Genotation: Actionable knowledge for the scientific reader

Panduka Nagahawatte¹, Ethan Willis¹, Mark Sakauye¹, Rony Jose¹, Hao Chen² and Robert L Davis¹

¹Center of Biomedical Informatics, Memphis, TN 38103, USA; ²Department of Pharmacology, University of Tennessee Health Science Centre, Memphis, TN 38103, USA

Corresponding author: Panduka Nagahawatte. Email: panduka@uthsc.edu

Abstract

We present an article viewer application that allows a scientific reader to easily discover and share knowledge by linking genomics-related concepts to knowledge of disparate biomedical databases. High-throughput data streams generated by technical advancements have contributed to scientific knowledge discovery at an unprecedented rate. Biomedical Informaticists have created a diverse set of databases to store and retrieve the discovered knowledge. The diversity and abundance of such resources present biomedical researchers a challenge with knowledge discovery. These challenges highlight a need for a better informatics solution. We use a text mining algorithm, Genomine, to identify gene symbols from the text of a journal article. The identified symbols are supplemented with information from the GenoDB knowledgebase. Self-updating GenoDB contains information from NCBI Gene, Clinvar, Medgen, dbSNP, KEGG, PharmGKB, Uniprot, and Hugo Gene databases. The journal viewer is a web application accessible via a web browser. The features described herein are accessible on www.genotation.org. The Genomine algorithm identifies gene symbols with an accuracy shown by .65 F-Score. GenoDB currently contains information regarding 59,905 gene symbols, 5633 drug–gene relationships, 5981 gene–disease relationships, and 713 pathways. This application provides scientific readers with actionable knowledge related to concepts of a manuscript. The reader will be able to save and share supplements to be visualized in a graphical manner. This provides convenient access to details of complex biological phenomena, enabling biomedical researchers to generate novel hypothesis to further our knowledge in human health. This manuscript presents a novel application that integrates genomic, proteomic, and pharmacogenomic information to supplement content of a biomedical manuscript and enable readers to automatically discover actionable knowledge.

Keywords: Genomic supplements, journal article viewer, annotation gene, annotation protein, pharmacogenomics

Experimental Biology and Medicine 2016; 241: 1202–1209. DOI: 10.1177/1535370216633795

Introduction

Biomedical researchers rely on peer-reviewed journal articles to improve their knowledge, stay abreast of current information, and with state of the art techniques and ideas. Journal articles often contain inter-related concepts of biological systems including biochemistry, immunology, molecular biology, genetics, diseases, and drug therapies, making it difficult for any single reader to recall details of all the content contained in each article. Readers may seek external knowledge for certain keywords to learn specific details about them. Genetics-related terms such as genes, proteins, mRNA accession numbers, and genomic events represent a subset of such keywords. Each term is implicitly linked to information from different domains such as functional genomics, clinical relevance, pharmacogenomics, and disease implications in disparate biomedical databases. These biomedical databases store structured scientific knowledge for convenient retrieval. Recent technological developments in high-throughput experiments have

contributed to an unprecedented rate of scientific knowledge discovery, and informaticians have introduced a diverse array of biomedical databases in order to facilitate access to discovered knowledge. According to the 2014 Nucleic Acid Database survey, there are 1552 molecular biology databases available online to the scientific community.¹

The objective of this study was to create an informatics solution enabling scientists to conveniently discover and share knowledge while they read scientific journal articles. The source code for this project is available under the MIT license deposited as a GitHub open source project available at https://github.com/pnagahawatte/genotation_public.git

Background

The multiplicity of genetic terms, the abundance of biomedical databases describing those terms and the rate at which

information is updated in these databases present biomedical researchers with a challenge in knowledge management. Therefore, biomedical researchers often resort to doing keyword web searches while reading to supplement their knowledge of the article content. This is an inefficient and relatively time consuming “hunter-gatherer” approach that may not always yield the desired results due to search engine rankings and the large number of search results returned. To address this challenge, there have been multiple attempts in the past to assist a reader by integrating-related content on to the document being read using computational methods.

PubMed “Related Citations” uses a word-weighted algorithm to compare words from the title, abstract, and the MeSH headings where the top matches for each article is pre-computed and stored. The top matches linked to an article are presented to a reader as a match to a user query, allowing the reader to identify similar articles of interest.² However, this approach only considers a part of the document content (title and abstract), and only relates that information to one biomedical database; *PubMed*. *Readcube* is a commercial article library manager applications that automatically provides links to citations and images from supplementary material.³ This application only uses biomedical document corpora when hyperlinking the citations, and supplementary images. It is a desktop application only available on Windows operating system, which creates issues for most scientists who use Linux as their main work operating system. *Reflect* is a web application that accepts an article as input where it highlights and links genetic terms to online resources such as *Wikipedia*.⁴ *Reflect* employed a semantic-web-based approach in finding and hyperlinking relevant resources to supplement keywords of the uploaded article. Europe *PubMed Central* uses an Extensible Markup Language-based annotation stream through a distributed server infrastructure in order to supplement article content.⁵

Here, we present a journal article viewer that automatically links genetic terms such as genes, proteins, and mutations to external information from multiple biomedical databases as actionable knowledge to the reader. The platform consists of three key components: (1) *Genomine*, a named entity recognition (NER) algorithm that searches through an entire document to identify genetic terms, (2) *GenoDB*, a self-updating database synthesized from biomedical databases that contain descriptive, functional, clinical, and pharmacogenomics information, and (3) *Genotation*, the journal article viewer where users will be able to read a journal article alongside supplementary knowledge accessible via an interactive menu. The application will enhance the reading experience by placing additional information relevant to the primary article at readers’ fingertips. The supplements will enable the reader to discover relevant knowledge on genetic terms, drug and disease names without leaving the article to perform time consuming web searches.

Challenges

There were three major challenges identified in the project: (1) Accurately identifying the subset of terms from an

article to be linked to external knowledge, (2) synthesizing a centralized knowledgebase containing information from multiple databases, and (3) presenting the supplementary knowledge intuitively to enhance the reader experience. Our solution, as detailed in Figure 1, is initiated by a user specifying a journal article in portable document format (PDF) or Hyper Text Markup Language (HTML). Once uploaded, the article is converted to plain text, and a copy is preserved in the original format to be displayed back to the user. The *Genomine* algorithm automatically searches through the plain text copy of the article to identify genetic terms that could be linked to supplementary information. Subsequently, each identified term is queried against *GenoDB*, the centralized knowledgebase, to gather descriptive, functional, clinical, and pharmacogenomics information as supplements. These terms are categorized into genes, diseases, and drugs. The original article along with the categorized information is rendered on the user interface; the *Genotation* journal article viewer. The screen area of the user interface is dedicated to displaying the article in the original format, which presents the user with an unhindered view for reading purposes. The supplements are stored inside an interactive menu accessible to the user when reading the document, and the user can expand the menu and search or navigate through the tree of annotations to locate term(s) of interest. While users can drill down to access granular details on terms, the granularity of details creates a challenge in presenting detailed textual descriptions using the restricted screen space. Therefore, the users are presented with hyperlinks that will redirect them to the details from the original source. This web application provides the ability for the user to switch scope when reading the article and to access information within the article negating the need for external searches.

Materials and methods

Genomine is a rules-based NER algorithm based on a dictionary of terms, specialized in identifying genetic terms for augmentation. As shown in Figure 2, it involves three major steps in producing a set of augmented genetic terms identified from a document. The journal articles are commonly published in PDF or HTML format. Searching a binary or mark-up document presents many challenges, such as stylistic components, text within tags, and selecting only the text from binary content. Therefore, the initial step of *Genomine* converts an article from HTML or PDF format to text, which eliminates all stylistic components, and tags while preserving white space and sentence structure. These are accomplished using free source software *Perl HTML::Parser* library (<http://search.cpan.org/dist/HTML-Parser/Parser.pm>) and *pdftotext* application (<http://linux.die.net/man/1/pdftotext>). The resulting plain text is used by *Genotation* algorithm to identify the candidate terms.

The overlap of genetic terms with natural languages creates a challenge in accurate identification of terms such as “APRIL” and “MICE.” A different challenge is presented where complex terms refer to an individual entity, such as “nuclear factor kappa-light-chain-enhancer of

Figure 1 Overview of Genotation. The web application accepts a document in PDF, HTML, and text formats from a user. A copy of the uploaded document is examined by Genome algorithm to identify genetic terms. Subsequently, GenoDB supplements the identified terms with multi-dimensional information and delivers to Genotation interface. Genotation interface is the article viewer application where the supplementary information is organized and presented to the user as interactive annotations. (A color version of this figure is available in the online journal.)

Figure 2 Genome concept. The objective of the algorithm is to accurately identify terms from the article that could be supplemented with further information. The journal article is initially converted to a plain text copy, where the algorithm compares individual words with a dictionary of all possible terms that have external knowledge. An automated validation is performed at this juncture in order to reduce false positives and negatives. The set of identified terms is the result of the algorithm. (A color version of this figure is available in the online journal.)

activated B cells" identified as *NFKB*. Another complex scenario is presented when multiple words are mentioned as a single term by using a connecting character, such as a hyphen. Gene fusions are an example where they are denoted by using a hyphen in between the two genes, such as "BCR-ABL1." In order to solve these challenges, Genome invokes a set of rules-based heuristics such as common word detection, abbreviation detection, neighboring word context analysis, and sentence analysis for

multiple genetic terms in order to increase accuracy. Although Genome is designed to intuitively identify such complex scenarios, there were false positives and negatives present after the initial step of the algorithm. Details of the false positives and negatives are mentioned in the "Systems Evaluation" section in Tables 1 to 4. An automated validation is performed as a secondary measure to minimize such errors in the identified terms. A set of terms that passes the automated validation is passed onto

the subsequent step for augmenting the term with information from GenoDB.

GenoDB is a self-updating, centralized knowledgebase synthesized using data from multiple freely available databases. The current version includes data from NCBI Gene,⁶ HUGO,⁷ ClinVar,⁸ MedGen, dbSNP,⁹ KEGG,^{10,11} Uniprot,¹² and PharmGKB.¹³ Preserving the original meaning, structure and relationships of data when combining data from disparate sources is a challenge. Documentation and meta-data provided by the data source were used to identify the fields to be integrated along with their structure and relationships. GenoDB database was designed to store the selected fields of multiple biomedical databases preserving the original meaning and relationships and is implemented using MySQL relational database management system as shown in Figure 3. GenoDB is populated with multi-dimensional information gathered from above sources, where each genetic term is a unique identifier linking information to functional genomics, genomic events, pharmacogenomics, and clinical relevance.

GenoDB as a knowledgebase must contain the current information to be most effective for the users. This is a

challenge, as the original databases such as ClinVar, MedGen, and dbSNP are constantly updated with novel information. GenoDB is populated using a set of Extract Transform Load (ETL) scripts that gather necessary information from downloaded plain text data files and loads into the database. GenoDB is a self-updating knowledgebase where downloading, extracting, and loading the data occurs once every two weeks. The set of key genetic terms identified by the Genomine algorithm is supplemented with the multi-dimensional information gathered from GenoDB. Genetic terms and their supplements are provided to Genotation interface for visualization.

Several biomedical repositories have made available Application Programming Interfaces (API) in order to programmatically query and retrieve information. Several APIs were evaluated in order to gather necessary information as an alternative option to GenoDB. The evaluation showed that an approach using APIs to be unusable due to latency introduced in the Genotation query design that negatively impacted the user experience.

Genotation results from Genomine and GenoDB are presented to the users on a journal article viewer web application named Genotation. This application is freely available for use on www.genotation.org. The main purpose of the application is to provide supplements to better comprehend the article content without creating visual obstacles

Table 1 Characteristics of BioInfer corpus

Number of sentences	1100
BioInfer tag of interest	Gene/protein/RNA
Total unique terms with tag	307
Number of terms with no overlap to Homo sapiens	206
Number of unique terms in test dataset	101

The sentences and entities were generated by using the *BioInfer* extract software on the *BioInfer* corpus 1.1.1. The annotations were tagged by different categories. Only terms with the *BioInfer* tag "Gene/protein/RNA" were selected to include genetic terms. There were 307 unique terms (greater than two character count) included in the *BioInfer* corpus. However, 206 terms belonged to other species and had no overlap with genetic terms relating to Homo sapiens. Therefore, eliminating those terms from the dataset left 101 unique terms in the dataset.

Table 3 Analysis of *BioInfer* corpus test results

True positives	92
False positives	9
False negatives	11
Precision	0.91
Recall	0.89
F-score	0.90

Performance metrics for the Genomine algorithm using the 1100 sentence annotated *BioInfer* dataset.

Table 2 Description of manually annotated test set

Article title	Category	Article reference	Article ID
Histone methylation: a dynamic mark in health, disease and inheritance.	Epigenetics	Doi: 10.1038/nrg3173	1
The malignant melanoma landscape.	Cancer – Melanoma	Doi: 10.1038/nrd4326	2
RB1 gene inactivation by chromothripsis in human retinoblastoma.	Cancer – Retinoblastoma	PMID: 24509483	3
Genetics of asthma and allergic disease	Immunology – Asthma	Doi: 10.1093/hmg/9.16.2359	4
Influenza: old and new threats	Infectious diseases – Influenza	Doi: 10.1038/nm1141	5
Prioritizing causal disease genes using unbiased genomic features	Cardiovascular disease	Doi: 10.1186/s13059-014-0534-8	6
Denosing perturbation signatures reveal an actionable AKT-signaling gene module underlying a poor clinical outcome in endocrine-treated ER + breast cancer	Cancer – Breast cancer	Doi: 10.1186/s13059-015-0630-4	7

Profile of the manually annotated dataset. These articles represent a wide variety of subjects that contain genetic terms.

Table 4 Analysis of manually annotated document set test results

Article ID	1	2	3	4	5	6	7	Comprehensive
True positives	95	4	65	8	2	20	60	90
False positives	28	0	21	0	3	24	10	37
False negatives	61	4	15	5	1	32	23	61
Precision	0.77	1.00	0.76	1.00	0.40	0.45	0.86	0.71
Recall	0.61	0.50	0.81	0.62	0.67	0.38	0.72	0.60
F-score	0.68	0.67	0.78	0.76	0.50	0.42	0.78	0.65

Performance metrics for the Genome algorithm using the seven article test dataset. Article ID refers to the documents detailed in Table 2. Comprehensive column summarizes the aggregate results of the test dataset.

Figure 3 GenoDB, the synthesized knowledgebase. There are multiple databases that contain valuable biomedical information in verbose detail. Documentation of each database is consulted in identifying the most valuable fields and their structure to be integrated into the knowledgebase. That information is used to create the technical specification for the synthesized database. The database is created on MySQL as detailed in the specification. (A color version of this figure is available in the online journal.)

Figure 4 The entire screen space is dedicated to displaying the journal article, and the supplements are hidden away in the conspicuous Genotation menu. User can expand the menu to view all the supplements in a collapsed state. Click on a term to expand and drill down to the necessary detail. Click on a link to be re-directed to the source website. (A color version of this figure is available in the online journal.)

(e.g. tagging, underlining, and coloring of terms) for the reader. Genotation hides the supplements using an interactive menu that is collapsed by default. The reader has the ability to expand the menu on demand and browse or search through supplements to locate the term of interest. Supplements comprise of descriptive genetic information, genomic events, diseases, pathways, drugs, gene interactions, and related literature relevant to the genetic term of interest. An example of the information available to the user is shown in Figure 4. A reader can drill down to granular levels of information once they identify a term. The verbosity of information increases with granularity, which makes it difficult to display the entire content using a side bar. Details that cannot be fully displayed on the interface will be presented as hyperlinks on the menu, where the user will be re-directed to a web page from the origin of the source describing full details.

Interactions with external web applications

Currently, readers can automatically interact with two external applications through Genotation; Chillibot¹⁴ and GoPubMed.¹⁵ Genotation allows a reader to interact with these external applications by providing hyperlinks to identified genetic terms. Chillibot is a web application to rapidly identify relationships between genes, proteins, and other keywords, where the relationships are inferred using information gathered from the PubMed literature database. This will allow Genotation users to identify relationships between two specific terms of interest. GoPubMed is a search engine that relies on background information using semantic knowledge gathered as concepts from authors, GO and MeSH ontologies, proteins, journals, geolocations, and real-time concepts. This application gives a Genotation user the ability to search related articles, and filter by concepts of interest to identify the most relevant documents.

User instructions

The journal article viewer can be accessed at www.genotation.org using a web browser. Genotation currently has been tested on *Mozilla Firefox* version 37.0.1 and *Chrome* version 42.0. *Internet Explorer* is not currently supported with the application due to the browser's JavaScript handling mechanism. A user has the ability to search for an article of interest or drag and drop an article onto the web application. The search feature currently is configured to only show results provided by the PubMed Central search API. Clicking on a search result will open the article in PDF format on the Genotation article viewer for reading. Copying and pasting a manuscript Universal Resource Locator in the search box will enable a user to activate Genotation in HTML mode. Alternatively, if a user has an article saved on the computer in PDF format, they can drag and drop the document onto the web application and click annotate, which will redirect the user to the Genotation article viewer to read the uploaded document.

A reader has the flexibility to read, navigate, print, and save the article through a native PDF menu as part of the Genotation article reader. The Genotation access menu contains the save, annotations (to view the supplementary

information), and help buttons. These buttons are on the left corner of the Genotation bar, which is displayed above the article, as shown in Figure 4. A user can expand and collapse the menu as they read the article to access pertinent information. A user has the ability to search or scroll to find a term of interest, and drill down to view details. A user may click on a link to read further details on an external site. The external site will be opened in a secondary tab or a browser window depending on the user's browser settings. Genotation provides a feature where a user can save the supplements to their desktop for later use or to be shared. This feature can be invoked by clicking the save icon on the Genotation menu, where the supplemental information will be saved as a graphic. The graphic is based on *d3.js*^{16,17} web graphics, and could be viewed using *Mozilla Firefox* or *Chrome* browsers. Readers are able to interact with the supplements on the saved document as they would be on the Genotation web application.

A video tutorial is available to users via the help button of the Genotation menu that will describe core features of the article reader. Another feature available through the help page is the Genotation blog (<http://www.Genotation.blogspot.com>), which can be used to comment on current features as well as request new features.

Results

The three main components of the framework were tested independently in order to verify individual features. Genomine was tested for accuracy of the algorithm in identifying genetic terms, GenoDB for validity in storing the information in the database and Genotation for usability features. Genomine algorithm was tested using an annotated corpus provided by BioInfer¹⁸ as detailed in Table 1, and a seven document set of documents manually annotated by our internal team as a test dataset, as detailed in Table 2. BioInfer corpus provided 1100 sentences annotated with the genomic entities found for each individual sentence. The BioInfer corpus provided the necessary power to analyze the algorithm, whereas the manually annotated full text document set provided a real-life testing dataset. The objective of Genomine is to identify at least one instance of a genetic term mentioned within the article because Genotation will only present supplementary information for one instance of each term to the user regardless of the incidence of that term within the article. Each document was processed using Genomine, and the identified genetic terms were listed using the document as a key. Subsequently, these results were compared against the annotations of the test dataset in order to identify true positives, false positives, and false negatives. The F-score was computed as follows for the Bioinfer corpus as shown in Table 3 and for the seven document test dataset as shown in Table 4.

$$\text{Precision} = \text{Truepositives} / (\text{Truepositives} + \text{Falsepositives})$$

$$\text{Recall} = \text{Truepositives} / (\text{Truepositives} + \text{Falsenegatives})$$

$$F - \text{score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

The validity of GenoDB was tested involving a multiple step process. A semi-automated evaluation procedure was adopted to test components of the database design. The implemented schema was manually compared to the details of the specification to validate the structure and relationships of stored data. Stored procedures and views were tested using automated scripts for validity. These tests will be conducted anytime a structural change occurs in the database. ETL scripts were validated using two sets of automated tests; one for direct move transformations involving checking consistency of the records copied from source to GenoDB, and another for derived move transformations that involved checking the transformations that happened between the source and GenoDB.

The Genotation article viewer interface was manually tested to ensure the number of genetic terms identified by Genomine was consistent with the items displayed on the interface. Subsequently, tests were conducted to check consistency of the supplements provided by GenoDB to information displayed on the interface. The user interface has to be tested by potential users, and surveys are available on the website that will allow users to comment on existing features based on a Likert scale.

Discussion and conclusion

The difference in F-score between the Bioinfer corpus and the seven document article dataset is mainly due to graphics in PDF files, non-*Homo sapiens* genetic terms, and gene names mentioned in reference section of articles. The graphics in PDF files such as figures and tables contain gene names. In some documents, such as article ID 6 and 7, these graphics' text were not extracted to text format, which led to not identifying the genetic terms, which led to false negatives. The test dataset was compiled by a scientist reading the articles, and denoting the genetic terms without regarding the species. Since GenoDB currently only contains genetic terms of *Homo sapiens*, this led to false negatives in documents such as article ID 1 and 3. False positives identified in the articles were due to genetic terms identified from citations in the Bibliography.

Currently, Genotation disseminates information from public biomedical databases such as NCBI Gene, ClinVar, and MedGen to enhance the experience of scientific readers. However, individual scientists may possess curated datasets specific to their biomedical research niche. These datasets are of great value as they include highly specific information with very little noise. Genotation will evolve into a platform where scientists can collaborate by sharing such datasets to enhance content of peer-reviewed publications. The current GenoDB only contains information relating to *Homo sapiens*' genetic terms. We are currently working on expanding GenoDB to include genetic terms for common organisms used in biomedical research such as *Mus musculus*, *Rattus norvegicus*, and *Saccharomyces cerevisiae*. Furthermore, the next version of Genotation will incorporate semantic web concepts, where the most recent information gathered from the "web of data" will serve a dual purpose: (1) enrich the supplemented knowledge by providing an additional layer of information and (2) to be

used in a graph database algorithm recommender engine that would customize the information to be presented based on a reader profile such as Immunologist, Virologist, Molecular Biologist, etc.

Authors' contributions: PN designed the algorithm and the database, analyzed the results, and drafted the manuscript. EW designed the error handling visualizations, and the browser plug-in. MS designed the user interface. RJ carried out the systems evaluation and analyzed results. HC provided the interface code. RD critically revised the manuscript.

ACKNOWLEDGMENTS

The authors acknowledge:

1. Dr Amali Samarasinghe, Department of Pediatrics at UTHSC for providing insightful comments on the manuscript.
2. Dr Charisse Madlock-Brown, Department of Health Information Management at UTHSC for input on integrating external applications in order to provide literature-based supplements to the user.
3. Jay Leafy and Peter Fox, ITS UTHSC for configuring the web and database servers to host the www.genotation.org website.
4. Todd Barber, ITS UTHSC for providing technical guidance on web application programming.

DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

1. Fernandez-Suarez XM, Rigden DJ, Galperin MY. The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. *Nucleic Acids Res* 2014;**42**:D1-6
2. Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics* 2007;**8**:423
3. Hughes M. ReadCube Desktop. *Charleston Advisor* 2014;**16**:34-8
4. Pafilis E, O'Donoghue SL, Jensen LJ, Horn H, Kuhn M, Brown NP, Schneider R. Reflect: augmented browsing for the life scientist. *Nat Biotechnol* 2009;**27**:508-10
5. Kirsch H, Gaudan S, Rebholz-Schuhmann D. Distributed modules for text annotation and IE applied to the biomedical domain. *Int J Med Inform* 2006;**75**:496-500
6. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res* 2011;**39**:D52-7
7. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res* 2015;**43**:D1079-85
8. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 2014;**42**:D980-5
9. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;**29**:308-11
10. Du JL, Yuan ZF, Ma ZW, Song JZ, Xie XL, Chen YL. KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. *Mol Biosyst* 2014;**10**:2441-7
11. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27-30

12. Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* 2011;**2011**:bar009
13. Altman RB. PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. *Nat Genet* 2007;**39**:426
14. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* 2004;**5**:147
15. Doms A, Schroeder M. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res* 2005;**33**:W783–6
16. Bostock M, Ogievetsky V, Heer J. D(3): data-driven documents. *IEEE Trans Visual Comput Graph* 2011;**17**:2301–9
17. Heer J, Bostock M. Declarative language design for interactive visualization. *IEEE Trans Visual Comput Graph* 2010;**16**:1149–56
18. Pyysalo S, Ginter F, Heimonen J, Bjorne J, Boberg J, Jarvinen J, Salakoski T. BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics* 2007;**8**:50

(Received September 21, 2015, Accepted January 21, 2016)