Bioinformatics Advance Access published August 26, 2014

BIOINFORMATICS **ORIGINAL PAPER**

2014, pages 1–8
doi:10.1093/bioinformatics/btu524

Data and text mining

Advance Access publication August 6, 2014

BioTextQuest^{+}: a knowledge integration platform for literature mining and concept discovery

Nikolas Papanikolaou[1,†], Georgios A. Pavlopoulos[1,†], Evangelos Pafilis[2], Theodosios Theodosiou[1], Reinhard Schneider[3], Venkata P. Satagopam[3], Christos A. Ouzounis[4,5], Aristides G. Eliopoulos[1,6], Vasilis J. Promponas[7] and Ioannis Iliopoulos[1,*]

[1]Division of Basic Sciences, University of Crete, Medical School, Heraklion 71110, Greece, [2]Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Hellenic Centre for Marine Research (HCMR), Heraklion, Greece, [3]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Campus Belval, 7, avenue des Hauts-Fourneaux, L-4362 Esch sur Alzette, Luxembourg, [4]Biological Computation & Process Laboratory (BCPL), Chemical Process & Energy Resources Institute (CPERI), Centre for Research & Technology Hellas (CERTH), PO Box 361, GR-57001 Thessalonica, Greece, [5]Donnelly Centre for Cellular & Biomolecular Research, University of Toronto, Toronto, Ontario, Canada, [6]Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology Hellas, 70013 Heraklion, Crete, Greece and [7]Department of Biological Sciences, Bioinformatics Research Laboratory, University of Cyprus, PO Box 20537, CY 1678, Nicosia, Cyprus

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: The iterative process of finding relevant information in biomedical literature and performing bioinformatics analyses might result in an endless loop for an inexperienced user, considering the exponential growth of scientific corpora and the plethora of tools designed to mine PubMed® and related biological databases. Herein, we describe BioTextQuest^{+}, a web-based interactive knowledge exploration platform with significant advances to its predecessor (BioTextQuest), aiming to bridge processes such as bioentity recognition, functional annotation, document clustering and data integration towards literature mining and concept discovery. BioTextQuest^{+} enables PubMed and OMIM querying, retrieval of abstracts related to a targeted request and optimal detection of genes, proteins, molecular functions, pathways and biological processes within the retrieved documents. The front-end interface facilitates the browsing of document clustering per subject, the analysis of term co-occurrence, the generation of tag clouds containing highly represented terms per cluster and at-a-glance popup windows with information about relevant genes and proteins. Moreover, to support experimental research, BioTextQuest^{+} addresses integration of its primary functionality with biological repositories and software tools able to deliver further bioinformatics services. The Google-like interface extends beyond simple use by offering a range of advanced parameterization for expert users. We demonstrate the functionality of BioTextQuest^{+} through several exemplary research scenarios including author disambiguation, functional term enrichment, knowledge acquisition and concept discovery linking major human diseases, such as obesity and ageing.

Availability: The service is accessible at http://bioinformatics.med.uoc.gr/biotextquest.

Contact: g.pavlopoulos@gmail.com or georgios.pavlopoulos@esat.kuleuven.be

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 20, 2013; revised on June 24, 2014; accepted on July 29, 2014

1 INTRODUCTION

The tremendous growth of biomedical literature and biological data repositories has become a true challenge for researchers who wish to follow in detail the developments in rapidly growing fields (Altman *et al*., 2008). Currently, PubMed contains >22 million publications and information about biological entities, such as genes, proteins, pathways and sequences, which is stored and distributed by several repositories worldwide. Therefore, targeted information retrieval from literature and other data collections often becomes a complicated and tedious task when searching with available engines (Lu, 2011) as it regularly results to hundreds or thousands of records, typically not sorted by relevance and often changing on parameterization or query rephrasing.

Despite the plethora of proposed tools aiming to make document searching easier, more efficient and targeted, literature mining is still an open area of research (Rodriguez-Esteban, 2009) Tools designed to tackle this problem can be classified in five categories: (i) *tools for ranking search results*, (ii) *tools for document clustering in several topics*, (iii) *tools for semantic annotation and display*, (iv) *tools for predicting relations between bioentities and/or concepts* and (v) *tools extending PubMed functionality by improving the search interface*. Notable tools for result ranking include Quertle (Giglia, 2011), MedlineRanker (Fontaine *et al*., 2009), Caipirini (Soldatos *et al*., 2012),

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

MiSearch (States *et al*., 2009), MScanner (Poulter *et al*., 2008), eTBLAST (Errami *et al*., 2007) and PubFocus (Plikus *et al*., 2006). Typically, they use a combination of classifiers, such as support vector machines or weight schemes, complemented by artificial intelligence algorithms to rank documents based on MeSH® terms, author names or structured vocabularies. Document clustering tools, such as Anne O'Tate (Smalheiser *et al*., 2008), McSyBi (Yamamoto and Takagi, 2007), PuRed-MCL (Theodosiou *et al*., 2008), GoPubMed (Doms and Schroeder, 2005) and XPlorMed (Perez-Iratxeta *et al*., 2002), aim to group documents into subjects for easier management of large unordered lists of results. Clustering is mostly performed by the use of MeSH terms, UMLS dictionaries, GO terms, titles, affiliations, keywords, authors, standard vocabularies, extracted terms or any combination of the aforementioned, including semantic annotation. Other tools, e.g. OnTheFly (Pafilis *et al*., 2013; Pavlopoulos *et al*., 2009) for full-text or MedEvi (Kim *et al*., 2008), EBIMed (Rebholz-Schuhmann *et al*., 2007), BioLit (Fink *et al*., 2008), BSQA (He *et al*., 2010) and PubNet (Douglas *et al*., 2005), for abstract-based analysis, extract relations between bioentities, such as proteins, chemicals, drugs, GO terms, and deliver them as visual network representations. Additionally, tools such as FACTA$^+$/FACTA (Tsuruoka *et al*., 2008, 2011) or PolySearch (Cheng *et al*., 2008) are designed to help users browsing and linking biomedical concepts. Finally, tools like HubMed (Eaton, 2006), askMEDLINE (Fontelo *et al*., 2005), PubCrawler (Hokamp and Wolfe, 2004) and iPubMed (Wang *et al*., 2010) take advantage of the newest web technologies to provide enhanced interfaces with various search options, filtering and parameterization, currently not available in PubMed.

While most of the aforementioned tools serve complementary purposes and are PubMed-based, only a few of them manage to bridge the gap between information retrieval and analysis. The iterative process, (i) *compose the most appropriate query*, (ii) *query several databases*, (iii) *retrieve information*, (iv) *filter down the most relevant results*, (v) *retrieve further information about the bioentities of interest from other repositories*, and (vi) *analyse the data*, can become tedious, demanding and laborious. Therefore, the development of efficient tools to automate and bridge such procedures is of major importance. BioTextQuest$^+$ provides a workflow to query PubMed and OMIM databases (Hamosh *et al*., 2005) and feed an automatic pipeline for document preprocessing, clustering, visualization and data integration with other repositories. It comes with an intuitive web-based graphical user interface and supports name entity recognition (NER) in abstracts and depiction of associations among them within the inferred clusters. In addition, it offers an array of visualization tools for efficient navigation among biomedical records and concept extraction/association. Through in-house services, BioTextQuest$^+$ provides links and further information from public repositories about the identified genes/proteins in one or more clusters. Such information includes homology, transcription factor analysis, protein clustering by domain content, generation of protein–protein/protein–chemical interactions and functional enrichment. BioTextQuest$^+$ is a complementary tool to PubMed and offers intuitive solutions for researchers who can navigate from a single query to concept discovery, knowledge management and bioinformatics analysis in a simple, controlled and automated way, thus making BioTextQuest$^+$ a powerful tool for the scientific community.

2 METHODS

2.1 System overview

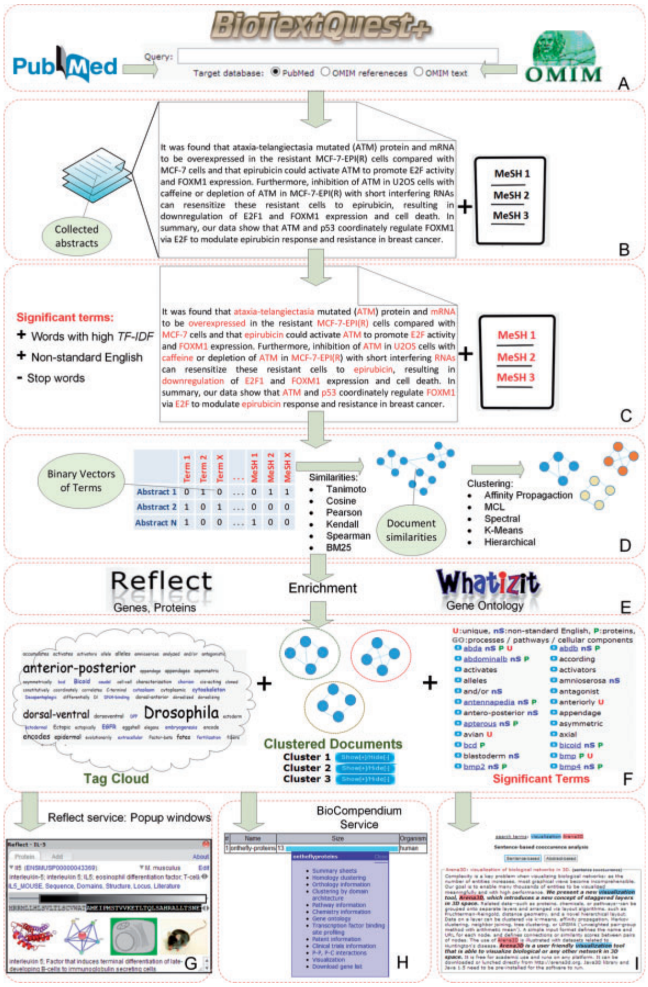
BioTextQuest$^+$ is an easy-to-use web application aiming to support literature management, knowledge discovery by concept association and data integration, aimed at both computational and experimental researchers. It is designed to maximize user experience by offering a simple Google-like web interface and graphical visualization of associations among biological terms. The core component of BioTextQuest$^+$ is based on a number of former implementations (Iliopoulos *et al*., 2001; Papanikolaou *et al*., 2011) aiming to extract significant biomedical terms from an abstract collection and subsequently cluster these abstracts into subjects according to their similarity based on the extracted terms. Currently, BioTextQuest$^+$ significantly extends its inherited functionality by supporting queries to both PubMed and OMIM repositories, and facilitates advanced text annotation services, employed to map the extracted terms of biomedical significance to the corresponding bioentities. The reported document clusters and their associated terms, along with the retrieved annotation, are combined to generate a series of views of results. In addition, BioTextQuest$^+$ incorporates web services to automatically enable in-depth bioinformatics analysis for the gene and protein set of each resulting cluster. Importantly, users can interactively sub-cluster and/or re-cluster the results of any completed analysis. The functionality of BioTextQuest$^+$ is summarized in Figure 1 and described in the Supplementary Materials and Methods file.

2.2 Query system

BioTextQuest$^+$ currently enables both PubMed and OMIM querying (Supplementary Figure S1). Both databases are locally stored and a daemon is implemented to monitor new MEDLINE® and OMIM releases to maintain them up to date weekly. The query field allows input of any valid PubMed and OMIM query supporting all features offered by their search mechanisms, such as field-tags, Boolean operators or grouping parentheses. In the case of PubMed, BioTextQuest$^+$ uses the Entrez utilities to post a query directly to PubMed and get back the PubMed identifiers of matching entries. In a subsequent step and depending on user-defined parameters, the platform uses these identifiers to retrieve the appropriate combination of abstracts and their associated MeSH terms from the local database (Supplementary Figure S1, Figure 1B), thus maximizing the speed of information retrieval. In the case of OMIM, BioTextQuest$^+$ follows a similar strategy. It initially allows the user to query the OMIM database and get back the corresponding identifiers to the query. The following step is to either query the local OMIM database for the OMIM text related to these identifiers or collect the PubMed identifiers for each OMIM record and subsequently query the local PubMed for their abstracts and MeSH terms (Figure 1B, Supplementary Figure S1). While PubMed retrieval is performed in a few seconds, OMIM full-text analysis may take a few minutes, as parsing the original text of the retrieved records is required. BioTextQuest$^+$ users may also specify the number of documents to be retrieved/processed (with a maximum of 5000 articles/entries as default). In cases of queries returning >5000 documents, the 5000 most recent are retained for analysis.

2.3 Document similarities

We represent each abstract with a binary vector indicating the presence or absence of the biologically significant terms found in the text collection (Figure 1D). Similarity metrics available in BioTextQuest$^+$ are the *Cosine similarity*, *Tanimoto coefficient*, *Pearson correlation coefficient*, *Spearman correlation* and *Kendall's Tau rank correlation coefficient*. Additionally,



It was found that ataxia-telangiectasia mutated (ATM) protein and mRNA to be overexpressed in the resistant MCF-7-EPI(R) cells compared with MCF-7 cells and that epirubicin could activate ATM to promote E2F activity and FOXM1 expression. Furthermore, inhibition of ATM in U2OS cells with caffeine or depletion of ATM in MCF-7-EPI(R) with short interfering RNAs can resensitize these resistant cells to epirubicin, resulting in downregulation of E2F1 and FOXM1 expression and cell death. In summary, our data show that ATM and p53 coordinately regulate FOXM1 via E2F to modulate epirubicin response and resistance in breast cancer.

It was found that ataxia-telangiectasia mutated (ATM) protein and mRNA to be overexpressed in the resistant MCF-7-EPI(R) cells compared with MCF-7 cells and that epirubicin could activate ATM to promote E2F activity and FOXM1 expression. Furthermore, inhibition of ATM in U2OS cells with caffeine or depletion of ATM in MCF-7-EPI(R) with short interfering RNAs can resensitize these resistant cells to epirubicin, resulting in downregulation of E2F1 and FOXM1 expression and cell death. In summary, our data show that ATM and p53 coordinately regulate FOXM1 via E2F to modulate epirubicin response and resistance in breast cancer.

	Term 1	Term 2	Term X	…	MeSH 1	MeSH 2	MeSH X
Abstract 1	0	1	0	…	0	1	1
Abstract 2	1	0	1	…	0	0	0
Abstract N	1	0	0	…	1	0	0

Fig. 1. Outline of the general functionality of BioTextQuest+. (**A**) Queries to PubMed and OMIM by keeping the original syntax. (**B**) Retrieval of abstracts and their MeSH terms related to the query. (**C**) Identification of significant terms by excluding stop-words, standard English words and terms with low TF-IDF. (**D**) Generation of binary vectors for each abstract showing the presence or the absence of significant terms and generation of similarity matrices for document clustering. (**E**) Enrichment of presented clusters for genes, proteins, pathways and GO terms. (**F**) Tag cloud example of highly representative terms for each document cluster (left). Document clustering and document categorization in subjects (middle). Annotated list of significant terms used to cluster documents (right). (**G**) Informative popup windows for genes/proteins and chemicals. (**H**) Functional annotation, bioinformatics analysis and data integration with a great variety of repositories by using the in-house BioCompendium service. (**I**) Sentence-based and abstract-based co-occurrence analysis

we use the *Okapi BM25* similarity, which is widely used by search engines to rank matching documents according to their relevance to a query (Sparck *et al.*, 2000) and is mostly preferred for large document sets.

2.4 Document clustering

Inspired by jClust (Pavlopoulos *et al.*, 2009) and NeAT (Brohee *et al.*, 2008) applications along with ClusterMaker (Morris *et al.*, 2011), we incorporated a plethora of clustering algorithms in BioTextQuest+ to cluster documents based on their significant terms. We have currently integrated the Affinity Propagation (Frey and Dueck, 2007), MCL (Enright *et al.*, 2002), *k*-Means (MacQueen, 1967), average linkage hierarchical clustering from SCPS (Nepusz *et al.*, 2010) and spectral (Paccanaro *et al.*, 2006) clustering algorithms. Notably, the aforementioned methods take as input the similarity matrices as defined in the previous section and their results are visualized in various ways as shown in Figure 1F. Comparisons between these algorithms are presented elsewhere (Boyack *et al.*, 2011; Moschopoulos *et al.*, 2011). In order not to overwhelm non-expert users, we chose to hide by default the relevant options from the main interface; MCL is selected as the default clustering algorithm with inflation value 1.8 and the cosine similarity metric. Some may consider that the plethora of the integrated algorithms can become confusing; however, it is a useful feature for clustering experts, as each algorithm carries a different philosophy on how to cluster data (Pavlopoulos *et al.*, 2009). An important feature of BioTextQuest+ is that it hosts two classes of clustering algorithms, depending on whether the number of resulting clusters is required as input. For example, MCL automatically detects the number of clusters formed by the data (depending on the choice of inflation parameter), whereas *k*-Means requires that the number of clusters *k* is known beforehand. Empirically, the MCL and *k*-Means clustering algorithms are among the fastest with MCL being able to handle bigger datasets due to its stochastic nature and its efficient memory handling.

2.5 Representation of results

The resulting page provides different views of the results organized under tabs, along with a frame holding a summary of the analysis. The '*Tag Clouds*' view displays a tag-cloud, computed for the terms characterizing each document cluster (Figure 1F, left). This type of familiar representation can be particularly informative, as the font size of each term is proportional to the fraction of documents with the specific term within a cluster. Additionally, in this view, users may highlight (i) terms appearing exclusively in any particular cluster (*unique*), (ii) terms that are not standard grammatical terms (i.e. they do not belong in the reference dictionary used by BioTextQuest+), (iii) terms describing protein names, detected and annotated using the Reflect web service and (iv) Gene Ontology-related terms (i.e. *processes/pathways/cellular components*) identified via the Whatizit web service. When users have opted for applying a stemming algorithm, terms are displayed in their stemmed form. However, the respective unstemmed terms are revealed for inspection by clicking on the tag-cloud. The '*Clustered Documents*' view (Figure 1F, middle) categorizes the documents in subjects corresponding to implicit concepts and provides the titles of the documents that belong to each cluster with a link to the respective article in PubMed. Finally, the '*Biomedical Terms*' view (Figure 1F, right) lists the significant terms identified by the BioTextQuest+ algorithm and subsequently used for the clustering procedure. Each term is graphically annotated and colored in the same manner as in the 'Tag Clouds' view. Users may directly interact with the results by requesting a new refined analysis of the corpus retrieved by the current query. This is achieved by any combination of (i) altering the available tunable options (e.g. stemming or clustering algorithms and their parameters), (ii) excluding terms from the term list available in the 'Biomedical Terms' view and (iii) removing clusters of documents from the 'Tag Clouds' view.

2.6 Co-occurrence analysis

Term co-occurrence analysis is offered when users select up to five terms of interest (mouse click). BioTextQuest+ subsequently retrieves the set of documents where the selected terms are co-mentioned. Users can switch between an abstract-based and sentence-based co-occurrence analysis, thus revealing associations between bioentities through their coexistence in single sentences or entire abstracts (Pavlopoulos *et al.*, 2014). We currently use simple rules to segment the sentences using periods as sentence delimiters, i.e. as indicators of sentence termination in English. When two or more terms are found between two successive delimiters we consider that they co-occurred in the same sentence. This functionality is presented in Figure 1I.

2.7 Gene, protein and chemical information at a glance

Often, resulting pages contain significant terms, such as genes and proteins appearing as names, symbols, database identifiers or free text. To automate and maximize information extraction, we highlight these terms and provide at-a-glance information about their functionality by enriching BioTextQuest's$^+$ resulting pages with the Reflect annotation service (Pafilis *et al*., 2009) through JavaScript feeds. In this manner, a researcher can obtain information about a protein in a popup window on-the-fly with a single mouse click. This popup window compacts information about the protein sequence, its 3D structure, related protein interactors, hosting organisms, subcellular localization, functional domains and several identifiers and synonyms. An example of such a popup window is presented in Figure 1G.

2.8 Data integration and functional enrichment

To automate functional enrichment and information retrieval from public repositories, we have merged BioTextQuest$^+$ with the in-house high-throughput experimental data analysis platform, BioCompendium (http://biocompendium.embl.de) (Figure 1H). BioCompendium is currently a publicly accessible high-throughput experimental data analysis platform and is designed to work with large lists of genes or proteins for which it collects a wide spectrum of biological information from public repositories. It facilitates the analysis, comparison and enrichment of experimental results; either proprietary or publicly available datasets are supported. Notably, the current version is designed to work for human, mouse and yeast. The main features of the BioCompendium system are summarized below: (i) *Comprehensive knowledge collection from different biological databases for a given list(s) of genes*. (ii) *Search interface to the knowledge collection to find information like gene annotations, disease associations, sequence domain architectures, related chemicals and involved pathways*. (iii) *Enrichment analysis for Gene Ontology terms, diseases, pathways and other biological concepts*. (iv) *Extraction of protein–protein (PPI), protein–chemical interactions networks*. (v) *Protein clustering data based on sequence similarity-based homology and sequence domain architectures in a given list(s) of genes*. (vi) *Automated analysis and clustering of transcription factor binding site profiles*. (vii) *Gene linking to orthology information, clinical trial and patent information*. (viii) *Deep comparison of results derived from different experimental conditions, time series or treatments*. By using in-house web services, BioTextQuest$^+$ is able to automatically feed the BioCompendium platform with the relevant identifiers for proteins mentioned in a set of selected clusters. The protein names are first translated to ENSEMBL identifiers, and a new analysis with all the features of BioCompendium is offered in a new tab. Genes belonging to a certain document cluster are mapped to ENSEMBL identifiers through Reflect and are fed into BioCompendium. Gene annotations are collected through ENSEMBL (Flicek *et al*., 2012), EMBL (Cochrane *et al*., 2009), GenBank (Benson *et al*., 2011), EntrezGene (Maglott *et al*., 2011), UniGene (Schuler, 1997), UniProt (Magrane and Consortium, 2011), IPI (Kersey *et al*., 2004), NCBI Protein (Sayers *et al*., 2012), RefSeq (Sayers *et al*., 2012), HGNC (Seal *et al*., 2011), GeneCards (Safran *et al*., 2010) and UCSC (Fujita *et al*., 2011) databases. Sequence domain architectures, structures and annotations are collected from PDB (Berman *et al*., 2007), HSSP (Schneider and Sander, 1996) and PSSH (Schafferhans *et al*., 2003). Functional enrichment is performed by querying the Gene Ontology repository (Ashburner *et al*., 2000) to collect all of the related biological processes, molecular functions and cellular components. Pathways containing any of the submitted genes are retrieved from KEGG (Kanehisa *et al*., 2012), and genes that belong to any cluster are highlighted within the target pathway. Biological networks (Pavlopoulos *et al*., 2011) such as PPIs are obtained from the STRING (Szklarczyk *et al*., 2011) database and protein–chemical interaction networks are gathered from STITCH (Kuhn *et al*., 2008). Relations with other chemicals are additionally retrieved from DrugBank (Wishart *et al*., 2008), HMDB (Wishart *et al*., 2009), PubChem (Bolton *et al*., 2008), chEBI (Degtyarenko *et al*., 2008), MATADOR (Gunther *et al*., 2008) and PDBLigand (Feng *et al*., 2004). Analysis and clustering of transcription factor binding site profiles is performed with the use of JASPAR (Bryne *et al*., 2008), and access to orthology information and clinical trials is given by the ENSEMBL (Vilella *et al*., 2009) and ClinicalTrials.gov (Zarin *et al*., 2011) resources, respectively. Finally, patent information is collected from the EPO Proteins (www.epo.org), JPO Proteins (www.jpo.go.jp), USPTO Proteins (www.uspto.gov) and KIPO Proteins (www.kipo.go.kr/en/) repositories.

3 RESULTS

3.1 Author disambiguation and use of MeSH terms

Searching for an author name in PubMed can often be confusing and misleading and might result in long unordered lists when there is a substantial overlap between the names of different authors working in different fields. To demonstrate the usefulness of the tool in author disambiguation, we query PubMed for '*Pavlopoulos G[AU]*' author. '*Pavlopoulos GA*' (a co-author of this article) works in the areas of bioinformatics and computational biology with emphasis on biological data analysis and visualization, whereas '*Pavlopoulos G*' is an ophthalmologist. Despite the name overlap and the apparent concept similarity between 'data visualization' and 'human vision/ophthalmology', BioTextQuest$^+$ is able to distinguish the author names with maximum accuracy. We have chosen the following parameters: cosine similarity, Markov clustering and, importantly, clustering based on significant terms without the use of MeSH terms for a less directed and unsupervised clustering. BioTextQuest$^+$ successfully detected the two authors in two distinct clusters for '*Pavlopoulos GA*' (17 articles) and '*Pavlopoulos G*' (7 articles). Over-represented significant terms in the tag cloud for the former author include the words: '*bioinformatics*', '*datasets*', '*visualization*', '*large-scale*', '*high-throughput*' and '*genomics*'. Similarly, over-represented significant terms in the second tag cloud for the latter author include the words: '*diopter*', '*keratometry*', '*corneal*', '*intraocular*' and others. We repeated the same analysis by additionally using MeSH terms for a more guided clustering and by triggering the same clustering parameters. Again, BioTextQuest$^+$ produces two distinct clusters for the two authors. New and more specialized over-represented terms give a clearer overview about the author's profile. Such terms include '*information retrieval*', '*cluster analysis*', '*user interface*', '*databases*' and '*computational biology*'. Similarly, new specialized words such as '*diagnosis*', '*laser therapy*', '*eye infection*', '*contact lenses*' and '*conjunctivitis*' clearly refer to the field of ophthalmology, thus providing an enriched picture about the second author's profile and background. This example clearly demonstrates the challenge of author name disambiguation in possibly semantically overlapping areas of activity and the drill through steps in term collections. Moreover, this example illustrates that the use of MeSH terms may highlight subtle concepts characterizing a specific corpus.

3.2 Functional enrichment

To demonstrate the bridging between text-mining and data integration and simultaneously benchmark the accuracy of the BioTextQuest$^+$ service, we have performed a case study based on an extensive cell cycle dataset (Jensen *et al.*, 2006). According to this study, 600 human genes were assigned to specific phases of the cell cycle, given their expression levels at 100 different time points (Supplementary Figure S2A). We performed four different MeSH-term–based queries (Supplementary Figure S2B) to the BiotextQuest$^+$ platform and retrieved all recent PubMed articles which uniquely mention a specific phase of the human cell cycle by excluding all others (phases M, G1, S, G2) (Soldatos and Pavlopoulos, 2012). BioTextQuest$^+$ retrieved 2431 articles for M-phase, 2475 for G1-phase, 6900 for S-phase and 2218 for G2-phase (Supplementary Figure S2B). The results were subsequently treated as one cluster (*k-means*, $k = 1$), and genes with high TF-IDF were automatically annotated by the BioCompendium service. According to the queries and the high TF-IDF threshold (>19), 24 genes were assigned to the first query ('M-phase'[MeSH Term]), 51 to the second ('G1-phase'[MeSH Term]), 25 to the third ('S-phase'[MeSH Term]) and 49 to the fourth ('G2-phase'[MeSH Term]). We have filtered these gene lists taking into consideration the proposed gene list (Jensen *et al.*, 2006) to detect which of the annotated genes belong to the reference cell cycle dataset. The process returned 8 genes for M-phase, 7 for G1-Phase, 8 for S-phase and 11 for G2-phase (Supplementary Figure S2C). To calculate the coverage of our method, we have highlighted the genes which correspond to the correct phase in their given time points. We observed 75% correctly assigned genes for M-phase, 50% for G1-phase, 75% for S-phase and 82% for G2-phase—largely corresponding to an overall precision of >70%. Similarly, recall can be calculated at 75, 67, 75 and 64%, respectively, again corresponding to a >70% value. While there is no genuine gold standard dataset to assess true accuracy in this kind of scenario, as the target categories are not mutually exclusive, we show the ability of BioTextQuest$^+$ to implement a workflow with genomics-like experiments for the discovery of relevant genes and proteins.

3.3 Knowledge discovery

To assess the power of BioTextQuest$^+$ for knowledge discovery, we have interrogated PubMed abstracts for the involvement of nucleophosmin (NPM) in human diseases. NPM (also known as B23) is a nucleolar protein that participates in a plethora of cellular processes, including ribosome biogenesis, response to genotoxic stress, maintenance of genomic stability and DNA-repair as well as regulation of chromatin modifications influencing transcription (Colombo *et al.*, 2011; Grisendi *et al.*, 2006). Using the query '*(NPM OR B23) AND disease*' and choosing to include both abstracts and MeSH terms with the k-Means clustering algorithm ($k = 3$) on the BioTextQuest$^+$ platform, enables us to focus on the most prevalent disease-related functions of NPM, we retrieved 258 abstracts, with 707 significant biomedical terms forming 3 clusters: Cluster 1 is dominated by the terms '*NPM-ALK*', '*anaplastic lymphomas*' and '*translocation*'; Cluster 2 is enriched for the terms '*NPM*', '*leukemia*' and '*myeloid*' but not for the term '*NPM-ALK*'; and Cluster 3 is characterized by the terms '*nucleolar*', '*NPM*', '*autoantibodies*', '*autoimmune*' and '*immunology*'. NPM is implicated in human carcinogenesis. In anaplastic large cell lymphoma, a t(2;5) (p23; q35) translocation leading to the production of a chimeric protein comprising NPM fused to the anaplastic lymphoma kinase (ALK) is a frequent genetic event and the oncogenic role of NPM-ALK has been established using a variety of experimental models (Grisendi *et al.*, 2006)—represented in Cluster 1. The chimera functions as a constitutively-active protein tyrosine kinase operating in the cytoplasm to activate proliferation-promoting signaling pathways. In acute myeloid leukemia with normal karyotype, a heterozygous mutation in exon 12 of NPM (termed NPMc) results in the localization of the mutated protein in the cytoplasm, thereby reducing the tumor-suppressing properties of nuclear NPM produced by the wild-type allele—represented in Cluster 2. The role of NPM in immune responses and autoimmune diseases is less-well studied but clear links between NPM antigens and susceptibility to a distinct pulmonary vascular phenotype in scleroderma and graft-versus-host disease have been established (Ulanet *et al.*, 2004)—represented in Cluster 3. With BioTextQuest$^+$, the three disease entities have thus been successfully distinguished, allowing a rapid and reliable assessment of the role of NPM in human disease and associations between NPM structure and pathogenic functions. This example elucidates the ability to blindly and reproducibly detect subtleties with little prior information provided. One may argue that the choice of $k = 3$ is biased and reflects some prior knowledge on what is expected to be 'discovered' for this specific case. Nevertheless, choosing larger values of the parameter k (e.g. $k = 5$) only changes the number of resulting clusters, without modifying their semantics. In fact, while a few abstracts seem to cluster within different concepts, such an approach reveals additional and more fine-grained characteristics of the specific corpus (data not shown).

3.4 OMIM-BioCompendium example

The potential of BioTextQuest$^+$ for concept discovery was validated by the examination of potential links between two seemingly unrelated human health conditions: *aging* and *obesity*. A PubMed search using these terms resulted in a significant number of 5939 publications, hindering the identification of commonalities in the pathophysiology and/or molecular features of these diseases. To overcome this bottleneck, the BioTextQuest$^+$ suite was used to query the OMIM reference database for the terms '*Obesity AND Aging*'. A total of 14 document clusters representing 1874 research papers were identified. The largest cluster (Cluster 1) comprising 1095 documents, is dominated by the terms *apolipoprotein*, *ApoE*, *allele*, *polymorphisms* and *Alzheimer's disease*. Apolipoprotein E (ApoE) is a major constituent of chylomicrons and the most abundant apolipoprotein in the central nervous system. The *ApoE* ε4 allele has been associated with hypercholesterolemia, a hallmark of obesity, and with increased risk of Alzheimer's disease, the commonest cause of dementia in the elderly. Thus, the expression of ε4 increases the risk of developing Alzheimer's disease by as much as 7–9 years per allele (Jarvik *et al.*, 1995). In addition to the ε4 haplotype, a promoter polymorphism has been independently associated with Alzheimer's disease risk, suggesting that not only qualitative variability between the various ApoE isoforms but also quantitative differences in ApoE levels may influence the risk of this pathology (Lambert *et al.*, 1998). Genes represented in Cluster 1 were interrogated for signaling pathway similarities through the BioTextQuest+ link to the BioCompendium service. Input genes significantly (adjusted *P*-value $< 10^{-6}$) mapped to inflammatory and metabolic disease-related signaling cascades, such as *Adipocytokine* (KEGG ID: hsa04920), *NOD-like receptor* (KEGG ID: hsa04621), *Toll-like receptor* (KEGG ID: hsa04620) and *RIG-I-like* (KEGG ID: hsa04622) signaling pathways. BioCompendium-based prediction of transcription factors regulating the input genes revealed common regulators, such as members of the NF-kB family (RelA, NF-kB1), the peroxisome proliferator-activated receptors (PPARs) and the retinoid X receptors (RXRs). NF-kB transcription factors are powerful orchestrators of pro-inflammatory gene expression and the PPAR-RXR transcriptional complex plays a critical role in inflammation and energy balance, including triglyceride metabolism and glucose homeostasis. Collectively, the aforementioned BioTextQuest+ analyses suggest that obesity and aging may share common pathogenic pathways and raise the possibility that pharmacological or dietary interventions aiming to control the inflammatory component of these diseases may contribute to healthy aging. The link between obesity and aging is further highlighted by Cluster 12 of this BioTextQuest+ study, comprising 12 documents which are characterized by the terms *telomere* and *chromosomal*. Indeed, obesity in humans has been associated with reduced telomere length in women, a molecular hallmark of aging and associated co-morbidities, such as dementia and cognitive decline (Martin-Ruiz *et al.*, 2006; Valdes *et al.*, 2005). This example illustrates the potential of BioTextQuest+ for rapid knowledge acquisition and concept discovery.

4 DISCUSSION

The field of text mining in life and health sciences expands rapidly, considering the exponential growth of biomedical literature. While techniques such as NER, information extraction, co-occurrence analysis have become more mature, the field is still in its infancy, as most of these approaches have only been used for PubMed-based abstract-centric searches. Currently, a limited number of tools succeed in automatically bridging literature mining, information extraction, integration with external repositories and implementation of workflows for further bioinformatics analysis. Herein, we present BioTextQuest$^+$, a platform principally developed to bridge the gaps between these complementary areas. With PubMed and OMIM repositories as starting points, BioTextQuest$^+$ currently offers automated literature extraction, identification of significant bioentity terms, term-based document clustering, co-occurrence analysis as well as integration with a rich collection of biological databases and automated bioinformatics analysis. While initial versions of this platform focused only on document clustering (Iliopoulos *et al.*, 2001; Papanikolaou *et al.*, 2011), BioTextQuest$^+$ adds a significant set of novel features (Supplementary Table S1). A description of the most salient characteristics of established tools in the field are presented in Supplementary Table S2 and a thorough comparison of BioTextQuest$^+$ with these tools is summarized in Supplementary Table S3. We do believe that a gold standard benchmark dataset or process is largely missing to objectively measure the performance of any such tool. It is worth stressing that this is a significant challenge, partly because different text mining tools are built to address different questions. Evidently, the BioCreAtIvE challenge (Krallinger *et al.*, 2008) or similar community-driven initiatives may become the proper forum for designing and implementing such a gold standard benchmark process. BioTextQuest$^+$ can thus serve as a powerful tool in the fields of biomedical literature mining and data integration by aiding users in concept discovery and new hypothesis generation, enhancing our arsenal in the efforts to tackle the complexity of biological text.

ACKNOWLEDGEMENTS

The authors wish to thank the four anonymous reviewers for invaluable comments. I.I. and V.P.S. would like to thank I. Manolakis for fruitful discussions.

Funding: This work was supported by the European Commission FP7 programmes INFLA-CARE (EC grant agreement number 223151), 'Translational Potential' (EC grant agreement number 285948) and MARBIGEN (EC grant agreement number 264089). This work was also supported in part by the FP7 collaborative projects MICROME (agreement no 222886-2) and CEREBRAD (agreement no 295552).

Conflict of interest: none declared.

REFERENCES

Altman,R.B. *et al.* (2008) Text mining for biology—the way forward: opinions from leading scientists. *Genome Biol.*, **9** (Suppl. 2), S7.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.

Benson,D.A. *et al.* (2011) GenBank. *Nucleic Acids Res.*, **39**, D32–D37.

Berman,H. *et al.* (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.

Bolton,E. *et al.* (2008) PubChem: integrated platform of small molecules and biological activities. *Ann. Rep. Comput. Chem.*, **4**, 217–240.

Boyack,K.W. *et al.* (2011) Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches. *PloS One*, **6**, e18029.

Brohee,S. *et al.* (2008) Network analysis tools: from biological networks to clusters and pathways. *Nat. Protoc.*, **3**, 1616–1629.

Bryne,J.C. *et al.* (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.

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**, W399–W405.

Cochrane,G. *et al.* (2009) Petabyte-scale innovations at the European nucleotide archive. *Nucleic Acids Res.*, **37**, D19–D25.

Colombo,E. *et al.* (2011) Nucleophosmin and its complex network: a possible therapeutic target in hematological diseases. *Oncogene*, **30**, 2595–2609.

Degtyarenko,K. *et al.* (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.

Doms,A. and Schroeder,M. (2005) GoPubMed: exploring PubMed with the gene ontology. *Nucleic Acids Res.*, **33**, W783–W786.

Douglas,S.M. *et al.* (2005) PubNet: a flexible system for visualizing literature derived networks. *Genome Biol.*, **6**, R80.

Eaton,A.D. (2006) HubMed: a web-based biomedical literature search interface. *Nucleic Acids Res.*, **34**, W745–W747.

Enright,A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.

Errami,M. *et al.* (2007) eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. *Nucleic Acids Res.*, **35**, W12–W15.

Feng,Z. *et al.* (2004) Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics*, **20**, 2153–2155.

Fink,J.L. *et al.* (2008) BioLit: integrating biological literature with databases. *Nucleic Acids Res.*, **36**, W385–W389.

Flicek,P. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.

Fontaine,J.F. *et al.* (2009) MedlineRanker: flexible ranking of biomedical literature. *Nucleic Acids Res.*, **37**, W141–W146.

Fontelo,P. *et al.* (2005) askMEDLINE: a free-text, natural language query tool for MEDLINE/PubMed. *BMC Med. Inform. Decis. Mak.*, **5**, 5.

Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.

Fujita,P.A. *et al.* (2011) The UCSC genome browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.

Giglia,E. (2011) Quertle and KNALIJ: searching PubMed has never been so easy and effective. *Eur. J. Phys. Rehabil. Med.*, **47**, 687–690.

Grisendi,S. *et al.* (2006) Nucleophosmin and cancer. *Nat. Rev. Cancer*, **6**, 493–505.

Gunther,S. *et al.* (2008) SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.*, **36**, D919–D922.

Hamosh,A. *et al.* (2005) Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.

He,X. *et al.* (2010) BSQA: integrated text mining using entity relation semantics extracted from biological literature of insects. *Nucleic Acids Res.*, **38**, W175–W181.

Hokamp,K. and Wolfe,K.H. (2004) PubCrawler: keeping up comfortably with PubMed and GenBank. *Nucleic Acids Res.*, **32**, W16–W19.

Iliopoulos,I. *et al.* (2001) Textquest: document clustering of medline abstracts for concept discovery in molecular biology. *Pac. Symp. Biocomput*, **2001**, 384–395.

Jarvik,G.P. *et al.* (1995) Interactions of apolipoprotein E genotype, total cholesterol level, age, and sex in prediction of Alzheimer's disease: a case-control study. *Neurology*, **45**, 1092–1096.

Jensen,L.J. *et al.* (2006) Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature*, **443**, 594–597.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Kersey,P.J. *et al.* (2004) The international protein index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.

Kim,J.J. *et al.* (2008) MedEvi: retrieving textual evidence of relations between biomedical concepts from medline. *Bioinformatics*, **24**, 1410–1412.

Krallinger,M. *et al.* (2008) Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biol.*, **9** (Suppl. 2), S1.

Kuhn,M. *et al.* (2008) STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.*, **36**, D684–D688.

Lambert,J.C. *et al.* (1998) A new polymorphism in the APOE promoter associated with risk of developing Alzheimer's disease. *Hum. Mol. Genet.*, **7**, 533–540.

Lu,Z. (2011) PubMed and beyond: a survey of web tools for searching biomedical literature. *Database*, **2011**baq036.

MacQueen,J.B. (1967) Kmeans some methods for classification and analysis of multivariate observations. In: *5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley CApp. 281–297.

Maglott,D. *et al.* (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.

Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**, bar009.

Martin-Ruiz,C. *et al.* (2006) Telomere length predicts poststroke mortality, dementia, and cognitive decline. *Ann. Neurol.*, **60**, 174–180.

Morris,J.H. *et al.* (2011) clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics*, **12**, 436.

Moschopoulos,C.N. *et al.* (2011) Which clustering algorithm is better for predicting protein complexes? *BMC Res. Notes*, **4**, 549.

Nepusz,T. *et al.* (2010) SCPS: a fast implementation of a spectral method for detecting protein families on a genome-wide scale. *BMC Bioinformatics*, **11**, 120.

Paccanaro,A. *et al.* (2006) Spectral clustering of protein sequences. *Nucleic Acids Res.*, **34**, 1571–1580.

Pafilis,E. *et al.* (2009) Reflect: augmented browsing for the life scientist. *Nat. Biotechnol.*, **27**, 508–510.

Pafilis,E. *et al.* (2013) OnTheFly 2.0: a tool for automatic annotation of files and biological information extraction. In: *Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference*. Chania, Crete, Greece.

Papanikolaou,N. *et al.* (2011) BioTextQuest: a web-based biomedical text mining suite for concept discovery. *Bioinformatics*, **27**, 3327–3328.

Pavlopoulos,G.A. *et al.* (2009a) jClust: a clustering and visualization toolbox. *Bioinformatics*, **25**, 1994–1996.

Pavlopoulos,G.A. *et al.* (2009b) OnTheFly: a tool for automated document-based text annotation, data linking and network generation. *Bioinformatics*, **25**, 977–978.

Pavlopoulos,G.A. *et al.* (2014) Biological information extraction and co-occurence analysis. In: *"Biomedical Literature Mining", Methods in Molecular Biology*. Humana Press, Springer, New York, pp. 77–92.

Pavlopoulos,G.A. *et al.* (2011) Using graph theory to analyze biological networks. *BioData Min.*, **4**, 10.

Perez-Iratxeta,C. *et al.* (2002) Exploring MEDLINE abstracts with XplorMed. *Drugs Today*, **38**, 381–389.

Plikus,M.V. *et al.* (2006) PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. *BMC Bioinformatics*, **7**, 424.

Poulter,G.L. *et al.* (2008) MScanner: a classifier for retrieving Medline citations. *BMC Bioinformatics*, **9**, 108.

Rebholz-Schuhmann,D. *et al.* (2007) EBIMed—text crunching to gather facts for proteins from medline. *Bioinformatics*, **23**, e237–e244.

Rodriguez-Esteban,R. (2009) Biomedical text mining and its applications. *PLoS Comput. Biol.*, **5**, e1000597.

Safran,M. *et al.* (2010) GeneCards version 3: the human gene integrator. *Database*, **2010**, baq020.

Sayers,E.W. *et al.* (2012) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **40**, D13–D25.

Schafferhans,A. *et al.* (2003) The PSSH database of alignments between protein sequences and tertiary structures. *Nucleic Acids Res.*, **31**, 494–498.

Schneider,R. and Sander,C. (1996) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **24**, 201–205.

Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.

Seal,R.L. *et al.* (2011) genenames.org: the HGNC resources in 2011. *Nucleic Acids Res.*, **39**, D514–D519.

Smalheiser,N.R. *et al.* (2008) Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. *J. Biomed. Discov. Collab.*, **3**, 2.

Soldatos,T.G. *et al.* (2012) Caipirini: using gene sets to rank literature. *BioData Min*, **5**, 1.

Soldatos,T.G. and Pavlopoulos,G.A. (2012) Mining cell cycle-related literature using support vector machines. *Lect. Notes Comput. Sci.*, **7297**, 278–284.

Sparck,J.K. *et al.* (2000) A probabilistic model of information retrieval: development and comparative experiments. Part I. *Inform. Process. Manag*, **36**, 779–808.

States,D.J. *et al.* (2009) MiSearch adaptive pubMed search tool. *Bioinformatics*, **25**, 974–976.

Szklarczyk,D. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.

Theodosiou,T. *et al.* (2008) PuReD-MCL: a graph-based PubMed document clustering methodology. *Bioinformatics*, **24**, 1935–1941.

Tsuruoka,Y. *et al.* (2011) Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics*, **27**, i111–i119.

Tsuruoka,Y. *et al.* (2008) FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics*, **24**, 2559–2560.

Ulanet,D.B. *et al.* (2004) Selective cleavage of nucleolar autoantigen B23 by granzyme B in differentiated vascular smooth muscle cells: insights into the association of specific autoantibodies with distinct disease phenotypes. *Arthritis Rheum.*, **50**, 233–241.

Valdes,A.M. *et al.* (2005) Obesity, cigarette smoking, and telomere length in women. *Lancet*, **366**, 662–664.

Vilella,A.J. *et al.* (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.

Wang,J. *et al.* (2010) Interactive and fuzzy search: a dynamic way to explore MEDLINE. *Bioinformatics*, **26**, 2321–2327.

Wishart,D.S. *et al.* (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, D901–D906.

Wishart,D.S. *et al.* (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.*, **37**, D603–D610.

Yamamoto,Y. and Takagi,T. (2007) Biomedical knowledge navigation by literature clustering. *J. Biomed. Inform.*, **40**, 114–130.

Zarin,D.A. *et al.* (2011) The ClinicalTrials.gov results database—update and key issues. *N. Engl. J. Med.*, **364**, 852–860.