, drugs annotated databases). Examples of these tools are: (i) Query Chem [15] that combines text-based IR on biochemical databases and WebAPI to retrieve the information and relationships between compound structures and

(ii) EBIMed [16] that retrieves sentences based on co-occurrences between biological entities and identifies relationships between protein/gene names and drugs.

Of course, IR should not rely only on methods for query terms matching, because term ambiguity may cause low precision and low recall. To address this issue, a number of IR tools that exploits established domain ontologies, to support semantic search in biomedical repositories and to guarantee more precision with respect to the Boolean search systems, have been proposed. For instance, GoPubMed (http://www.gopubmed.org/) [17] classifies abstracts using GO terms, GoWeb (http://www.gopubmed.org/goweb) [18] combines keyword-based Web search with text mining and ontologies to organize and navigate the results and facilitate question answering, whereas Textpresso (http://www.textpresso.org/) [19] uses a custom ontology to query a collection of documents for information on specific classes of biological concepts (gene, cell, etc.) and their relations.

IR systems are currently focusing on how to present and distill the search results and how to cross-link these results with the biological databases used in the retrieval process [17, 20]; in fact, long lists of retrieved papers provide a scarce overview of the problem and may create confusion in the users on which sources of data have been used. For example, iHOP [21] converts the information in PubMed into a navigable multi-sources network of genes and proteins (that also includes phenotypes, pathologies and gene function), thus providing an intuitive alternative way of accessing the ten million of abstracts in PubMed.

NER

Biological entities are the backbone of any text-mining system, but often the naming of the entities is inconsistent and imprecise [22] since they are cited with a variety of terms. Therefore, the main goal of a NER system is to find the biological entities (mainly genes and proteins) that are mentioned within a text and to associate them with known names or identifiers (IDs). Usually, this task is performed in two steps: first, the recognition of the words that refer to entities and then, the unique identification of such entities.

The earliest NER systems relied on rule-based approaches (e.g. in [23]), i.e. they were based on manually crafted rules that described common naming structures for certain term classes, based on morphological, orthographic and syntactic characteristics [24]. As annotated corpora (in which gene and protein names are categorized) have become available, the newer systems have relied on machine-learning algorithms [25, 26], to recognize the names on the basis of their peculiar features. Differently, methods relying on dictionaries [27, 28] depend on lists of synonyms of entities names that are matched in documents using algorithms that recognize variations in how the names appear (e.g. gene 'BRCA1' may be written as 'Brca1', 'BRCA 1', 'brca1', etc.). However, the most effective and recent NER systems are based on the curation of entities name lists to reduce the aliases [6,29], even though their main difficulty is the lack of standardization of names (e.g. each gene has many names and abbreviations). Under this scenario, ontologies, taxonomies and controlled vocabularies are of strategic importance for NER systems since they provide semantic interpretation of bio-entities [30–32].

A relevant example of controlled vocabulary that can be used for NER is Medical Subject Headings (MeSH) containing about 30 000 terms and mainly used for indexing articles in MEDLINE (MEDLINE is one of the component of Pubmed that indexes the records using MESH controlled vocabularies.) (i.e. each article is summarized by a set of controlled terms). MeSH covers protein functions in cellular systems, but it is not exhaustive.

Currently, the tendency for NER systems is to integrate different vocabularies or ontologies in order to provide a structured, accurate and complete list of the biological entities that can overcome the aforementioned drawback. In this direction, one valuable approach, based on integration of several controlled vocabularies, is SemCat [33] consisting of a large number of semantically categorized terms coming from different biomedical knowledge resources (e.g. Unified Medical Language System (UMLS) [34], Gene Ontology (GO) [35], Entrez Gene [36], ProtScan [37] and ChemID [38]) and open-domain corpora [39]. An example of NER system based on SemCat was developed by Tanabe *et al.* [40]. This approach builds a priority model for entity recognition based on the position of the words in a sentence, i.e. a word on the right side of a sentence is more likely an entity with respect to a word on the left side. Similarly, knowledge-based NER approaches using platforms that integrate different

types of ontologies from gene terms to genetic pathways (A genetic pathway is a linear sequence of gene activities resulting from the functional interactions between different genes.), to proteins, to clinical trials], such as BioPortal (http://www.bioontology.org/ncbo/faces/index.xhtml) and Open Biological Ontologies (http://www.obofoundry.org/), have been investigated [41].

A recent approach based on web-services is Whatitiz [42] that implements a NER based on morphological variability of terms. In particular, it is provided with numerous modules for annotating different entities: chemical entities (whatizit Chemical), diseases (whatizitDiseaseUMLS for accessing the UMLS Metathesaurus using the tool MetaMap (http://metamap.nlm.nih.gov/), drugs (whatizitDrugs maps drugs in the text with terms of a controlled vocabulary built using Drugbank (http://redpoll.pharmacy.ualberta.ca/drugbank/), and genes (whatizitGO searches for gene ontology terms).

Most recently, algorithms able to identify and disambiguate acronyms automatically, even if these are not mapped in any standard nomenclature, have been investigated to improve NER systems performance [43, 44].

IE

IE from literature aims at extracting pre defined types of facts in the form of relationships between biological entities from the retrieved documents. The inputs to this step are sentences, whereas the outputs are relationships among biological entities. Generally, two main approaches exist:

- Co-occurrences processing: these approaches identify entities that co-occur within the text, i.e. terms that appear in the same texts tend to be related. Often these methods are able to detect co-occurrences to extract single relationships of a certain type: gene–gene, gene–disease, protein–protein, etc. Several works have used co-occurrence frequencies for extracting known single relationships [45]. For example, Al-Mubaid and Singh in [46] proposed a text mining approach based on co-occurrence and term frequency analysis, by which they found and validated six significant genes for Alzheimer's disease.

  Co-occurrences approaches have been investigated also for extracting facts that involve multi-type data, in line with the current research's trend of text and biological data integration. For instance, Mukhopadhyay *et al.* [47] identifies multi-way relationships involving more than two biological entities, i.e. genes, proteins, diseases, drugs and chemicals, etc. An example of identified relationship is 'gene A activates protein B in disease C for organ D under influence of chemical E'. Co-occurrence approaches tend to provide better recall than precision and errors arise in complex sentences containing multiple relationships. These approaches are unable to extract directional relationships (i.e. A involves B but B does not involve A) and to distinguish different types of relationships, e.g. they cannot identify relationships in the form 'A is not connected to B'. Precision can be improved by integrating co-occurrence methods with rule or pattern-based approaches. However, these approaches tend to be dataset dependent, i.e. the rule or pattern sets are derived from training data often not applicable to other data different from the ones used during the training [48].
- Patterns parsing by Natural Language Processing (NLP): all the above mentioned issues are addressed by NLP approaches, which combine the analysis of syntax and semantics in a text for obtaining relationships between facts. The workflow of these approaches is: first, the text is tokenized to identify the boundaries of the words and sentences, then a part-of-speech tagging (e.g. [49, 50]) system assigns labels such as noun, verb, adjective to each word. Afterwards, a syntax tree is computed for each sentence to detect noun phrases and represent their relationships [6]. NER is then used to tag the relevant biological entities in these relationships. Finally, in order to identify the evidences for entities relationships, a rule set based on the syntax tree and on the semantic labels [51, 52] is used. For example Fundel's *et al.* [53] developed RelEx to obtain dependency trees from MEDLINE abstracts by using Stanford Lexicalized Parser (http://nlp.stanford.edu/downloads/lex-parser.shtml). These trees are then enriched with genes and proteins by using ProMiner [54], a dictionary-based NER. Finally, a set of three simple rules (e.g. 'A activates B', 'Activation of A by B' and 'Interaction between A and B') is applied to obtain candidate relationships that are then submitted to a filtering module that uses negation check, enumeration resolution

and restriction to the domain of interest for screening the candidate relationships. Often partial language-parsing approaches are unable to detect relationships that span multiple sentences [6], and full parsing, providing more elaborate syntactic information, is adopted to achieve potentially better results [55]. A typical full grammar parsing example is the Pro3Gres dependency parser [56] that integrates hand-written grammar with a statistical language model for parsing unrestricted text by using deep knowledge of the English language.

Full parsing approaches have been recently integrated with multi-type data for extracting facts involving more concepts: InfoPubMed [57] recognizes different types of interaction between gene and proteins on MEDLINE abstracts by combining full parsing, machine learning techniques and ontologies for NER; Pharmspresso [58], identifies important pharmacogenomics facts in articles referenced to human genes, polymorphisms, drugs and diseases by full text parsing and by exploring biological, chemical and drugs databases.

An approach based on full-text processing that also uses the 'not' concept in proteins relationships, i.e. protein A binds protein B but not protein C, was developed by Kim [59]. They found 41 471 protein–protein contrasts available at the web-address http://biocontrasts.biopathways.org/.

The current trend is to favor full texts over abstracts since biological entities identified from mining only abstracts can be strongly underestimated because of abstracts' concise nature. Methods for mining full biomedical texts need to be improved substantially, especially in converting PDF or HTML documents to plain text and in handling grammatical errors [60]. Another shortcoming of current methods is that they do not consider information hidden in tables and figures. Recently, approaches that integrate text data, biological databases and non-textual data (e.g. images, graphics, etc.) have been proposed and a comparative list is provided in [7]. An example is SLIF [61] that combines figure's caption mining, image processing and specific domain ontologies to extract biomedical information from fluorescence microscopy images. A biological entity recognition system finds protein and cell type names in the mined captions and these entities are associated with the patterns extracted from the related images. Finally, a web-interface and a XML-based web-service allow users to investigate and query the derived information.

KD

Trying to discover hidden or implicit biomedical links and to propose them as potential scientific hypotheses is the main goal of knowledge discovery systems. In fact, the previously described IE systems extract only pre identified or explicit relationships. Swanson, pioneer of the research in knowledge discovery from text, in [62] demonstrated, by using the semi-automated Arrowsmith system [63], how new knowledge can be inferred from existing literature. Inferring indirect relationships implies to use facts in the form A leads to B and B leads to C, then a relationships may be inferred between A and C. In detail, the user provides a hypothesis between two biological entities (A is related to C) that is further proved by searching for related terms (B) supporting the given hypothesis. An example of inferred relationships is the one 'fish oil - Raynaud's disease' discovered by Swanson [64] or the relationship between magnesium deficiency and migraine headache [65]. These two discoveries were confirmed experimentally [66, 67]. Several methods relying on natural language processing exist to discover knowledge about gene regulation [68], protein phosphorylation [69, 70], gene–disease or gene–gene interaction [71–73].

One of the most complete systems that uses NLP is GeneWays [74] that examines entire articles to extract the physical interactions among disease and genes hidden in the literature. Differently, many other systems are based on co-occurrence, i.e. the idea that two concepts (biological entities) are related if they occur in the same contexts in the literature. They can be based either on (i) first order co-occurrences, e.g. entity A co-occurs with entity B [73], [10] or (ii) second-order co-occurrences [75, 76], i.e. entity A co-occurs with entity B which co-occurs with entity C, therefore there is a relationship between entities A and C. These approaches share the assumption that hidden and valid relationships may be found by suitably screening the huge number of facts retrieved by the co-occurrence approaches. For instance, Jelier *et al.* [76] propose the associative concept space (ACS) to filter the irrelevant relationships and terms obtained by applying a second order co-occurrence approach. In detail, ACS reflects not only the co-occurrence of

two entities, but also indirect, multi-step relationships between entities.

Very few systems have been designed to extract complex and multiple types of relationships (e.g. find all the genes involved in a disease and all the related proteins) that necessarily require different types of data to be integrated. Anni 2.0 [77] uses an ontology-based interface to MEDLINE to identify different types of associations between biomedical concepts, including genes, proteins and diseases. It resorts to the idea of concept profiling, i.e. a list of concepts is presented where each concept is associated to the analyzed text together with a score describing its importance. An example of association derived with Anni 2.0. is: 'Gene KLK3 is bound to the prostate cancer, more specifically with malignant neoplasm of prostate'. Polysearch [78] is a recently developed web-tool able to identify associations from published abstracts and many well-annotated databases. It enables users to perform queries in the form: 'Find all genes associated with a prostate cancer'. Up to date it supports more than 50 classes of queries mining more than a dozen of text, abstracts and bioinformatics databases. A key functionality of PolySearch is that it extracts and analyzes text data not only from PubMed but also from other databases such as DrugBank [79] and Human Gene Mutation Database (HGMD) [80]. Gendoo [81] identifies disease relevant genes and aims at understanding their mechanisms by interpreting data provided by genome sequences and transcriptomics. In detail, in Gendoo the On-line Mendelian Inheritance in Man (OMIM) (http://www.ncbi.nlm.nih.gov/omim/) knowledge-based system, that contains about 20 000 entries for human genes and for genetic diseases, is re-organized by using MeSH, thus improving OMIM's exploitability by computer automation.

In summary, several KD methods have been proposed in the last years where integration between biological data and unstructured/structured text has been achieved in at least one of the IR, NER and IE sub-systems. However, to realize the full potential of text mining, new methods that integrate complex texts, biological and also raw experimental data are needed, with a focus on enabling biologists to exploit biological knowledge more effectively. This is necessary because any knowledge discovery method generates hypotheses about relationships to be validated empirically, and reusing available experimental data is an effective strategy to speed up scientific progress.

COMBINING TEXT AND MICROARRAY DATA

Literature text-mining methods are useful to discover hidden or indirect relationships, however their integration with high-throughput methods (e.g. microarray) is heavily demanded. To fulfill this need, in the last years bioinformatics efforts have been directed toward the implementation of tools supporting the integration of biological and experimental data with literature information in order to infer biological hypothesis that can assume the form of pathways, gene regulatory networks, or, more in general, biological networks involving different entities such as genes, proteins, diseases, drugs from experimental data gathered from high-throughput methods.

DNA microarray technology, one of the most common high-throughput methods, allows researchers to come across biological functions on a genomic scale. However, the list of the produced down and upregulated genes is very cryptic, thus requiring a huge effort in data interpretation [13, 82]. Moreover, the selection of such lists of genes (i.e. the clusters to be analyzed) is demanded to the researchers that, given the amount of data involved, might not pursue the best selection since it is difficult to catch the correlation between the cluster and the biological aspect to be investigated.

Therefore, in the last 10 years, the attention of the bioinformatics community has been directed mainly to eliciting such correlations to understand the biological meaning of the produced lists of genes, instead of investigating novel clustering and statistics methodologies. Understanding the biological meaning of a set of up and downregulated genes derived from microarray experiments is one aspect of current bioinformatics efforts in data integration; the other one foresees text mining combined with microarray data targeted to the generation and to the evaluation of biological hypotheses, which can be obtained either by mining microarray data that involves data clustering and manual selection of the clusters to be analyzed or by mining the literature using the knowledge discovery systems before described or by exploiting the knowledge of the biomedical researchers.

Understanding biological meaning

The most natural way to assign a biological meaning to a set of genes that has been obtained by mining microarray data is to project it onto biological

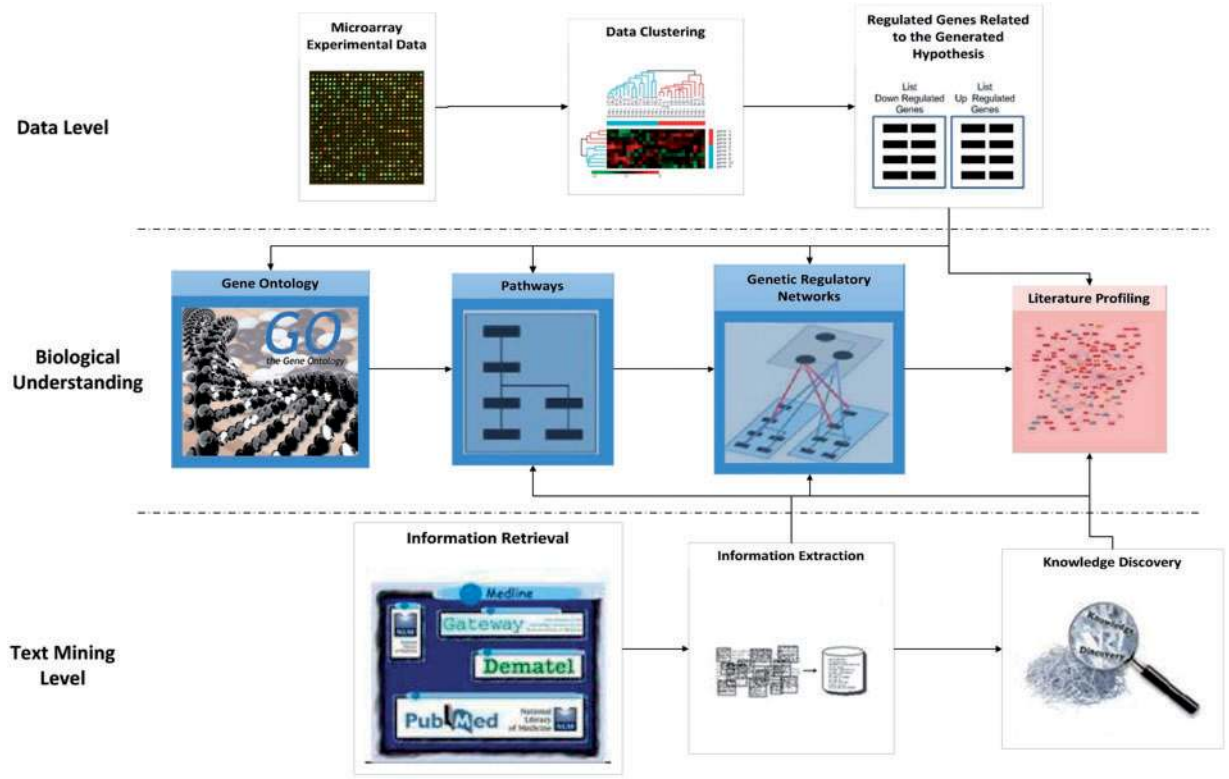

Figure 1: Understanding biological meaning of a set of regulated genes. The most common ways for understanding the biological meaning of a set of genes are: (i) to project it onto biological processes represented in the form of GO terms, pathways and gene regulatory networks (blue rectangle) and/or (ii) to annotate the lists of regulated genes based on literature profiling (red rectangle). Pathways and gene regulatory networks are usually derived by manual literature analysis.

processes that can be represented in the form, in order of increasing complexity, of GO terms, pathways and gene regulatory networks (A gene regulatory network represents a collection of segments of DNA that maps gene regulations in living cells.) previously identified or manually compiled by researchers [82] (see Figure 1).

The most common (and simple) approach for gene list projection is to use GO for interrelating a list of genes with a biological process and/or a molecular function and/or a cellular component. GO is also used to rank significant genes (produced in the microarray experiment) in relationship to the GO categories. A list of about 70 methods and tools that carry out GO-based microarray analysis is reviewed in [83]. An approach that goes beyond simple GO classification is Onto-Express [84], since it associates lists of up and down regulated genes with functional profiles built by correlating GO terms (biological processes, chemical components, molecular functions) with expression profiles.

However, the GO classification does not provide exhaustive information about the biological context of a given set of up and downregulated genes. This can be achieved by pathway analysis and/or by regulatory network analysis. Pathway analysis mainly investigates the functional and physical interaction among genes instead of using the gene-centered view as GO-based approaches. These systems try to map genes derived from microarray experiments onto precompiled pathways derived by manually analyzing the literature. Most non-commercial systems for pathway analysis rely on the KEGG database (http://www.genome.jp/kegg/) that contains a collection of pathways representing the current knowledge on gene and molecular interaction. A comprehensive list of tools for pathways analysis can be found at the weblink http://www.geneontology.org/GO.tools.microarray.shtml. Pathway mapping of microarray data, usually, generates more than one pathway, therefore, it is necessary to rank them according to their relevance to the dataset. Pathways ranking is provided in GenMAPP

2 [85], where the users can rank, and at the same time, customize the pathways; an extension of GenMAPP2 [85] proposes the Fisher's exact test for ranking the relationships between genes and pathways. PathExpress [86, 87], instead, identifies the most relevant metabolic pathways associated with a subset of genes using *P*-values. The KEGG-based web-tool KOBAS [88] proposes a controlled vocabulary for gene pathways mapping and the relevance of the discovered pathways is estimated using binomial, Chi-square and hypergeometric distribution test. Although pathway-based approaches provide deeper information on biological processes possibly relevant to a set of genes, their main shortcoming is that biological processes usually depend on more than one pathway and the connections between such pathways is related to the biological context. The interconnection of pathways is defined as gene regulatory network. This network cannot be easily derived by simply combining pre compiled pathway because the networks' morphology changes with the biological context. The earliest attempts for building gene regulatory networks have been successful only for lower eukaryotes with simple genomes [89, 90]. Current approaches (both stand alone and also combined with GO classification and pathways), instead, are directed toward more complex mammalian systems. For instance, ARACNe [91] builds regulatory networks in mammalian cells by identifying transcriptional interactions among genes from microarray expression profiles. An interesting effort is represented by MONET [92] a method based on Bayesian networks for inferring gene regulatory networks. It mainly consists of two steps: the first aims at splitting the whole gene set into overlapped groups that contain genes whose GO annotations or microarray expression patterns are highly correlated. Finally, the second step infers Bayesian networks over each group and integrates such groups into global regulatory networks. BioCAD [93] integrates both the above inference tools (ARACNe and MONET) for building gene regulatory networks. The tool also supports validation of the inferred networks by integrating gene and protein regulatory networks derived from MEDLINE abstracts using a text-mining system based on STRING-IE [94].

The described approaches provide as outcomes precomputed relationships between genes and biological processes. However, the literature may enrich the information about relationships regulated genes-biological processes much more than structured ontologies or precompiled pathways can do. To extract the additional information hidden in the literature, several methods that annotate the lists of regulated genes based on literature profiling have been proposed [95, 96]. Most of these approaches are based on keywords over-representation of a set of genes, similarly to GO-based microarray analysis, but where the keywords to be associated to the gene set are gathered by mining directly MEDLINE and they are used to interpret genes in domains scarcely covered by GO. In detail, such methods retrieve a subset of MEDLINE abstracts associated with one or more genes, e.g. a cluster of genes derived by gene set analysis methods [97, 98]. Then, these abstracts are used to identify relevant keywords in the text or annotated MeSH terms (medical subject heading terms), thus helping the gene sets characterization. For example, GenClip [99], one of the most recent tools, builds functional clusters of genes related to disease pathogenesis starting from a list of genes from microarray. The tool first identifies keywords as terms that co-occur in at least two of the analyzed genes by mining literature abstracts and then clusters the list of genes based on keyword occurrences, thus obtaining functional clusters. Differently, Chagoyen *et al.* [100] proposes a system for literature profiling of large sets of genes or proteins that can be used to find similarities among genes. The method starts from creating a pool of documents related to a specific gene. Afterwards, the pool of documents is converted into a vector space representation and finally, the non-negative matrix factorization [101] is applied to the vector space, thus obtaining for each gene a literature profile (A literature profile can be seen as a picture of the functional relationships, derived from scientific papers, between set of genes). CoPub [102], provides an insight into the biological mechanisms related to a set of regulated genes for liver pathologies by calculating statistics for gene-keyword co-occurrences using the entire MEDLINE abstracts, instead of only a subset, as the previous approaches do. The inputs of the tool are a subset of genes obtained by microarray data processing and a set of keywords, whereas a navigable network of MEDLINE abstracts where the genes and the keywords co-occur is provided as output. The text mining method extracts networks of abstracts by analyzing the co-occurrences of human, mouse and rat genes with keywords describing liver pathologies, pathways, GO terms, diseases, drugs and tissues. An

Table 1: List of the web available tools for understanding biological meaning of a set of regulated genes derived from experimental data

GenMAPP2	**Description** **Visualize gene expression data biological pathways**	**Used Resources** **GO, KEGG**	**Available at** **http://www.genmapp.org/**
PathExpress	Mapping of a set of Genes onto Pathways.	KEGG Swiss-Prot database[a] Blastx[b]	http://bioinfoserver.rsbs.anu.edu.au/utils/PathExpress/
KOBAS	Identify statistically enriched pathways for a set of genes or proteins	Pathways Database: KEGG, PID Curated[c] BioCyc[d] and Panther[e]	http://kobas.cbi.pku.edu.cn/home.do
ARACNe	Estimate gene regulatory networks in mammalian cells using microarray expression profiles	Expression profile dataset of human B lymphocyte cells built by the authors	http://wiki.c2b2.columbia.edu/califanolab/index.php/Software/ARACNE
GenCLIP	Clustering of gene lists by literature profiling	NCBI EUtilities for Text Mining Gene List: HUGO[f] Entrez Gene, or Unigene[g]	http://www.genclip.com
CoPub	Find biomedical concepts from Medline linked to a gene set (Affymetrix identifiers)	NCBI E-Utilities for Text Mining	http://services.nbic.nl/cgi-bin/copub/CoPub.pl

[a]http://expasy.org/sprot/; [b]http://blast.ncbi.nlm.nih.gov/; [c]http://pid.nci.nih.gov/; [d]http://biocyc.org/; [e]http://www.pantherdb.org/pathway/; [f]http://www.genenames.org/; [g]http://www.ncbi.nlm.nih.gov/unigene.

approach that exploits literature for supporting gene list interpretation is the one proposed by Jelier *et al.* [103], which uses associations derived from literature (using the Anni 2.0 tool) to provide an interpretation of gene expression changes. In detail, they propose the literature-weighted global test to compute the correlation between associations (in the form gene-biomedical concept) obtained by mining the literature and list of genes extracted from microarray and they provide as output the scores reflecting the importance of a gene for a concept of an association. Table 1 shows some of the web available tools for understanding biological meaning of a set of regulated genes derived from experimental data.

Hypothesis generation

Hypothesis generation has the objective to suggest undiscovered associations between biomedical concepts; this is different from the attempts at providing a biological meaning to a set of facts (e.g. lists of genes) extracted from experimental data. The most explored venue for hypothesis generation concerns the discovery of genes and other biological entities (together with their role) involved in a specific disease (disease-centric analysis). In fact, predicting biological entities (and their role) involved in a disease before experimental analysis may save time and effort by indicating where the research should look into. Text mining and microarray data have been combined in two main ways to achieve this goal: (i) starting from a microarray related to a specific disease, a list of genes (the hypothesis) is extracted (e.g. one or more cluster) and then the role of such genes (i.e. the prioritization) in the given disease is explored using information extracted from literature and (ii) the hypothesis is generated from literature mining in the form of associations between genes and a given disease and then these associations are filtered and validated by resorting to microarray data.

The first approaches follow this workflow: given a set of genes (the hypothesis) either gathered directly from microarray data analysis or previously stored in public databases (such as Gene Expression Omnibus), the literature (mainly MEDLINE) is mined, starting from this set of genes, in order to elicit a gene prioritization for the given disease or to find out other biological concepts involved in the same disease (Disease modeling) (see Figure 2). The workflow is similar to the biological understanding approach's one with the difference that in this case the outcome is a refinement and a close examination of the input hypothesis regarding concepts and their role in the given disease, whereas in the context of biological understanding the output is the association of a meaning to a list of genes (facts).

One of the most complete tools for hypothesis generation from microarray is G2D [104, 105]. It performs genes prioritization related to inherited diseases by combining Mesh annotations in MEDLINE and a set of genes with the GO annotations of entries

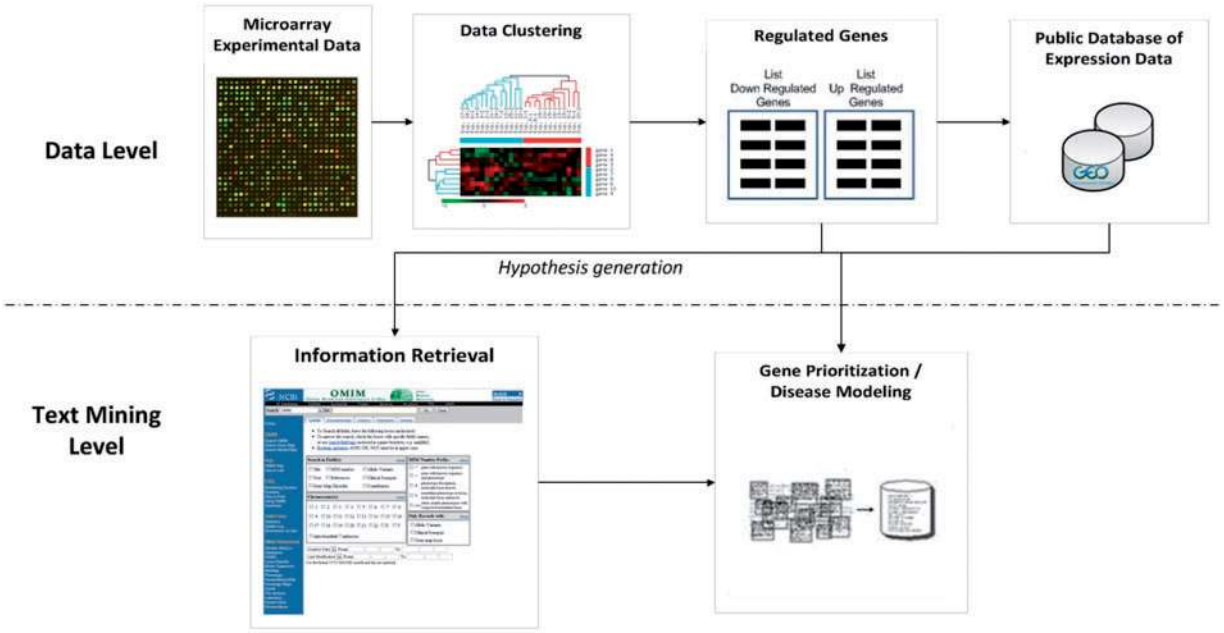

Figure 2: Hypothesis generation by microarray data analysis. The first approach follows this workflow: given a set of genes (the hypothesis) either gathered directly from microarray data analysis or previously stored in public databases, the literature is mined, starting from this set of genes, in order to elicit a gene prioritization for the given disease or to find out other biological concepts involved in the same disease (Disease modeling).

in NCBI RefSeq [106] (collection of annotated sequences, including genomic DNA, transcripts and proteins). More specifically, it receives as input a genomic region and an OMIM disease identifier and provides as output the genes potentially involved in the given disease. In detail, for the disease under analysis the MESH terms from the 'Disease Category' associated to publications in OMIM are retrieved. These terms are then associated with chemical, drugs and molecular functions by using GO. The detected molecular functions are used to identify a sequence of DNA by querying the RefSeq protein database. This sequence is integrated with a chromosomal location for the given disease provided by the OMIM database in order to obtain a list of genes related to the analyzed disease. G2D was originally developed only for Mendelian diseases; currently, it also works for complex genetic diseases [107]. Likewise, Tiffin *et al.* [108] propose genes prioritization according to the relationship disease-affected tissue. The method integrates literature discoveries (co-occurring disease and tissue names in MEDLINE) and human gene expression data from the Ensemble database [109] to link gene expressions to diseases by using an anatomical ontology. First, the tool associates anatomical terms from an ontology for human anatomical systems and cell types (eVOC [110]) to diseases names, based on the co-occurrence in Pubmed abstracts. Each term of the eVOC ontology is then ranked according to the frequency of annotation. The top-scoring terms are compared with the terms already annotated to candidate disease genes using the Ensemble database. The genes that mismatch with the already annotated genes represent the list of genes to be explored.

The second approaches (Figure 3) generate hypotheses (in the form of associations gene-disease) by mining literature, then they validate such hypotheses by checking if there is any evidence of the discovered relationships in the experimental data. To the best of our knowledge, few methods use knowledge gathered from the literature for hypotheses generation and validate these sets using high-throughput methods, thus allowing the identification of novel biological entities relationships.

An approach in this direction is the one proposed by Faro *et al.* [11], where hypothesis generation about gene–diseases relationships is made by mining specialized literature using the co-occurrence processing approach described in [10] and the inferred relationships are selected and then validated by means of microarray data analysis. The used text-mining algorithm tends to provide better recall than precision, i.e. it provides more relationships

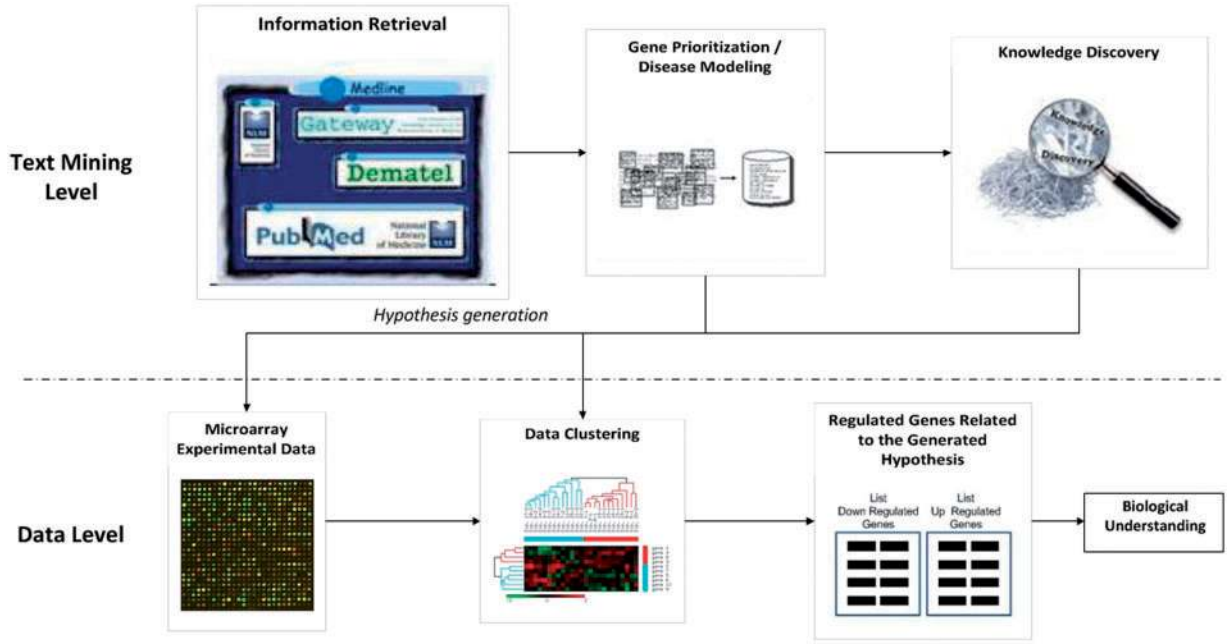

Figure 3: Hypothesis generation by literature text mining. These approaches generate biological hypothesis by mining literature, then they check if there is any evidence of the discovered relationships in the experimental data. If the found relationship is validated, then they investigate the other biological entities (mainly genes) involved in this relationship. They also use the methods described in previous section to provide a biological insight of the achieved findings.

than the ones obtained with other existing methods. The first selection from the pool of obtained relationships is performed on the basis of the availability of relevant raw experimental data. Resorting to microarray data, of course, serves both as a verification of the derived relationships and as a discovery of novel lists of genes related to a specific disease. This latter, in particular, is achieved by exploring down regulated and upregulated genes through a gene relevance network (A GRN is a group of genes whose expression levels in a microarray dataset are highly predictive of others genes in the group.) (GRN) related to the gene of the discovered relationship. In order to understand the biological meaning behind the obtained genes, molecular, biological processes, cellular components and molecular functions are pointed out by querying the GO database. The approach is implemented in 'GeneWizard', a tool that will be discussed in detail in the next section.

Another tool for biological discovery that validates hypothesis by integrating multiple types of data is ENDEAVOUR [111], which filters a set of candidate genes indirectly connected to a given disease according to chromosomal-mapping data about the disease. In detail, the method takes as input a list of genes (possibly extracted by literature text mining) potentially involved in the given disease and provides as output the prioritized genes list. This list includes all the genes involved in the disease ranked according to a score for a specific data source. For example, with ontology-based data sources, the genes are ranked according to the significance of the related terms (the ones over represented in the input gene lists), whereas with a microarray data source the genes are ranked according to the probability of being involved in a disease. Currently, the data sources supported by ENDEAVOUR are ontologies, interactions, gene expressions, regulatory information, sequence-based data and literature data.

The above approaches (GeneWizard and ENDEAVOUR) can be differentiated on the basis of how they combine experimental and literature data. In fact, ENDEAVOUR performs gene prioritization by integrating heterogeneous and multiple data sources, whereas GeneWizard integrates literature facts in the microarray mining loop, i.e. the selection and the analysis of the microarray data clustering is 'literature-driven'. In detail, the selection of the cluster to be explored is based on the presence of the gene of the mined association. This is a

Table 2: List of the web-available tools for knowledge discovery based on integration between literature data and experimental data

	Description	Used Resources	Available at
GeneSeeker	Gene Prioritization located on a specified human genetic location and expressed in a specified tissue	Expression Data: MEDLINE, OMIM, SwissProt Cytogenetic data: MIMMAP, GDB Cytogenetic data: MIMMAP, GDB	http://www.cmbi.ru.nl/GeneSeeker/
G2D	Gene Prioritization related to an inherited disease	NCBI RefSeq for annotated DNA sequences and MESH, OMIM	http://www.ogic.ca/projects/g2d2/
GeneWizard	Discovery of gene–disease associations and biological understanding	NCBI E-Utilities for Text Mining GEO for expression data profiles GO for genes annotation	http://i3s-lab.ing.unict.it/GeneWizard
ENDEAVOUR	Prioritization of genes list underlying biological disease using several sources of data	GO, SwissProt, Blast CisRegModule and SonEtAl for expression data	http://www.esat.kuleuven.be/endeavour

novelty in bioinformatics tools and it is one the major strengths of GeneWizard, because it uses experimental data-evidence for knowledge discovery, whereas ENDEAVOUR uses what already exists in form of annotations. An approach that performs literature-driven gene clustering for biological understanding of regulated genes is GenClip (already mentioned in the previous section), where genes are clustered according to their literature profiles. This way of proceeding is different from the GeneWizard's one; in fact, GenClip creates functional clusters of genes according to their co-occurrences in literature abstracts thus leading to discoveries in the form: 'the genes G_1, G_2, G_3 are involved in diseases D_1, D_2 and are related to the Biological Processes P_1 and P_2', but it is not possible to elicit all the concepts related to only one disease and, moreover, there is no evidence that the genes are really involved in the extracted diseases (not experimental data driven). Differently, GeneWizard creates cluster of genes involved in the disease D of the inferred relationship $D - G_1$ (obtained through text mining) starting from the data-driven evidence that the gene G_1 is possibly involved in the given disease D (i.e. in the microarray data the gene G_1 is differentially expressed). Moreover, the derived list of genes (due to the clustering) is then associated with GO terms in order to explore the terms involved in the disease D. Therefore GeneWizard leads to discoveries in the form: 'the genes G_1, G_2, G_3 are surely involved in disease D and possibly are responsible of the biological processes P_1 and P_2 in the disease D'.

Table 2 lists the web available tools analyzed in this review that combine literature data with experimental data for biological hypothesis generation. The comparison among these tools is only functional (what they achieve and which resources they use), and it is not based on performance, since the performance's evaluation of knowledge discovery tools is still challenging, especially because the definition of 'discovery' is controversial [112, 113].

In the next section a description of the tool proposed by the authors, GeneWizard, is given.

GENEWIZARD

GeneWizard is a user-centered application that allows the users to produce easily new biological hypotheses through an intuitive and guided interface without requiring knowledge of text-mining and data-mining methods. It retrieves automatically gene–disease relationships by mining Pubmed abstracts and validates them with microarray experiments, gathered from the public GEO database (http://www.ncbi.nlm.nih.gov/geo/). Moreover, it is able to build a GRN by microarray data analysis and, finally, to provide biological insights by mapping the obtained gene relevance network onto GO. GeneWizard is a five-step wizard system that leads the user during the experiments and its workflow (shown in Figure 4) is as follows:

(1) retrieval of the information needed to search and discovery biological relationships starting from any disease as query term. A set of genes (Entrez Gene), a set of diseases (MeSH), and a set of biomedical scientific abstracts (PubMed) are identified for a specific disease by querying,

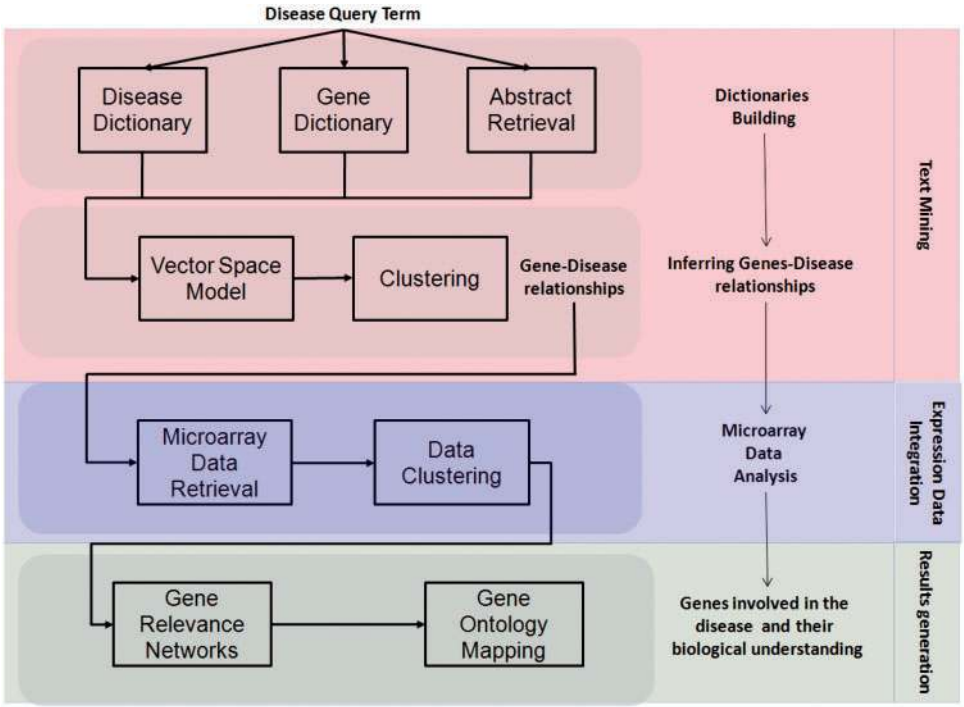

Figure 4: GeneWizard's workflow.

respectively, Entrez Gene, MeSH and PubMed. For each dataset a suitable dictionary is built;

(2) text mining of the retrieved scientific abstracts to build the relationships based on co-occurrences, according to the methodology proposed by Faro *et al.* [10];

(3) scrutiny and validation of each relationship through the analysis of specific microarray datasets available in public repositories of gene profile expressions (GEO database);

(4) analysis of selected gene expression data to generate GRNs for each gene–disease relationship; and

(5) finally, the genes involved in the relevance networks are mapped onto specific biological processes, molecular functions and cellular components using GO.

The above five steps are mapped into three main sections of the tool, namely, *Text Mining*, *Expression Data Integration* and *Results Generation* that will be described in the next subsections. Table 3 lists the resources used by each module shown in Figure 4.

Text mining

The text-mining approach implemented in GeneWizard builds gene–diseases relationships following the co-occurrences method proposed by Faro *et al.* [10]. It consists of four steps: (i) Pubmed abstracts querying and retrieval; (ii) parsing and indexing using term identification; (iii) abstract clustering based on document similarity; and (iv) relationships discovery between meaningful entities. In particular, the retrieved abstracts are first converted into a sequence of words (parsing), then each abstract is represented by vectors (Vector Space Model) containing how many times each gene and each disease appear in it (indexing). Gene and disease identification is carried out using a dictionary-based approach, i.e. dictionaries for genes and disease are built by accessing available external web data sources. In detail, the Entrez web services (http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi) are used to query the biological, chemical and medical databases available through MeSH and Entrez Gene. MeSH is used as the dictionary for diseases, whereas Entrez Gene to build the gene dictionary. The abstracts are retrieved by querying MEDLINE (see Figure 5).

After dictionary building, three subsets are created: (i) vectors indexed on gene terms with null components on the disease space, (ii) vectors indexed on diseases terms with null components on the gene space, and (iii) vectors with non null components on both the gene and disease spaces. Finally, vector similarity matrices are built for the first and second

Table 3: List of the resources used by GeneWizard

Section	Functionality	Used Resource	Available at
Text mining	Disease dictionary building	MESH	http://www.ncbi.nlm.nih.gov/mesh
	Gene dictionary building	Gene Entrez	http://www.ncbi.nlm.nih.gov/gene
	Abstracts retrieval	Pubmed	http://www.ncbi.nlm.nih.gov/pubmed
Expression data integration	Microarray data retrieval	GEO datasets	http://www.ncbi.nlm.nih.gov/gds
	Data clustering	MeV Java classes	http://www.tm4.org
Results generation	GRN	MeV Java classes	http://www.tm4.org
	Gene ontology mapping	Gene DAVID	http://david.abcc.ncifcrf.gov

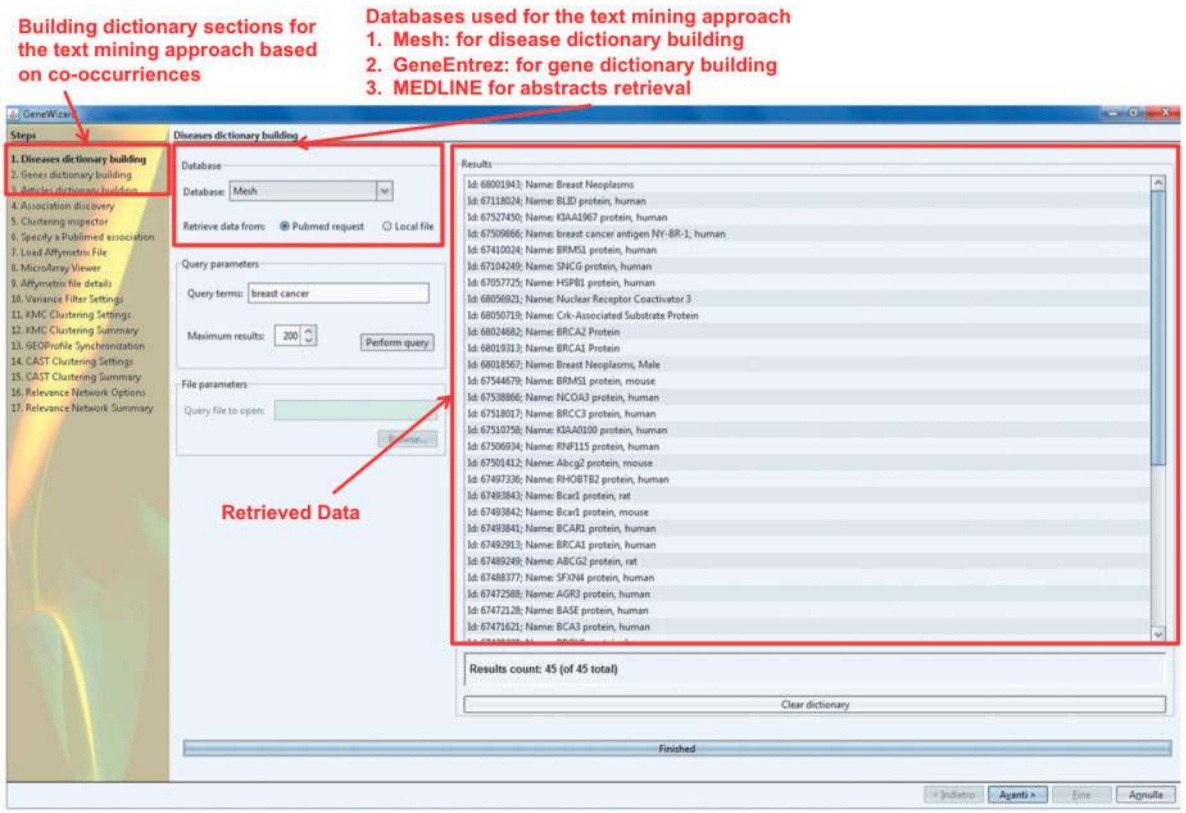

Figure 5: Dictionary building sections. The text mining approach developed in GeneWizard is based on the Vector Space Model (VSM) representing the retrieved MEDLINE abstracts as vectors whose elements are the frequency of the gene/disease, retrieved from the gene and the disease dictionaries, in the document. To do that, gene and disease dictionaries must be built. This is achieved by means of Entrez web services: namely, MeSH for disease dictionary, Entrez Gene for gene dictionary and Pubmed for retrieving the abstracts to be mined.

subset, whereas for the third set two similarity matrices are computed, i.e. respectively, the genes components and the diseases components. Each of the four similarity matrices can be clustered by either the k-means or the hierarchical clustering. For each cluster the set of its positive features, i.e. the terms occurring in a cluster with a frequency above a prefixed threshold, is evaluated; then the relationships between genes and diseases are inferred by intersecting the clusters of the two similarity matrices derived from the third set of similarity matrices. Figure 6 shows an example of such clusters, obtained by querying the system using the term 'Breast Cancer'.

The mining approach implemented in GeneWizard is based on terms co-occurrences, but it differs substantially from the ones described in 'KD' section. In fact, generally, these approaches

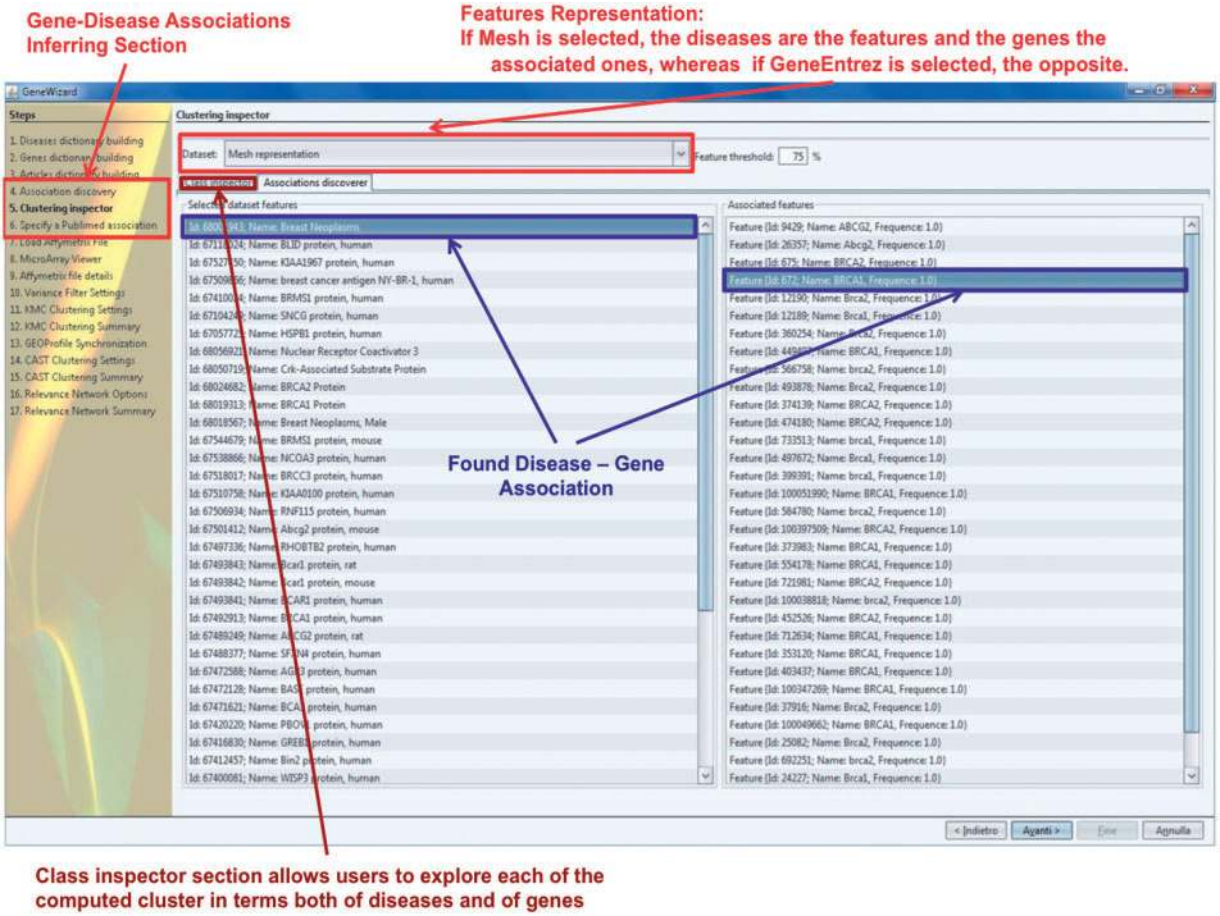

Figure 6: Inferred gene–disease relationships. This section allows users to visualize all the discovered relationships (in the figure related to 'Breast Cancer'). In particular, by clicking on the disease on the left window the tool shows all the related genes (extracted by the text-mining algorithm) on the right window (if Mesh representation is selected, otherwise the opposite). Class inspector section allows users to explore each of the computed cluster in terms both of diseases and of genes. In figure the relationship between the disease 'Breast cancer' and the gene BRCAI is discovered and will be further investigated by integrating microarray data.

rely on the fact that a relationship $T_1 \leftrightarrow T_2$ between two biomedical terms T_1 and T_2 can be derived by using these inferences: $T_1 \leftrightarrow T_x$, $T_x \leftrightarrow T_2$, and the relationships $T_1 \leftrightarrow T_x$, and $T_x \leftrightarrow T_2$ are explicitly derived from a text. An example is the relationship 'migraine – magnesium' discovered by Swanson [65] by identifying the intermediate medical term (the one we called T_x) 'calcium channel blockers' that occurs frequently in the magnesium literature and the migraine literature. Differently, GeneWizard's approach infers a relationship $T_1 \leftrightarrow T_2$ by finding two relationships $T_1 \leftrightarrow T_x$ and $T_y \leftrightarrow T_2$ with T_x and T_y belonging to the same cluster. For example, the association 'migraine – magnesium' is derived if GeneWizard finds a term related to migraine (T_x) and a term related to magnesium (T_y) and if T_x and T_y are clustered together. This, of course, produces numerous relationships (high recall), and the screening of the most promising ones is achieved by exploring experimental data.

Expression data integration

The relationship selected by the user from the set of relevant relationships proposed by the tool is then evaluated/validated by resorting to the microarray data available from the GEO database using the disease of the given relationship as query term.

Once a microarray dataset has been selected, the tool starts the analysis to obtain a GRN (see the left side of Figure 7) (a list of relevant genes for the given disease) that contains the gene of the selected relationship. The microarray analysis modules are based on the Java classes from the MEV (MultiExperiment Viewer) software [114]. The first step is to cluster, by

Figure 7: Microarray data analysis: microarray data retrieval and data clustering. GeneWizard allows the users to retrieve microarray datasets for the disease of the discovered relationship (in our example Breast Neoplasms – BRCA1). For example, in this figure the microarray data is related to the disease Breast Neoplasms (left part of the image), whereas the screenshot on the right side shows the clustering results after the application of KMC. The cluster highlighted is the one that includes gene BRCA1 (whose AFFY-ID is 204531_s_at) and is analyzed by the next step in order to find a gene relevance network that may be involved in the given disease.

K-means or Hierarchical Clustering (KMC Section), the microarray data to obtain homogeneous gene sets. If the number of genes of the selected dataset is still large, a filtering may be applied using the variance of the gene expression levels. Then the cluster that contains the gene under examination (i.e. the one of the chosen relationship) is automatically selected with the assistance of another call to the GEO Profiles to match the gene identifier (AFF-ID) in the microarray experiment with the real gene name. Therefore, the outcome of this step is a cluster containing the gene of the selected relationship. Starting from it, in the next step the list of genes (see the right side of Figure 7) (and its biological meaning) related to the given disease will be explored.

Results generation

Starting from the selected gene set (the cluster) GeneWizard provides Cluster Affinity Search Technique (CAST) [115] to compute gene relevance networks (Figure 8). Affinity is a similarity measure between a gene and all the genes in a cluster, based on the expression profile. Therefore, starting from a relationship between a gene and a disease, GeneWizard is able to extract a list of genes involved in the given disease. As stated in the previous section, GeneWizard differs from other approaches that combine experimental and literature data since it follows a 'literature-driven' microarray data analysis. Usually, microarray data analysis applies the CAST algorithm [115] in a blind manner to the entire gene set (each microarray may contain more than 100 000 of rows and columns) obtaining a large set of gene relevance networks (GRN) difficult to be understood. Instead, in GeneWizard's approach, after a simple KMC clustering, the selection of the cluster is performed by taking into account the discovered gene–disease relationship. After the application of the CAST

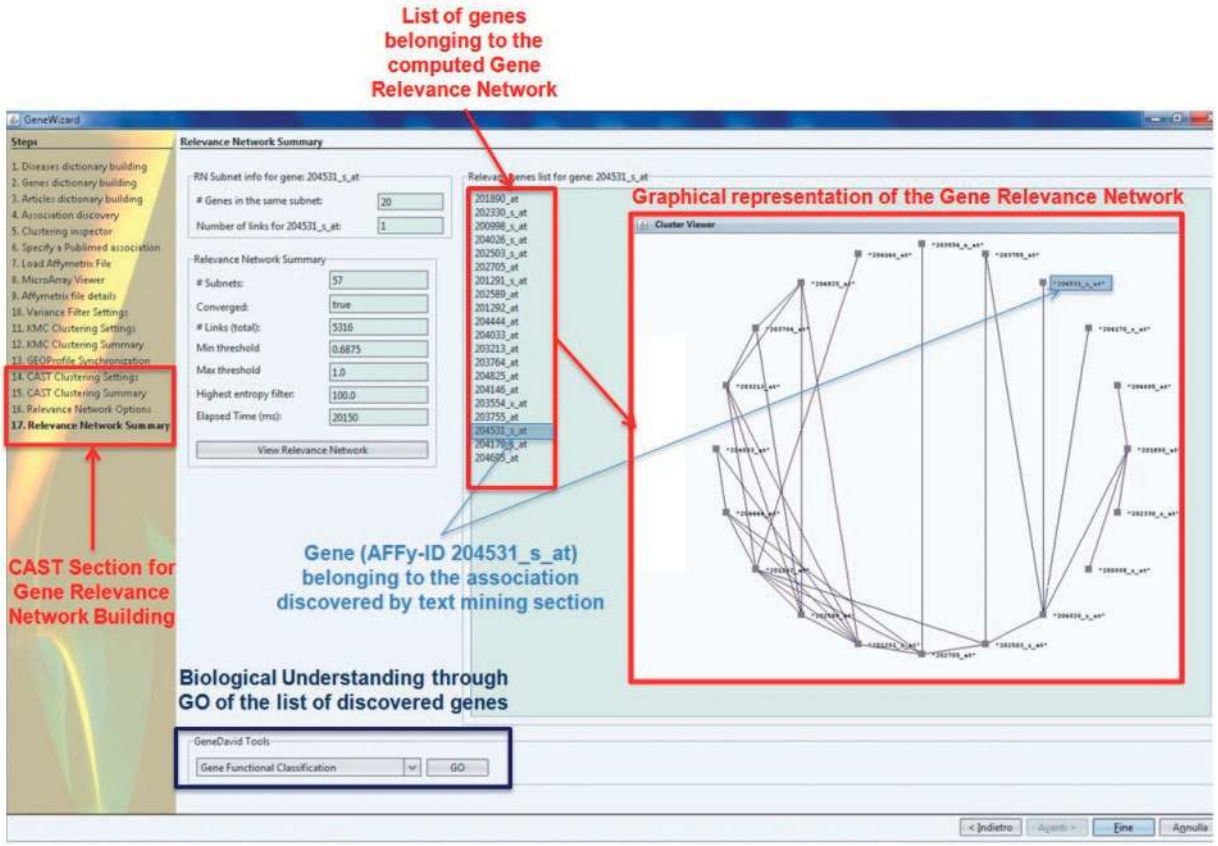

Figure 8: Result generation. Starting from the selected gene set (BRCA1 whose AFFY-ID is 204531_s_at), GeneWizard provides CAST to compute a GRNs involved in the analyzed disease (in this example, Breast Neoplasm). Finally, the list of the genes belonging to the computed GRN is automatically mapped onto GO for biological insights.

algorithm to the selected cluster, only one GRN is obtained. This GRN represents a particular aspect to be investigated, i.e. it indicates from which angle the disease should be analyzed.

The final step is to resort to the GO database to investigate the biological meanings associated to the genes belonging to the identified GRN. Accordingly, GeneWizard allows to define 'rules' to link functionally the transcriptional profile of the discovered list of genes with respect to molecular functions, biological processes and cellular components. These rules assume the form of 'All the genes that share the expression profile with the gene G (of the discovered relationship $G - D$) for the given disease D are related to the molecular function MF, the chemical component CC and the biological process BP'.

Implementation notes

The application has been developed using Sun Wizard API, since the overall model of the analysis is a predefined workflow. Some Java classes have been reused from MEV: i.e. TMEV, Multiple ArrayViewer, IslideData, Experiment, IViewer, AbstractAlgorithm, AlgorithmFactory, Algorithm Data and GEOSeriesMatrixLoader. Other important resources were the Entrez Programming Utilities and their SOAP interface. GeneWizard runs on any operating system (Windows, Linux and Mac OS) provided with a Java Virtual Machine version above of 1.6.0 and it is freely available at the link http://i3s-lab.ing.unict.it/GeneWizard.

CONCLUSIONS

Text mining of the scientific literature has been widely researched and the current availability of tools is satisfying, although there is no evident reason to prefer methods based on co-occurrences (higher recall) or methods based on natural language processing (better precision) when the researcher's goal is knowledge discovery to formulate novel

hypotheses about the relationships of biological entities. However, the full potential of text-mining approaches can be realized only through integration with other data sources, such as ontologies, regulatory information and high-throughput methods outputs. This review mainly focused on the integration between literature text-mining tools/methods and microarray data. In this direction, a major challenge in bioinformatics to support discoveries in biology is to include in the tools functionalities to carry out a contextualized analysis of all the available experimental data (i.e. several microarrays relating to the same relationship) in order to derive biological networks that are compatible with all the available experimental evidence. Another line of development would be the inclusion of functionalities that support explicitly the definition of strategies for exploring the identified relationships according to their probability of being biologically valid. A promising approach to achieve this goal is to modify onset the algorithms that perform the clustering to work not only on the basis of mathematical criteria but also on the basis of what is known on the biological entities that are gradually aggregated, to improve on the overall biological plausibility of the final clusters [116].

We expect to assist in the near future to the development of such approaches, and a desirable contribution from the bioinformatics community would be the development of easy-to-use and freely accessible tools such as GeneWizard. To date, the viability of general-purpose approaches and tools, as compared to domain-specific tools such as CoPub for liver pathologies, that integrate all the information related to the different biological aspects (genes, drugs, pathways, tissues, etc.) of a specific disease, is still unclear.

Still, it seems important that if generalization across domains is sought, it is not achieved at the expenses of tool usability and of a clear representation of the workflow underlying the mining run. In fact, one of the major barriers to the use of such tools is the required technical knowledge about the choice of algorithms, the setting of parameters and the strategies for composing the steps of the mining run, and how all of the above do impact on the obtained results. Clearly, this problem is further complicated when adding the complexity of dealing with several, heterogeneous sources. Thus a well-designed, user-center tool should address the challenges of making clear and explicit the methodological approach supported by the tool, leaving to the users flexibility in exploring the results, and yet providing assistance in managing the complexity of the analysis. This implies that bioinformatics tools should be provided with interactive interfaces for handling annotations, linking across resources or highlighting relevant portions of text, to make the process of data analysis and knowledge discovery more targeted to the users' goals. If these criteria are satisfied, these tools could also provide interesting opportunities to be used as teaching tools and sources to generate compelling teaching case, and be conveniently integrated even in an undergraduate curriculum. This would be in line with current pedagogical models [117] that favor problem-based learning in authentic contexts. The development of such tools will depend on how much close the collaboration between biologists and bioinformaticians will be.

DESCRIPTION OF THE ORGANISATION

The University of Catania, Italy (http://www.unict.it) was founded in 1434. Today more than 55 000 students attend lessons given by over 1500 professors in the 12 faculties, which in turn are staffed by over 1500 administrative employees. The authors are with the Dipartimento di Ingegneria Elettrica, Elettronica ed Informatica (DIEEI) of the Engineering Faculty. The Department aggregates two main ICT areas: computer engineering and telecommunications. Nowadays the Department's ICT research activities are widely differentiated and address subjects such as medical informatics, bioinformatics, multimedia systems, distributed computing, industrial informatics, embedded systems, human-computer interaction, pattern recognition and knowledge management.

Key Points

- The current state of art of text-mining approaches is satisfying and the current trend is to integrate text with multi-type data (biological, chemical, etc.).
- A number of tools supporting the integration of microarray data and literature information have been proposed both to understand the lists of down and upregulated genes and to generate novel biological hypotheses.
- Bioinformatics tools should be intuitive to use and should not require technical knowledge of underlying technology; rather they should assist the user in the process of data integration and results interpretation.
- GeneWizard is an easily usable and freely accessible tool that supports researchers in discovering gene–disease relationships by fusing data resulting from text mining and microarray data analysis.

References

1. Hunter L, Cohen KB. Biomedical language processing: what's beyond PubMed? *Mol Cell* 2006;**21**:589–94.
2. Kell DB. Metabolomics and systems biology: making sense of the soup. *Curr Opin Microbiol* 2004;**7**:296–307.
3. Kell DB. Metabolomics, modelling and machine learning in systems biology - towards an understanding of the languages of cells. *FEBS J* 2006;**273**:873–94.
4. Wheeler DL, Barrett T, Benson DA, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2005;**33**:39–45.
5. Altman RB, Bergman CM, Blake J. Text mining for biology–the way forward: opinions from leading scientists. *Genome Biol* 2008;**9**(Suppl 2):S7.
6. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet* 2006;**7**:119–129.
7. Zweigenbaum P, Demner-Fushman D, Yu H, *et al*. Frontiers of biomedical text mining: current progress. *Brief Bioinformatics* 2007;**8**:358–75.
8. Roberts PM. Mining literature for systems biology. *Brief Bioinformatics* 2006;**7**:399–406.
9. Ananiadou S, Kell D, Tsuj J. Text mining and its potential applications in systems biology. *Trends Biotechnol* 2006;**24**: 571–79.
10. Faro A, Giordano D, Spampinato C, *et al*. Discovering genes-diseases associations from specialized literature using the grid. *IEEE Trans Inf Technol Biomed* 2009;**13**:554–60.
11. Faro A, Giordano D, Spampinato C. Discovery and assessment of gene-disease associations by integrated analysis of scientific literature and microarray data. In: *Proceedings of the 10th Internation Conference on Information Technology and Applications in Biomedicine*, ITAB 2010, Corfu, Greece, November 2-5, 2010.
12. Hearst MA. Untangling text data mining. In: ACL '*99: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999. pp. 3–10. Association for Computational Linguistics, Morristown, NJ, USA.
13. Yang Y, Adelstein SJ, Kassis AI. Target discovery from data mining approaches. *Drug Discov Today* 2009;**14**: 147–54.
14. Kobayashi M, Aono M. Vector space models for search and cluster mining. In: Berry MW, Castellanos M, (eds). *Survey of Text Mining II*. London: Springer, 2008;109–27.
15. Klekota J, Roth FP, Schreiber SL. Query chem: a Google-powered web search combining text and chemical structures. *Bioinformatics* 2006;**22**:1670–3.
16. Rebholz-Schuhmann D, Kirsch H, Arregui M, *et al*. EBIMed-text crunching to gather facts for proteins from Medline. *Bioinformatics* 2007;**23**:e237–44.
17. Doms A, Schroeder M. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res* 2005;**33**: W783–6.
18. Dietze H, Schroeder M. GoWeb: a semantic search engine for the life science web. *BMC Bioinformatics* 2009;**10**(Suppl. 10):S7.
19. Muller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol* 2004;**2**:e309.
20. Perez-Iratxeta C, Bork P, Andrade MA. XplorMed: a tool for exploring MEDLINE abstracts. *Trends Biochem Sci* 2001; **26**:573–5.
21. Hoffmann R, Valencia A. iHOP, a new gene and protein analysis tool. *Cancer Biol Ther* 2007;**6**:7–8.
22. Ananiadou S, Freidman C, Tsujii J. Introduction: named entity recognition in biomedicine. *J Biomed Inform* 2004; **37**(6):393–5.
23. Fukuda K, Tamura A, Tsunoda T, *et al*. Toward information extraction: identifying protein names from biological papers. *Pac Symp Biocomput* 1998;707–18.
24. Narayanaswamy M, Ravikumar KE, Vijay-Shanker K. A biological named entity recognizer. *Pac Symp Biocomput* 2003(1);427–38.
25. Tsuruoka Y, Tsujii J. Improving the performance of dictionary-based approaches in protein name recognition. *J Biomed Inform* 2004;**37**:461–70.
26. Yeganova L, Smith L, Wilbur WJ. Identification of related gene/protein names based on an HMM of name variations. *Comput Biol Chem* 2004;**28**:97–107.
27. Crim J, McDonald R, Pereira F. Automatically annotating documents with normalized gene lists. *BMC Bioinformatics* 2005;**6**(Suppl. 1):S13.
28. Hirschman L, Morgan AA, Yeh AS. Rutabaga by any other name: extracting biological names. *J Biomed Inform* 2002;**35**: 247–59.
29. Fundel K, Guttler D, Zimmer R, *et al*. A simple approach for protein name identification: prospects and limits. *BMC Bioinformatics* 2005;**6**(Suppl. 1):S15.
30. Tsujii J-I, Ananiadou S. Thesaurus or logical ontology, which one do we need for text mining? *Lang Resou Eval* 2005;**39**(1):77–90.
31. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:D267–70.
32. Bamidis P, Kaldoudi E, Pattichis C. From taxonomies to folksonomies: a roadmap from formal to informal modeling of medical concepts and objects. In: *Proceedings of the 9th Internation Conference on Information Technology and Applications in Biomedicine* 2009. Larnaca, Greece, November 5–7, 2009.
33. Tanabe L, Thom LH, Matten W, *et al*. SemCat: semantically categorized entities for genomics. *AMIA Annu Symp Proc* 2006;**2006**:754–8.
34. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:D267–70.
35. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.
36. Maglott D, Ostell J, Pruitt KD, *et al*. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res* 2007;**35**:26–31.
37. Egorov S, Yuryev A, Daraselia N. A simple and practical dictionary-based approach for identification of proteins in Medline abstracts. *J Am Med Inform Assoc* 2004;**11**:174–8.
38. Wexler P. The U.S. National Library of Medicine's Toxicology and Environmental Health Information Program. *Toxicology* 2004;**198**:161–8.
39. Burchfield R. Frequency analysis of English usage: Lexicon and Grammar by W. Nelson Francis and Henry Kucera

with the assistance of Andrew W. Mackie. Boston: Houghton Mifflin. 1982. x + 561. *J Engl Linguist* 1985;**18:**64–70.

40. Tanabe L, Wilbur WJ. A priority model for named entities. In: *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, BioNLP 2006. pp. 33-40. Association for Computational Linguistics, Morristown, NJ, USA.
41. Rajapakse M, Kanagasabai R, Ang WT, *et al*. Ontology-centric integration and navigation of the dengue literature. *J Biomed Inform* 2007;**41**:806–15.
42. Rebholz-Schuhmann D, Arregui M, Gaudan S, *et al*. Text processing through web services: calling Whatizit. *Bioinformatics* 2008;**24**:296–8.
43. Stevenson Guo Y. Disambiguation in the biomedical domain: the role of ambiguity type. *J Biomed Inform* 2010; **46**(6):972–81.
44. Stevenson M, Guo Y. Disambiguation of ambiguous biomedical terms using examples generated from the UMLS metathesaurus. *J Biomed Inform* 2010;**43**(5):762–73.
45. Tsuruoka Y, Tsujii J, Ananiadou S. Facta: a text search engine for finding associated biomedical concepts. *Bioinformatics* 2008;**24**:2559–60.
46. Mubaid HA, Singh RK. A text mining technique for extracting gene disease associations from the biomedical literature. *Int J Bioinform Res Appl* 2010;**6**(3):270–86.
47. Mukhopadhyay S, Palakal M, Maddu K. Multi-way association extraction and visualization from biological text documents using hyper-graphs: applications to genetic association studies for diseases. *Artif Intell Med* 2010;**49**: 145–54.
48. Kabiljo R, Clegg AB, Shepherd AJ. A realistic assessment of methods for extracting gene/protein interactions from free text. *BMC Bioinformatics* 2009;**10**:233.
49. Katukuri JR, Xie Y, Raghavan VV. Biomedical relationship extraction from literature based on bio-semantic token subsequences. In: *Proceedings of the 2009 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2009*. pp. 366–70. IEEE Computer Society, Washington, DC, USA.
50. Masseroli M, Kilicoglu H, Lang FM, *et al*. Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease. *BMC Bioinformatics* 2006;**7**:291.
51. Barnickel T, Weston J, Collobert R, *et al*. Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *PLoS ONE* 2009;**4**:e6393.
52. Miyao Y, Sagae K, Saetre R, *et al*. Evaluating contributions of natural language parsers to protein-protein interaction extraction. *Bioinformatics* 2009;**25**:394–400.
53. Fundel K, Kuffner R, Zimmer R. RelEx-relation extraction using dependency parse trees. *Bioinformatics* 2007;**23**: 365–71.
54. Hanisch D, Fundel K, Mevissen HT, *et al*. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics* 2005;**6**(Suppl. 1):S14.
55. Miyao Y, Sagae K, Saetre R, *et al*. Evaluating contributions of natural language parsers to protein-protein interaction extraction. *Bioinformatics* 2009;**25**:394–400.
56. Rinaldi F, Schneider G, Kaljurand K, *et al*. Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. *Artif Intell Med* 2007;**39**: 127–36.
57. Ohta T, Matsuzaki T, Okazaki N, *et al*. Medie and Info-Pubmed: 2010 update. *BMC Bioinformatics* 2010; **11**(Suppl. 5):P7.
58. Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC Bioinformatics* 2009; **10**(Suppl. 2):S6.
59. Kim JJ, Zhang Z, Park JC, *et al*. BioContrasts: extracting and exploiting protein-protein contrastive relations from biomedical literature. *Bioinformatics* 2006;**22**:597–605.
60. Dai HJ, Chang Y-C, Tsai RT-H, *et al*. New challenges for biological text-mining in the next decade. *J Comput Sci Technol* 2010;**25**(1):169–79.
61. Coelho P, Ahmed A, Arnold A, *et al*. Structured literature image finder: extracting information from text and images in biomedical literature. *Lect Notes Comput Sci* 2010;**6004**: 23–32.
62. Swanson DR. Medical literature as a potential source of new knowledge. *Bull Med Libr Assoc* 1990;**78**:29–37.
63. Smalheiser NR, Torvik VI, Zhou W. Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *Comput Methods Programs Biomed* 2009;**94**:190–7.
64. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med* 1986;**30**: 7–18.
65. Swanson DR. Migraine and magnesium: eleven neglected connections. *Perspect Biol Med* 1988;**31**:526–57.
66. Swanson DR. Intervening in the life cycles of scientific knowledge Patrick Wilson, the value of currency. *Library Trends* 1993;**41**(4):606–31.
67. Swanson DR, Smalheiser N. Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. *Neuro-Sci Res Commun* 1994;**15**(4):1–9.
68. Saric J, Jensen LJ, Ouzounova R, *et al*. Extraction of regulatory gene/protein networks from Medline. *Bioinformatics* 2006;**22**:645–50.
69. Yang Z, Lin H, Li Y. BioPPISVMExtractor: a protein-protein interaction extractor for biomedical literature using SVM and rich feature sets. *J Biomed Inform* 2010;**43**:88–96.
70. Narayanaswamy M, Ravikumar KE, Vijay-Shanker K. Beyond the clause: extraction of phosphorylation information from medline abstracts. *Bioinformatics* 2005;**21**(Suppl. 1): i319–27.
71. Ozgur A, Xiang Z, Radev DR, *et al*. Literature-based discovery of IFN-gamma and vaccine-mediated gene interaction networks. *J Biomed Biotechnol* 2010;**2010**: 426479.
72. Bandy J, Milward D, McQuay S. Mining protein-protein interactions from published literature using Linguamatics I2E. *Method Mol Biol* 2009;**563**:3–13.
73. Jelier R, Jenster G, Dorssers LC, *et al*. Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation. *BMC Bioinformatics* 2007;**8**:14.
74. Rzhetsky A, Iossifov I, Koike T, *et al*. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J Biomed Inform* 2004; **37**:43–53.

75. Demaine J, Martin J, Wei L, *et al*. Litminer: integration of library services within a bio-informatics application. *Biomed Digit Libr* 2006;**3**(1):11.
76. Jelier R, Jenster G, Dorssers LC, *et al*. Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics* 2005;**21**: 2049–58.
77. Jelier R, Schuemie MJ, Veldhoven A, *et al*. Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biol* 2008;**9**:R96.
78. Cheng D, Knox C, Young N, *et al*. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res* 2008;**36**:399–405.
79. Wishart DS, Knox C, Guo A, *et al*. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 2008;**36**:D901–6.
80. Stenson PD, Mort M, Ball EV, *et al*. The human gene mutation database: 2008 update. *Genome Med* 2009;**1**(1):1–6.
81. Nakazato T, Bono H, Matsuda H, *et al*. Gendoo: functional profiling of gene and disease features using MeSH vocabulary. *Nucleic Acids Res* 2009;**37**:W166–9.
82. Werner T. Bioinformatics applications for pathway analysis of microarray data. *Curr Opin Biotechnol* 2008;**19**:50–4.
83. Ochs MF, Peterson AJ, Kossenkov A, *et al*. Incorporation of gene ontology annotations to enhance microarray data analysis. *Methods Mol Biol* 2007;**377**:243–54.
84. Draghici S, Khatri P, Martins RP, *et al*. Global functional profiling of gene expression. *Genomics* 2003;**81**:98–104.
85. Salomonis N, Hanspers K, Zambon AC, *et al*. GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics* 2007;**8**:217.
86. Ekins S, Nikolsky Y, Bugrim A, *et al*. Pathway mapping tools for analysis of high content data. *Method Mol Biol* 2007;**356**:319–50.
87. Goffard N, Weiller G. PathExpress: a web-based tool to identify relevant pathways in gene expression data. *Nucleic Acids Res* 2007;**35**:W176–81.
88. Wu J, Mao X, Cai T, *et al*. KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res* 2006;**34**:W720–4.
89. Pilpel Y, Sudarsanam P, Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet* 2001;**29**:153–9.
90. Sudarsanam P, Pilpel Y, Church GM. Genome-wide co-occurrence of promoter elements reveals a cis-egulatory cassette of rRNA transcription motifs in Saccharomyces cerevisiae. *Genome Res* 2002;**12**:1723–31.
91. Margolin AA, Nemenman I, Basso K, *et al*. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 2006;**7**(Suppl. 1):S7.
92. Lee PH, Lee D. Modularized learning of genetic interaction networks from biological annotations and mRNA expression data. *Bioinformatics* 2005;**21**:2739–47.
93. Lee D, Kim S, Kim Y. BioCAD: an information fusion platform for bio-network inference and analysis. *BMC Bioinformatics* 2008;**8**(Suppl. 9):S2.
94. Jensen LJ, Kuhn M, Stark M, *et al*. String 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* 2009;**37**:D412–6.
95. Jelier R, Jenster G, Dorssers LC, *et al*. Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation. *BMC Bioinformatics* 2007;**8**:14.
96. Kuffner R, Fundel K, Zimmer R. Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts. *Bioinformatics* 2005;**21**(Suppl. 2):i259–67.
97. Nam D, Kim SY. Gene-set approach for expression pattern analysis. *Brief Bioinformatics* 2008;**9**:189–97.
98. Park I, Lee KH, Lee D. Inference of combinatorial Boolean rules of synergistic gene sets from cancer microarray datasets. *Bioinformatics* 2010;**26**:1506–12.
99. Huang ZX, Tian HY, Hu ZF, *et al*. GenCLiP: a software program for clustering gene lists by literature profiling and constructing gene co-occurrence networks related to custom keywords. *BMC Bioinformatics* 2008;**9**:308.
100. Chagoyen M, Carmona-Saez P, Shatkay H, *et al*. Discovering semantic features in the literature: a foundation for building functional associations. *BMC Bioinformatics* 2006;**7**(1):41.
101. Lee DD, Seung HS. Algorithms for non-negative matrix factorization. *NIPS* 2000;**1**:556–62.
102. Frijters R, Heupers B, van Beek P, *et al*. CoPub: a literature-based keyword enrichment tool for microarray data analysis. *Nucleic Acids Res* 2008;**36**:W406–10.
103. Jelier R, Goeman JJ, Hettne K, *et al*. Literature-aided interpretation of gene expression data with the weighted global test. *Brief Bioinform* 2010. doi:10.1093/bib/bbq082.
104. Perez-Iratxeta C, Wjst M, Bork P, *et al*. G2D: a tool for mining genes associated with disease. *BMC Genet* 2005;**6**: 45.
105. Perez-Iratxeta C, Bork P, Andrade-Navarro MA. Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Res* 2007;**35**:W212–6.
106. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequence project: update and current status. *Nucleic Acids Res* 2003;**31**:34–7.
107. Tremblay K, Lemire M, Potvin C, *et al*. Genes to diseases (G2D) computational method to identify asthma candidate genes. *PLoS ONE* 2008;**3**:e2907.
108. Tiffin N, Kelso JF, Powell AR, *et al*. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res* 2005;**33**:1544–52.
109. Fernandez-Suarez XM, Schuster MK. Using the ensembl genome server to browse genomic sequence data. *Curr Protoc Bioinformatics* 2010. **Chapter 1**:Unit 1.15.
110. Kelso J, Visagie J, Theiler G, *et al*. eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res* 2003;**13**:1222–30.
111. Tranchevent L-C, Barriot R, Yu S, *et al*. Endeavour update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res* 2008;**36**(Suppl. 2):W377–84.
112. Yetisgen-Yildiz M, Pratt W. A new evaluation methodology for literature-based discovery systems. *J Biomed Inform* 2009;**42**:633–43.
113. Kostoff RN. Validating discovery in literature-based discovery. *J Biomed Inform* 2007;**40**:448–50.
114. Saeed AI, Sharov V, White J, *et al*. TM4: a free, open-source system for microarray data management and analysis. *BioTechniques* 2003;**34**:374–8.

115. Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. *J Comput Biol* 1999;**6**(3/4):281–97.

116. Tan M, Smith E, Broach J, *et al*. Microarray data mining: a novel optimization-based approach to uncover biologically coherent structures. *BMC Bioinformatics* 2008;**9**(1): 268–89.

117. Bamidis PD, Konstantinidis ST, Kaldoudi E, *et al*. New approaches in teaching medical informatics to medical students. In: *Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems, Vol. 1, June 17–19 2008,* pp. 385–390. Jyväskylä, Finland.