Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W779–W782***
doi:10.1093/nar/gki417

LitMiner and WikiGene: identifying problem-related key players of gene regulation using publication abstracts

Holger Maier, Stefanie Döhr, Korbinian Grote, Sean O'Keeffe, Thomas Werner, Martin Hrabé de Angelis and Ralf Schneider*

GSF-National Research Center for Environment and Health, Institute of Experimental Genetics, AG BIODV, Ingolstädter Landstrasse 1, D-85768 Neuherberg, Germany

Received February 12, 2005; Revised and Accepted March 21, 2005

ABSTRACT

The LitMiner software is a literature data-mining tool that facilitates the identification of major gene regulation key players related to a user-defined field of interest in PubMed abstracts. The prediction of gene-regulatory relationships is based on co-occurrence analysis of key terms within the abstracts. LitMiner predicts relationships between key terms from the biomedical domain in four categories (genes, chemical compounds, diseases and tissues). Owing to the limitations (no direction, unverified automatic prediction) of the co-occurrence approach, the primary data in the LitMiner database represent postulated basic gene–gene relationships. The usefulness of the LitMiner system has been demonstrated recently in a study that reconstructed disease-related regulatory networks by promoter modelling that was initiated by a LitMiner generated primary gene list. To overcome the limitations and to verify and improve the data, we developed WikiGene, a Wiki-based curation tool that allows revision of the data by expert users over the Internet. LitMiner (http://andromeda.gsf.de/litminer) and WikiGene (http://andromeda.gsf.de/wiki) can be used unrestricted with any Internet browser.

INTRODUCTION

The rapid increase in scientific publications makes it almost impossible for the individual scientist to cope with the large amount of data and to keep up to date with the current literature knowledge. Gaining an insight into a new field of interest without prior knowledge is even more difficult. Web-based literature search portals, such as NCBI's PubMed (1), can support the scientists in finding their respective problem-related publications. But they still have to locate, access and finally read large numbers of papers to find out, for example, the genes that are related to a certain disease.

Here, we describe our LitMiner server, which aims at helping scientists to speed up and facilitate the process of identifying key players in new fields of interest, such as the identification of related genes for a specific disease. For any given query key term that can describe a gene, a disease, a tissue or a chemical compound, LitMiner returns a ranked list of potentially related key terms. Recently the LitMiner system has been used successfully to generate the initial list of disease-related genes that are involved in the gene regulation network of the MODY syndrome (Maturity Onset Diabetes of the Young), which was required for the reconstruction of the gene-regulatory network of MODY by promoter modelling (2).

Several similar approaches have been described previously (3–5). In addition to these approaches and in order to overcome quality problems observed with stand-alone co-occurrence based literature mining, we linked LitMiner to a curation tool (WikiGene) that allows expert users to annotate or improve LitMiner predictions and even to add additional data and detail not available from the abstract. We have made LitMiner and WikiGene fully accessible to the scientific community.

APPROACH AND FEATURES

The basic assumption is that frequent occurrence of pairs of different 'key terms' together in the same abstracts reflects a relationship between them, rather than being mere coincidence. In order to compute ranked lists of key terms potentially related to a given 'key term', the following two procedures are performed.

*To whom correspondence should be addressed. Tel: +49 89 3187 4060; Fax: +49 89 3187 4400; Email: ralf.schneider@gsf.de
Present addresses:
Korbinian Grote and Thomas Werner, Genomatrix Software GmbH, Landsberger Straße 6, D-80339 München, Germany

Annotation of 'key terms' in abstracts that are available from PubMed

'Key terms' belonging to four different categories are annotated:

Genes. Names of genes and gene products. Only gene names and their aliases contained in Ensembl database tables (human, mouse and rat) are used for annotation. Gene names and aliases that also have a non-biological meaning are filtered out manually to avoid annotation of non-relevant key terms.

Compounds. Names and aliases of chemical compounds.

Diseases. A list of disease names and relevant key terms has been adapted manually from the Ensembl database.

Tissues. A list of tissue and organ names and relevant key terms has been adapted manually based on the 'Anatomical Dictionary Browser' located at the Mouse Genome Informatics resource of the Jackson Laboratory.

The annotation of key terms from each of the four categories in PubMed abstracts is done as follows: for every occurrence of 'key terms' x in abstract y, a simple $x \rightarrow y$ entry is added to a MySQL database table. Approximately 4.7 million gene annotations are contained within the database (compounds: 7.6 million; diseases: 2.6 million; tissues: 15.4 million annotations).

Calculation of a score value for every pair of key terms

For any pair of key terms KT1 and KT2, a co-occurrence score is calculated in the following way: $\mathrm{OVS(KT1-KT2)} = \mathrm{TNA} \times \mathrm{NCO(KT1-KT2)}/[\mathrm{NA(KT1)} \times \mathrm{NA(KT2)}]$ where OVS(KT1−KT2) is the overrepresentation score for key terms KT1 and KT2; TNA is the total number of abstracts examined; NCO(KT1−KT2) is the number of abstracts in which key terms KT1 and KT2 occur together; NA(KT1) is the number of abstracts in which key term KT1 occurs; and

GSF – Forschungszentrum für Umwelt und Gesundheit in der Helmholtz-Gemeinschaft

LitMiner

AG BIODV

Home | Genes | Compounds | Diseases and Phenotypes | Tissues and Organs | Contact

Genes co-annotated with disease/phenotype: "MODY" (900 articles)

Gene	Species	Locus	Is TF Filter: all / TF	Articles Filter > 10	Number of co-annotated articles Filter > 10	Overrepresentation score Filter > 3	Apply Filters
HNF4A	*Homo sapiens*	20q13.12	X	20	13	9295	Co-annotated tissues/organs
HN4A_RAT	*Rattus norvegicus*	3q42		18	11	8739	Co-annotated tissues/organs
Neurod1	*Mus musculus*	2D		21	6	4086	Co-annotated tissues/organs
NDF1_RAT	*Rattus norvegicus*	3q23		22	6	3900	Co-annotated tissues/organs
Ipf1	*Mus musculus*	5G2		85	23	3869	Co-annotated tissues/organs
Hnf4	*Mus musculus*	2H3		77	12	2229	Co-annotated tissues/organs
GCKR	*Homo sapiens*	2p23.3		21	3	2043	Co-annotated tissues/organs
Tcf2	*Mus musculus*	11B5		21	3	2043	Co-annotated tissues/organs
TCF2	*Homo sapiens*	17q12	X	65	9	1980	Co-annotated tissues/organs
IPF1	*Homo sapiens*	13q13.1	X	218	26	1706	Co-annotated tissues/organs
Gck	*Mus musculus*	11A1		542	42	1108	Co-annotated tissues/organs
GCK	*Homo sapiens*	7p13		542	42	1108	Co-annotated tissues/organs

Figure 1. Example output from LitMiner. The search query key term was 'MODY', a disease term describing the 'maturity onset diabetes of the young' variant of diabetes. The table shows a list of genes that are potentially related to MODY. The headline indicates that the disease key term 'MODY' occurs in 900 abstracts. The top-scoring gene 'HNF4A' only occurs in 20 abstracts, but in 13 of these abstracts together with 'MODY', which leads to the high scoring value of 9295. Results can be filtered manually to adjust sensitivity of the co-occurrence analysis.

NA(KT2) is the number of abstracts in which key term KT2 occurs. In other words, the co-occurrence score is the factor by which the observed co-occurrence frequency of two given 'key terms' KT1 and KT2 exceeds the frequency that could be expected if both 'key terms' were equally distributed among all abstracts.

Curation of LitMiner-predicted relations

To allow curation of LitMiner predictions, we imported a pre-computed set of relations covering most key terms from all categories with default parameters into WikiGene, the LitMiner data curation tool. Generally, a so-called 'Wiki' is a web-based editor system that allows rapid modification of web page contents. The idea behind 'Wiki' is that a common-interest community curates, increases and validates the data. This has been successfully implemented for the Wikipedia encyclopedia (http://www.wikipedia.org). Editing can be done easily by clicking the 'Edit' button on the page and using a simplified HMTL notation to alter the content of the page. WikiGene applies the 'Wiki' idea to the field of gene regulation and allows expert users not only to verify or to reject statements, but also to add additional information. A simple syntax is used to describe basic gene-regulatory events. The interface of WikiGene is currently very simple and focuses on functionality because our resources did not allow a more sophisticated implementation at this time. However, we prefer to provide scientists with a simple working system immediately rather than postpone the important curation issue till a more full-fledged interface has been developed.

The LitMiner CGI and annotation components are written in PERL using the CGI, DBI and DBD::mysql modules. A C program carries out calculation of co-occurrence scores. WikiGene is based on the MediaWiki software. LitMiner and WikiGene are running on a Compaq XP1000 workstation.

USAGE

LitMiner

The access to LitMiner and its use are free of charge for all users and registration is not required. On the LitMiner start page, users have the option to initiate their analysis of the publication abstract with a gene, a chemical compound, a disease or a tissue. For any of the four categories, the user can either browse an alphabetical list of key terms or search the key terms.

HNF4A links to other genes (please add more lines):
Click here to see predicted gene-gene links on LitMiner

- (Details) **HNF4A** <--> GCK (--LitMiner 12:00, 18 Mar 2004 (CET))
- (Details) **HNF4A** <--> IPF1 (--LitMiner 12:00, 18 Mar 2004 (CET))
- (Details) **HNF4A** <--> NEUROD1 (--LitMiner 12:00, 18 Mar 2004 (CET))
- (Details) **HNF4A** <--> NM_017569 (--LitMiner 12:00, 18 Mar 2004 (CET))
- (Details) **HNF4A** <--> NM_133379 (--LitMiner 12:00, 18 Mar 2004 (CET))
- (Details) **HNF4A** <--> Q15636 (--LitMiner 12:00, 18 Mar 2004 (CET))
- (Details) **HNF4A** <--> SPG3A (--LitMiner 12:00, 18 Mar 2004 (CET))
- (Details) **HNF4A** <--> TTN (--LitMiner 12:00, 18 Mar 2004 (CET))

HNF4A links to tissues (please add more lines):
Click here to see predicted gene-tissue links on LitMiner

- (Details) **HNF4A** <--> Beta cell (--LitMiner 12:00, 18 Mar 2004 (CET))
- (Details) **HNF4A** <--> Hepatocyte (--LitMiner 12:00, 18 Mar 2004 (CET))
- (Details) **HNF4A** <--> Liver (--LitMiner 12:00, 18 Mar 2004 (CET))

HNF4A links to diseases (please add more lines):
Click here to see predicted gene-disease links on LitMiner

- (Details) **HNF4A** <--> Diabetes mellitus, insulin-dependent, 4 (--LitMiner 12:00, 18 Mar 2004 (CET))
- (Details) **HNF4A** <--> Diabetes mellitus, noninsulin-dependent (--LitMiner 12:00, 18 Mar 2004 (CET))
- (Details) **HNF4A** <--> Diabetes mellitus, type II (--LitMiner 12:00, 18 Mar 2004 (CET))
- (Details) **HNF4A** <--> Insulin-dependent diabetes mellitus (--LitMiner 12:00, 18 Mar 2004 (CET))
- (Details) **HNF4A** <--> MODY (--LitMiner 12:00, 18 Mar 2004 (CET))
- (Details) **HNF4A** <--> Non-insulin-dependent diabetes mellitus (--LitMiner 12:00, 18 Mar 2004 (CET))
- (Details) **HNF4A** <--> Non-insulin dependent diabetes mellitus, susceptibility to (--LitMiner 12:00, 18 Mar 2004 (CET))

HNF4A links to chemical compounds (please add more lines):
Click here to see predicted gene-compound links on LitMiner

- (Details) **HNF4A** <--> INSULIN (--LitMiner 12:00, 18 Mar 2004 (CET))

Figure 2. Example page from WikiGene showing LitMiner-predicted relations of the human gene 'HNF4A'. Clicking on (details) will link to a page where information on the particular relation can be viewed or added by the user. The user can add new relations by just adding a new line to the list using the standard Wiki editing function.

Having identified the start key term, e.g. a gene name, the user can start the abstract analysis by choosing, as a set of results, a list of potentially related genes or a list of potentially related tissues. Upon request, LitMiner either accesses a pre-calculated ranked results table or calculates the co-occurrence scores, which may take some time. An example output is shown in Figure 1.

Resulting lists can be filtered by setting limits for the minimum number of key term citations or minimum number of co-occurrences required. This feature allows adjustment of the sensitivity of the relationship prediction. In addition, filtering for transcription factors is available for gene lists.

We have shown a more extensive application of LitMiner as part of a recently completed project where the system was used to compile initial gene lists for a pathway-oriented analysis (2).

WikiGene

The Wikipedia project (http://www.wikipedia.org) applies the very simple idea, that the community of users that are experts in their respective field can improve the amount and quality of data in any database. Therefore, they developed a web-based system that allows the management of encyclopedia entries over the Internet.

To improve the quality of data in our LitMiner database, we took advantage of this idea and implemented WikiGene as the data curation user interface for gene-related data.

This will allow expert users in the field of gene regulation to edit the information that has been generated automatically by LitMiner. To further improve the data, the interface also gives the user the unique opportunity to add new data and information not available from the PubMed abstract. It also enables the expert users to add valuable experimental data from their own laboratory research, but that has not yet been published (usually because it was deemed surplus to the paper). The interface also makes it easier to obtain high-quality data via direct input, rather than by way of transforming existing data into natural language text for publication and then trying to reconstruct the original information (by automatic data extraction and processing). We are aware that even an expert user might enter data that others cannot reproduce. However, we hope to gain enough users, so that others working in the same field will correct errors. Figure 2 illustrates an example page from WikiGene, which shows LitMiner-predicted relations of the human gene 'HNF4A'.

We therefore believe that the LitMiner and WikiGene package is a valuable combination of tools, which may support researchers in the fields of molecular biology and biomedical research.

ACKNOWLEDGEMENTS

The authors thank the Deutsche Forschungsgemeinschaft (DFG) for funding of the project (WE2370/1-1, WE2370/1-2 and SCHN 746/1-3), the European Molecular Biology Laboratory (EMBL) for developing and providing free access and use of the Ensembl database and the US National Library of Medicine (NLM) for providing free access to their databases. Funding to pay the Open Access publication charges for this article was provided by DFG.

Conflict of interest statement. None declared.

REFERENCES

1. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, D39–D45.
2. Döhr,S., Klingenhoff,A., Maier,H., Hrabé de Angelis,M., Werner,T. and Schneider,R. (2005) Linking disease-associated genes to regulatory networks via promoter organization. *Nucleic Acids Res*., **33**, 864–872.
3. Becker,K.G., Hosack,D.A., Dennis,G.,Jr, Lempicki,R.A., Bright,T.J., Cheadle,C. and Engel,J. (2003) PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics*, **4**, 61.
4. Jenssen,T.K., Laegreid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet*., **28**, 21–28.
5. Wren,J.D. and Garner,H.R. (2004) Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics*, **20**, 191–198.