Published online 20 May 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W153–W159***
doi:10.1093/nar/gkp392

SENT: semantic features in text

Miguel Vazquez[1], Pedro Carmona-Saez[2], Ruben Nogales-Cadenas[2], Monica Chagoyen[3], Francisco Tirado[2], Jose Maria Carazo[3] and Alberto Pascual-Montano[2,*]

[1]Software Engineering Department, [2]Computer Architecture Department, Complutense University and [3]Biocomputing Unit, National Center for Biotechnology, CNB-CSIC, Madrid, Spain

Received January 31, 2009; Revised April 20, 2009; Accepted April 30, 2009

ABSTRACT

We present SENT (semantic features in text), a functional interpretation tool based on literature analysis. SENT uses Non-negative Matrix Factorization to identify topics in the scientific articles related to a collection of genes or their products, and use them to group and summarize these genes. In addition, the application allows users to rank and explore the articles that best relate to the topics found, helping put the analysis results into context. This approach is useful as an exploratory step in the workflow of interpreting and understanding experimental data, shedding some light into the complex underlying biological mechanisms. This tool provides a user-friendly interface via a web site, and a programmatic access via a SOAP web server. SENT is freely accessible at http:\\sent.dacya.ucm.es.

INTRODUCTION

Advances in biomedical technologies such as DNA microarrays have enabled researchers to identify a large number of molecules simultaneously, opening the path to study biological systems from a global perspective. These techniques have been routinely used in research labs all around the world, generating huge amounts of data. The methods used to analyze and process this data have evolved significantly in the recent years, to the point that they can be considered mature. The interpretation of the results of the analysis, however, still remains one of the main challenges in bioinformatics, mainly due to the inherent complexity of biological systems.

One of the most notable initiatives to help the interpretation of a list of genes is the Gene Ontology (GO) (1). Several approaches use GO annotations to discover what biological terms are significantly enriched in a list of genes. This is an example of an annotation based approach to functional interpretations, a good review of the topic can be found in (2). Annotation based approaches provide a fast, easy and statistically sound interpretation of a list of genes. Although this information is extremely useful for the analysis of gene sets, its scope is limited by structured vocabularies and curated annotations.

Literature mining offers an interesting alternative to annotation based methods. The rationale behind it is that it contains much richer information about the function of genes that can be captured in structured vocabularies. Biomedical literature covers almost all aspects of biology and biochemistry, and with almost no limit to the types of information that may be recovered through careful and exhaustive mining (3). Many researchers have focused their attention in the use of text mining, with methodologies that go from determining protein–protein interactions from biomedical texts (4–7), to providing summary descriptions for genes or determining their similarities (8–12). Even though lots of works in this area have been reported, the practical use by the scientific community is hindered by the lack of efficient and easy to use software.

In a previous work we introduced a technique, based on Non-negative Matrix Factorization (NMF), to extract semantic features from the biomedical literature associated to a list of genes (13). The use of the term 'semantic features' was first introduced by Lee and Seung (14) to describe the NMF factors that group semantically related words, and has been used in this work to follow this nomenclature. These semantic features were able to characterize the biological meaning of the gene list by capturing be main biological topics that where discussed in the articles. Relationships between the genes could be established on the basis of their relationship to these semantic features. The technique has shown a great potential to analyze large literature collections, and has centered the attention of several works in the field (15–17).

This contribution presents a working, usable implementation of a methodology based on (13). SENT (semantic features in text) allows users to explore the biomedical literature associated to a list of genes by summarizing its contents in semantic features, and allowing the user to browse intelligently the relevant articles. It also includes several assisting functionalities like GO enrichment analysis, provided by the GENECODIS web server (18). SENT offers its services through an easy to use web site, and

*To whom correspondence should be addressed. Tel: +34 913 944 420; Fax: +34 913 944 687; Email: pascual@fis.ucm.es

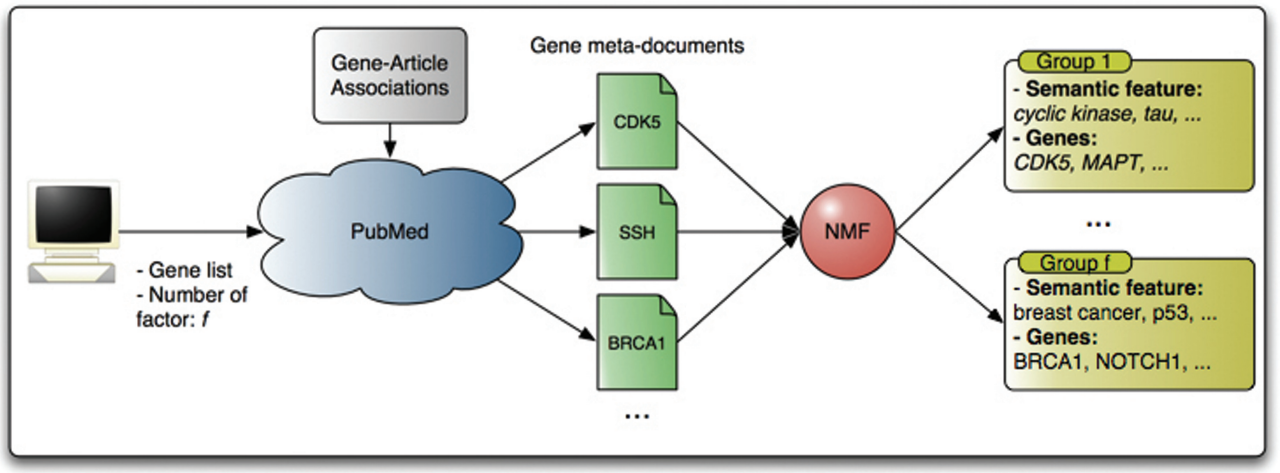

Figure 1. General schematic view of SENT. A set of meta-documents (merged documents associated to each gene) are decomposed by the NMF algorithm to produce groups of semantic features (sets of semantically related words) with their associated genes.

through an SOAP API that allows researchers to use it inside their own scripts and workflows.

METHODS

A general overview of the data analysis workflow implemented in SENT is presented in Figure 1. The input of the system is a set of gene identifiers and the number of semantic features (factors) to use in the NMF analysis. Titles and abstracts from articles associated to each gene are used to produce a meta-document and from all gene meta-documents a term frequency matrix is created. This matrix is then analyzed by means of the NMF algorithm yielding a set of semantic features and a way to associate genes to these semantic features.

The collection of articles used in the analysis is built into an index. This index can be queried to retrieve articles that mention certain terms. In particular, it can be used to find the articles that are most relevant to each semantic feature and, by extension, most relevant to understand the list of genes. This way the user can clarify and ground his interpretation of the semantic features by contrasting the literature. Coupled with the GO enrichment analysis, also provided in the web site, SENT serves as a guide in the examination of the literature.

Therefore, the methodology in SENT can be divided in three steps: Finding the articles to associate to each gene, processing the text into a vector representation that can be analyzed, and finally, analyzing that data to find the semantic features. The next sections describe these three steps in more detail, plus the indexing of documents for the literature examination.

Literature retrieval

To determine the set of articles to associate to each gene, SENT uses two sources:

(i) Curated resources. Databases such as GeneRIF (19) and the GO (1) already provide gene-articles associations. GeneRIF (Gene Reference Into Function) links any article about the basic biology of a gene or protein to the corresponding entry in Entrez Gene, while GO includes references to articles to support the associations of genes to GO annotations.

(ii) Associations automatically derived from the literature. SENT uses PubMed to find articles in which a given gene is explicitly mentioned in the abstract. To this end, a PubMed query is executed containing the organism name and, optionally, to narrow down the search, the words 'gene' or 'protein'. This retrieves a broad corpus of literature in which Named Entity Recognition (NER) and Normalization methodologies are used to find explicit mentions to genes in the article abstracts. NER is the process of finding mentions of entities in text, which in the particular context of gene names is called Gene Mention Recognition. Finding mentions to genes is only the first part of the problem, once the mentions are found the system needs to determine the specific gene they refer to. This problem is known as Gene Mention Normalization, and is aggravated by the often ambiguous ways in which genes are mentioned in free text. Both gene mention and normalization have received a considerable attention by the bioinformatics text mining community, and are a central task on several competitions, such as TREC and, specially, BioCreative (20,21). In SENT we have implemented solutions to both these problems, following many of the central ideas on the state of the art methodologies. Since this is a computationally time consuming process, the collection of articles examined for each organism is limited to the 30 000 most recently published.

Data processing

We construct a meta-document for each gene merging the text from the titles and abstracts of all its

associated articles. A vector representation (22) is generated from these meta-documents following a standard text-mining procedure: We remove words appearing in a list of very common English words, called a stop-word list. The rest of the words are reduced to their stem using the Porter stemmer (23), to group related words with a common stem, like 'telomer' and 'telomeric'. The stems resulting from the previous step are collected individually (unigrams) and in all overlapping pairs (bigrams) to form the bag-of-words representation of the document, called that way because the order of appearance the terms is no longer considered.

The number of individual terms that can appear in the bag-of-words is very large, and this would cause a problem in most text-mining applications. To select the most useful subset, we apply a filter to remove the terms that appear in too many documents ($\geq$80%), as they can be seen as a background, non-informative signal, or in to few documents ($\leq$0.5%), as they are to rare to be useful. We score each of the remaining terms using the Term Frequency, Inverse Document Frequency (TF–IDF) measure (24). This measure balances favoring a words frequency (TF, if it appears many times) and its specificity (IDF, if it does so in only a few documents). With this score we do another filtering to select only the top 3000 best terms (w). The genes are represented as w-dimensional vectors where each coefficient is the frequency of a particular term in the genes meta-document, multiplied by the IDF value for that term.

The above filtering process was done by looking at the frequency of appearance of each term in a collection of documents. SENT supports two options to define what this collection of document should be:

- the meta-documents for the complete list of genes in the organisms' genome. This will select broad and general terms for the word vectors.
- The meta-documents for just the genes in the input query. The terms selected are those important to the specific query, and thus, may be of a higher level of detail.

The first option is the default way the analysis is done, what we call a *standard analysis*. The second option is used in what we call the *fine-grained analysis*. There is an additional type of analysis, called the *custom analysis*, in which the user may provide a list of entities with their associated articles. The custom analysis allows the user to explore genes from unsupported organisms or even entities of other nature, like diseases, authors, journals, etc. The custom, like the fine-grained, also uses the collection of articles in the actual query to perform the filtering.

Note that in fine-grained and the custom analysis the computations must be done on-line, as opposed to the standard analysis in which the vectors may be precomputed. These computations are actually the most resource consuming part of the process, and may delay the jobs considerably.

Analysis

The previous step left us with a collection of n w-dimensional word-vectors representing the n genes in the input list query. The w dimensions represent the w selected terms, ideally 3000. These vectors are arranged as columns of a matrix **M** of dimensions wxn. We use NMF to factor the **M** matrix into two non-negative lower rank (f) matrices

$$\mathbf{M} = \mathbf{WH}$$

where f is the number of factors or semantic features. **W** is a wxf projection matrix and **H** is the coefficient matrix of dimension fxn. The column vectors of the **W** projection matrix are called semantic features, due to the fact that they are collections of semantically related terms. The columns of **H** project the original gene vectors in this new low rank space spanned by the **W** matrix. These vectors are known as the gene profiles, since they can be seen as expressing the genes as combinations of the semantic features. To calculate the NMF model we use the bioNMF web-server application reported in (25).

The NMF algorithm does not necessarily find the best solution. To cope with this situation several initializations strategies have been proposed to improve the convergence rate and eventually find better solutions under certain conditions. The NNDSVD (26) and CENTROID (27) methods are good examples that have proved to be suitable for this problem and will deserve our attention in future versions of this application.

In SENT we decided to take advantage of the nondeterministic nature of NMF to assess the stability of the factorizations at different ranks, using the approach introduced in (13). This approach is based on the collection of a series of results from repeated executions with random initializations. The rationale behind this is that strong signals that are present in the data will be captured by some factor and maintained from execution to execution. Weak signals, on the other hand, or inappropriate choice of the number of factors, will result in noisy factors across executions and won't be maintained. The extent to which factors are reproduced between executions can be used as a measure of appropriateness of the factorizations. We use the cophenetic correlation coefficient as well as a clustering heat-map as assessment of this appropriateness. The actual factors reported by the application are the results of clustering together the factors resulting from 10 separate runs, attending to their semantic features, and averaging both the semantic features and the correspondent gene semantic profiles for each cluster. We use as many clusters as factors originally specified. We will use the terms 'factors' and 'semantic features' to refer to the correspondent cluster averages for simplicity.

We select the most representative terms for each semantic feature as its description. A more sensible selection of terms picks the ones that are both important and exclusive for each factor. We use the following score function for the score of a term t in the i semantic feature

$$S_{i,j} = W_{t,i} / \left\{ average\left(W_{t,j}\right) | \forall_{j \neq i} \right\}$$

The representative terms are those with the 15 best scoring.

In this context, the well-known singular value decomposition (SVD) can also be used to find a low rank approximation of a data matrix. This technique and its application in texts is know as latent semantic indexing (LSI) (28), and show some similarity with the methodology described here.

The main benefit of NMF over LSI is interpretability. The non-negativity of the factors and the projection of the rows in the factor space make them easier to interpret as opposed to when there are negative coefficients as it happens with SVD-based approaches. In addition; the factors found by SVD are designed to incrementally capture all the variance in the data and thus larger factors explain most of data while the rest of factors explain the residuals. On the contrary NMF factors try to capture local signals, and thus they might have equivalent importance. Furthermore, NMF factors show certain degree of overlap as they are not required to be orthogonal. These characteristics make NMF factors more attractive than SVD for interpretability

Literature indexing

SENT builds an index with the titles and abstracts from all the articles associated to the genes in the input list using Ferret, a version of the popular indexing engine Lucene for the Ruby programming language. The index can be queried and will assign scores to the articles. These scores are used, for example, to sort the articles for relevance to a semantic feature.

SOFTWARE USAGE

The input for the application consists in a list of gene identifiers for a given organism. SENT supports ids from several databases, which are listed under the 'Supported ids' in the help section. Currently SENT supports the following organisms: *Candida albicans*, *Caernorhabditis elegans*, *Homo sapiens*, *Arabidopsis thaliana*, *Rattus norvegicus*, *Mus musculus*, *Saccharomyces cerevisiae* and *Saccharomyces pombe*.

SENT allows users to simultaneously explore the data at different resolutions determined by different ranks in the factorization, which must be between 2 and 32. The factorizations at the selected ranks will be produced sequentially in increasing order and made available as they are finished. Once all the analysis are completed another batch of factorizations can be scheduled, either to try with other ranks or to recalculate the results for a given value.

The creation of the literature index can be also scheduled to build it in the same analysis or to leave it for latter. In addition, the application offers two types of analysis: standard and fine grained analysis, being the former the default option. The 'fine grained' analysis produces semantic features whose terms are more specific, but it may take some time (average estimations are from 15 to 30 min). The job name can be used to access the results page at any latter time. Also, if an email is specified it will be used to notify the completion of the first factorization, this is useful for lengthy fine grained jobs.

Once the results are available users may explore the different resolutions to find which of them show a better stability. This can be done looking for high cophenetic correlation values for each factorization or by looking at the heat-maps. At this point one can detect that certain genes are sparsely annotated and produce clear blocks of useless terms relating to analysis methodologies, for example gels, microscopy, spectrometry, etc. If this is the case the genes can be removed and the analysis can be redone. Also, interesting genes may bundle together in large general groups; instead of increasing the resolution of the analysis an interesting option would be to use these genes in a separate analysis. Because the gene list may use some cleaning, it is advised to leave the fine grained analysis and build the literature index (both costly operations) for a second analysis job, after these issues have been assessed.

Once a factorization is found, the different factors in the results provide a general idea of the topics latent in the literature. From this point there are several alternatives to further interpret the groups of genes associated to each factor. One option is to examine the 'Gene Details' page to view results from a GO term enrichment analysis of the genes in the group. Another option is to look at the literature explorer for that group. The literature explorer will show the articles related to the genes in the group sorted for relevance to the terms in the semantic feature. To determine the role of a certain gene in the group we may visit the outside description page (e.g. Entrez Gene for the human), or use the literature explorer to review the articles related to that gene. It is also worth noting that the literature explorer also supports custom queries to the index, which will be used to rank the collection of articles according to that particular query. This is useful to find how a gene relates to each particular term in the feature for example.

With these functionalities the user should be able to find interesting nuggets of knowledge in the data and retrieve the relevant articles that support the findings.

USE CASE

To exemplify the type of results that can be obtained we present the analysis of 50 genes reported by Homayouni *et al.* (29). The 50-gene set is based on the manual selection of genes related to cancer biology, Alzheimer's disease and development, and includes five genes that are involved in the Reelin pathway (RELN, VLDLR, LRP8, DAB1 and FYN). Reelin is an extracellular protein that controls neuronal positioning, formation of laminated structures and synapse structure in the developing central nervous system. In addition some components in the Reelin signaling pathway are associated with Alzheimer disease.

SENT offers this list of genes as the *H. sapiens* example dataset, and can be loaded from the main form. Table 1 shows the semantic features and associated genes for reelin dataset from one analysis in which four groups were formed using the fine-grained analysis.

Table 1. Reelin dataset summarized into four groups using fine-grained analysis

1	**Terms:** cdk5, tau, cyclin depend, gsk, calpain, gsk 3beta, depend kinas, 3beta, microtubul, cyclin, ser, cdc2, microtubul associ, cdk2, kinas activ **Genes:** CDK5, CDK5R1, CDK5R2, FYN, MAPT
2	**Terms:** hedgehog, brca1, wnt, kit, breast, egfr, sonic, myc, breast cancer, basal cell, ptc, p53, notch, patch, renal **Genes:** ABL1, ATOH1, BRCA1, BRCA2, DLL1, DNMT1, EGFR, ERBB2, ETS1, FOS, GLI1, GLI2, GLI3, JAG1, KIT, MYC, NOTCH1, NRAS, PAX2, PAX3, PTCH1, ROBO1, SHH, SMO, SRC, TGFB1, TP53, WNT1, WNT2, WNT3
3	**Terms:** reelin, dab1, apo, lrp, lipoprotein, apolipoprotein, densiti lipoprotein, lipoprotein receptor, low densiti, ldl, macroglobulin, ldl receptor, schizophrenia, receptor relat, apolipoprotein apo **Genes:** A2M, APOE, DAB1, LRP1, LRP8, RELN, VLDLR
4	**Terms:** app, amyloid, presenilin, fe65, precursor protein, amyloid precursor, abeta, secretas, gamma secretas, alzheim, alzheim diseas, beta amyloid, protein app, ptb, amyloid beta **Genes:** APBA1, APBB1, APLP1, APLP2, APP, PSEN1, PSEN2, SHC1

The first group contains the cyclin-dependent kinase 5 and receptors and two additional genes, FYN and MAPT. These genes were also grouped together in the original work of Homayouni *et al.* (29). The semantic feature contains terms related to kinases and terms such as 'tau' and 'microtubl associ'. Indeed, Cdk5 is one of the major kinases that phosphorylate the microtubule-associated protein tau, which is encoded by the MAPT gene. Exploring the literature associated to this group we found articles that discus the role of the cyclic dependent kinases in the phosphorylation of the tau protein among the first ranked abstracts.

The second group is related with cancer and development while the third group contains genes from the Reelin pathway (except FYN) and some genes related with Alzheimer. The fyn kinase has been largely associated with cancer and its association to Reelin has only recently been demonstrated. The terms in this semantic feature also provide insights into the role of reelin in 'binding to lipoprotein receptors and especially low density lipoprotein receptors' as claimed in the highest ranked article (30) in the literature associated to this feature. In addition, there is apolipoprotein E blocks the interaction of Reelin with its receptors and is also considered a risk factor for late-onset Alzheimer disease.

Finally, the fourth group is associated to a semantic feature that clearly captured terms related with Alzheimer disease and most of the genes originally included in this category.

Comparison with similar tools

Other tools that enable literature-based knowledge discovery include GenCLiP (31) and FAUN (17). GenCLiP is a Windows application that offers clustering of genes and terms based on the literature. It also generates gene networks based on co-occurrence of genes in the literature associated to certain keywords. The clustering is done to identify sets of terms related to sets of genes as described in Chaussabel and Sher (8). GenClip allows users to find specific information based on interesting keywords. The clustering step is comparable in principle to the analysis in SENT, while the network generation based on co-occurrence could be considered a different application. Compared with SENT we think both applications focus on different goals and could complement each other. One of the main benefits of SENT over GenCLiP is processing time and simplicity in the interaction with the user.

FAUN on the other hand is a NMF-based text mining tool similar to SENT. The FAUN application is available at http://grits.eecs.utk.edu/faun (17) and at the time of this writing it only contains models for the same 50 gene dataset from Homayouni *et al.* (29) that was used in our use case. They provide visualizations at three resolutions; high, low and medium, meaning 10, 15 and 20 factors respectively. However, FAUN lacks the possibility to explore new datasets in almost real-time manner. With regards to the actual features extracted, the comparison of both applications seems to show the most relevant information at the first ranked items (results not shown).

Another important difference among both tools is the way in which genes are associated to features. In FAUN this is done in a fuzzy way, where each gene can belong to several groups at the same time, while in SENT genes are assigned to one and only one group. Both applications provide with tools to investigate the results further. In this area FAUN stands out showing the relation of each gene to each of the terms of the feature and also shows a gene-gene correlation matrix. In SENT we can use the literature explorer to find articles associated to the genes containing the terms of interest. In addition, SENT provides a collection of result files for downloading, in particular the semantic profiles for each gene, which can be used to calculate the gene–gene correlation matrix.

One of the most interesting features in FAUN is the sentence highlighting, which marks relevant sentences for each gene in relation to each of the features. SENT offers a similar functionality that highlights, not sentences, but complete abstracts. Both alternatives are based on the same idea but work at different resolutions.

SOAP WEB SERVER

All the jobs that are issued from the web site are forwarded to a SOAP server, so the web site can be seen as a front-end to this server. The SOAP server offers an API that can be used to access the functionality programmatically from other work-flows or scripts. Any job issued in the SOAP web server can be examined in the web site using its unique identifier. The web site help section includes an API description, the WSDL file, which is an XML file that most SOAP libraries can use to automatically set up a client for the server, and an example script, that can be used as a command line tool to launch jobs.

CONCLUSIONS

We have developed SENT, a web-based tool for the functional analysis of gene lists extracted from the biomedical literature. The main motivation to construct this tool is the lack of available user-friendly software to automatically analyze large amounts of documents related to genes or proteins. This is a very complex research area that is still in its infancy and we are all aware of the fact the methodologies to solve the full-text mining problematic are still under development. However, precisely because of this, any contribution in this area is more than welcome.

This tool offers several advantages in the area of biomedical text mining: first, SENT is oriented to give researchers a global functional picture of their genes of interest by summarizing the associated literature content in a small set of semantic topics. Second, SENT is able to categorize the list of genes or proteins according to these topics and also associated to Biological Processes terms in GO. Finally this functionality, and the way it is implemented using web-services technology, allows researchers to easily include this analysis into their workflows, providing their research with one more piece of information to be taken into account.

As any other system SENT is not without limitations and we are working in improving both the results and their interpretability. Several ideas gathered from FAUN and GenCLiP are been considered, as well as the possibility of automatically mapping semantic features to biomedical dictionaries or ontologies. We will work to have future versions of SENT updated with these enhancements.

We hope this application is useful to the biomedical community.

ACKNOWLEDGEMENTS

The authors thanks the support of Integromics, S.L. A.P.M. acknowledges the support of the Spanish Ramón y Cajal program.

FUNDING

Spanish grants [BIO2007-67150-C03-02, S-Gen-0166/2006, TIN2005-5619, PS-010000-2008-1]; European Union Grant [FP7-HEALTH-F4-2008-202047]. Funding for open access charge: Spanish Grant number BIO2007-67150-C03-02.

Conflict of interest statement. None declared.

REFERENCES

1. Ashburner,M., Ball,C., Blake,J., Botstein,D., Butler,H., Cherry,J., Davis,A., Dolinski,K., Dwight,S., Eppig,J. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet., **25**, 25–29.
2. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
3. Shatkay,H. and Feldman,R. (2003) Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol.*, **10**, 821–855.
4. Blaschke,C., Andrade,M.A., Ouzounis,C. and Valencia,A. (1999) Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **1999**, 60–67.
5. Jenssen,T.K., Laegreid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.*, **28**, 21–28.
6. 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.
7. Hoffmann,R. and Valencia,A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, **21**, 252–258.
8. Chaussabel,D. and Sher,A. (2002) Mining microarray expression data by literature profiling. *Genome Biol.*, **3**, 1–16.
9. Jelier,R., Jenster,G., Dorssers,L.C.J., Wouters,B.J., Hendriksen,P.J.M., Mons,B., Delwel,R. and Kors,J.A. (2007) Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation. *BMC Bioinformatics*, **8**, 14.
10. Raychaudhuri,S., Schütze,H. and Altman,R.B. (2002) Using text analysis to identify functionally coherent gene groups. *Genome Res.*, **12**, 1582–1590.
11. Huang,Z.X., Tian,H.Y., Hu,Z.F., Zhou,Y.B., Zhao,J. and Yao,K.T. (2008) GenCLiP: a software program for clustering gene lists by literature profiling and constructing gene co-occurrence networks related to custom keywords. *BMC Bioinformatics*, **9**, 308.
12. Frijters,R., Heupers,B., van Beek,P., Bouwhuis,M., van Schaik,R., de Vlieg,J., Polman,J. and Alkema,W. (2008) CoPub: a literature-based keyword enrichment tool for microarray data analysis. *Nucleic Acids Res.*, **36**, W406.
13. Chagoyen,M., Carmona-Saez,P., Shatkay,H., Carazo,J.M. and Pascual-Montano,A. (2006) Discovering semantic features in the literature: a foundation for building functional associations. *BMC Bioinformatics*, **7**, 41.
14. Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.
15. Pehkonen,P., Wong,G. and Toronen,P. (2005) Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC Bioinformatics*, **6**, 16.
16. Heinrich,K.E., Berry,M.W. and Homayouni,R. (2008) Gene tree labeling using nonnegative matrix factorization on biomedical literature. *Comput. Intelligence and Neuroscience*, **2008**, 12.
17. Tjioe,E., Berry,M. and Homayouni,R. (2008) *First Workshop on Data Mining in Functional Genomics, IEEE International Conference on Bioinformatics and Biomedicine*, November 3–5, 2008, Philadelphia, pp. 185–192.
18. Carmona-Saez,P., Chagoyen,M., Tirado,F., Carazo,J.M. and Pascual-Montano,A. (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol.*, **8**, R3.
19. Mitchell,J.A., Aronson,A.R., Mork,J.G., Folk,L.C., Humphrey,S.M. and Ward,J.M. (2003) Gene indexing: characterization and analysis of NLM's GeneRIFs. *AMIA\ldots Annu. Symp. Proc. [electronic resource]*, **2003**, 460.
20. Yeh,A., Morgan,A., Colosimo,M. and Hirschman,L. (2005) BioCreAtIvE task 1A: gene mention finding evaluation. *BMC Bioinformatics*, **6**, 1.
21. Wilbur,J., Smith,L. and Tanabe,L. (2007) Biocreative 2. gene mention task. *Proc. Second BioCreative Challenge Eval. Workshop*, **1**, 7–16.
22. Salton,G., Wong,A. and Yang,C. (1975) A vector space model for automatic indexing. *Commun. ACM*, **18**, 613–620.
23. Porter,M. (1980) An algorithm for suffix stripping. *Program*, **14**, 130–137.
24. Sparck,J.K. (1988). In Willett,P. (ed.) A statistical interpretation of term specificity and its application in retrieval. *Document Retrieval Systems, Taylor Graham Series in Foundations of Information Science*, Vol. 3, Taylor Graham Publishing, London, UK, pp. 132–142.

25. Mejia-Roa,E., Carmona-Saez,P., Nogales,R., Vicente,C., Vazquez,M., Yang,X.Y., Garcia,C., Tirado,F. and Pascual-Montano,A. (2008) bioNMF: a web-based tool for non-negative matrix factorization in biology. *Nucleic Acids Res.*, **36**, W523–W528.
26. Boutsidis,C. and Gallopoulos,E. (2008) SVD based initialization: a head start for nonnegative matrix factorization. *Pattern Recogn.*, **41**, 1350–1362.
27. Wild,S., Curry,J. and Dougherty,A. (2004) Improving non-negative matrix factorizations through structured initialization. *Pattern Recogn.*, **37**, 2217–2232.
28. Deerwester,S., Dumais,S., Furnas,G.W., Landauer,T.K. and Harshman,R. (1990) Indexing by latent semantic analysis. *J. Am. Soc. Inform. Sci.*, **41**, 391–407.
29. Homayouni,R., Heinrich,K., Wei,L. and Berry,M.W. (2005) Gene clustering by latent semantic indexing of MEDLINE abstracts. *Bioinformatics*, **21**, 104–115.
30. D'Arcangelo,G., Homayouni,R., Keshvara,L., Rice,D., Sheldon,M. and Curran,T. (1999) Reelin is a ligand for lipoprotein receptors. *Neuron*, **24**, 471–479.
31. Huang,Z.X., Tian,H.Y., Hu,Z.F., Zhou,Y.B., Zhao,J. and Yao,K.T. (2008) GenCLiP: a software program for clustering gene lists by literature profiling and constructing gene co-occurrence networks related to custom keywords. *BMC Bioinformatics*, **9**, 308.