Efficient functional bioinformatics tools: towards understanding biological processes


Alberto Pascual-Montano[1,2,*], Jose M. Carazo[1]

[1]National Center for Biotechnology. CNB-CSIC, Madrid Spain
[2]Institute for Molecular Medicine Príncipe de Asturias. IMMPA-CSIC, Madrid Spain
*Corresponding author: pascual@cnb.csic.es

Abstract

Experimental high-throughput techniques in biology are generating large amounts of data related to genes and proteins at different levels. The functional analysis of such datasets is a necessary key step in their interpretation. In this contribution we review different methodologies and applications for functional bioinformatics in the context of automated data and text analysis in biology.

Introduction

The post-genomics era has been characterized by the emergence of several high-throughput techniques such as DNA microarrays, protein chips, chip-on-chip, and more recently, the new generation sequencers [1-4], which are allowing the scientific community to study biological processes from a global perspective to understand the basis of development and diseases. Nevertheless, the great potential of these technologies is resulting in a problem when researchers have to deal with the vast amounts of data that are generated by these technologies.

Life sciences in general are generating huge amounts of data with the accelerated development of high-throughput technologies. Therefore, in parallel to the accumulation of the experimental information, there is an obvious need to analyse this huge amount of data. The challenge lies not only in the data-processing area, in which the bioinformatics community has made significant advances, but also in the interpretation of the experimental data. An essential task in this analysis is to translate gene signatures into information that can assist in understanding the underlying biological mechanisms.

During the last few years, the Functional Bioinformatics Group from the National Center for Biotechnology in Madrid (http://bioinfo.cnb.csic.es) has focused on the development of several methods and tools to assist researchers in the analysis and interpretation of genomics and proteomics data. One of our main goals has been to provide the research community with easy-to-access and easy-to-use tools, implementing complex data-analysis methodologies that allow a more complete understanding of cellular processes and diseases. In this way, we have generated a set of web-based applications that addresses three main fields: gene-expression data analysis, functional interpretation of large lists of genes or proteins and automatic knowledge extraction from the biomedical literature.

In our effort to develop applications for a broad number of users able to deal with computing-intensive problems, we have designed these tools incorporating efficient implementations of the algorithms and their variants, most of them using parallel and grid computing. In addition, most of these tools can be accessed via web-services, which allow users to integrate them in their analysis workflows.

Applications

The mission of our research group is the development and implementation of algorithms that facilitate the understanding of biological processes through the application of statistical and data-mining techniques. Figure 1 shows an overview of the applications and databases that we have developed, can be divided into three different groups:

Data analysis

This comprises applications for data clustering and biclustering, as well as data accommodation, normalization and pre-processing. Tools like Engene (http://engene.cnb.csic.es) [5] were designed for microarray data analysis, from preprocessing to clustering and classification. bioNMF (http://bionmf.cnb.csic.es) [6], on the other hand, performs more complex analysis (clustering and biclustering) using specific matrix factorization. Finally, MARQ was designed to retrieve ex-

Figure 1. General overview of the applications available at the Functional Bioinformatics group from the National Center for Biotechnology. A brief explanation of the content of each web-based tool is provided.

perimental datasets from the GEO database [7] that have expression patterns similar or opposite to a given query. This is useful for meta-analysis studies.

Functional analysis

This group is composed of different analysis tools that aim at providing functional interpretation to a list of genes or proteins. Usually, this is a secondary step, after primary analysis of gene or protein expression. Applications for functional analysis in gene or protein sets use annotations from different data sources and calculate statistics to determine their significant presence or absence in the list. The GeneCodis tool searches for significant concurrent annotations in a list of genes (http://genecodis.cnb.csic.es [8]). Another application, ChipCodis (http://chipcodis.cnb.csic.es [9]), mines regulatory systems in yeast using chip-on-chip data using a similar approach. Text mining tools, like SENT (http://sent.cnb.csic.es [10]) and Moara (http://moara.cnb.csic.es [11]), also belongs to this category, as the final goal is to provide functional interpretation of gene sets using the scientific literature as the main source of information.

Databases

Databases that collect and centralize interesting unique information about biological events are also a central part of our research interest. Centrosome:db (http://centrosomedb.cnb.csic.es [12]) is a database that contains a set of human genes encoding proteins that are localized in the centrosome. It offers different perspectives to study the human centrosome, including evolu-

tion, function, and structure. The ProteoPathogen [13] and Compluyeast databases were designed to handle and manage proteomic studies.

In this report, we will focus our attention on four representative tools taken from data-analysis and functional-analysis categories. These tools are widely used by the scientific community, handling several thousand submissions every month:

GeneCodis

Gene Annotation Co-occurrence Discovery

An essential task in functional analysis of genes and proteins is the translation of gene signatures into information that can assist in understanding the underlying biological mechanisms. Several methods and tools have been developed to interpret large lists of genes or proteins using information available in biological databases. The common idea used in most of these methods is to find functional descriptors that are significantly enriched in the gene signature.

The first type of techniques that emerged in this field were focused on evaluating the frequency of individual annotations and apply a statistical test to determine which annotations are significantly enriched in a input list with respect to a reference list, usually the whole genome or all genes in a microarray. Annotations from different sources like the Gene Ontology (GO) [14] or KEGG [15] are commonly used in this context. Several tools have been developed following this approach [16,17]. A fresh line of research appeared with the observation that the use of thresholds to select the significant genes could underestimate the effect of significant biological effects during the functional analysis. This idea derived in a new and different analytical concept in which the distribution of annotations is evaluated over the whole list of genes, sorted by their correlation with the phenotype: the gene set enrichment analysis (GSEA) [18].

Nevertheless, both approaches, the standard enrichment analysis and GSEA methods, evaluate each annotation independently from the others without taking into account the potential relationships among them. However, most of the annotations in biological databases are interconnected because they are associated to common genes. Patterns that contain these relationships can provide invaluable information and extend our understanding of biological events associated with the experimental system. It is therefore highly desirable to evaluate these associations in the functional analysis of gene lists [19].

In 2007, we introduced GeneCodis [20], a tool for modular enrichment analysis oriented to integrate information from different sources and find enriched combinations of annotations in large lists of genes or proteins. Modular enrichment represents a step forward in the functional analysis because of its capacity to integrate heterogeneous annotations and to discover significant combinations among them. A new version of this tool has recently been reported [8].

The application is simple in its concept: it takes a list of genes as input, and determines biological annotations or combinations of annotations that are over-represented with respect to a reference list. The novelty of this tool relies on the fact that, before computing the statistical test, it incorporates new functionality to extract all combinations of annotations that appear in at least x genes, x being a user-defined threshold [20]. To extract combinations of gene annotations, GeneCodis uses a modification of the methodology reported in [21], which implements an efficient variant of the apriori algorithm to extract associations among gene annotations and expression patterns.

Once all combinations of annotations that appear in at least x genes have been extracted, the method counts the occurrence of each set of annotations in the list of genes and in a reference list. Then, a statistical test is applied to identify categories, and their combinations, that are significantly enriched in the list of genes. Therefore, the result of the analysis performed by GeneCodis consists of a list of annotations or combinations of annotations with their corresponding p-values. Annotations showing p-values below a certain threshold can be considered significantly associated with the list of genes under study and can be used to discern the biological mechanisms relevant to the experimental system.

GeneCodis works with most of the model organisms in biological research. The whole list includes *Arabidopsis thaliana*, *Bos taurus*, *Caenorhabditis elegans*, *Candida albicans*, *Danio rerio*, *Drosophila melanogaster*, *Escherichia coli*, *Gallus gallus*, *Homo sapiens*, *Leishmania major*, *Mus musculus*, *Rattus norvegicus*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and

Vibrio cholerae. Alternatively, to facilitate the usability of the application we have extended the type of gene identifiers that are supported, including proprietary identifiers from commercial platforms such as Affymetrix, Agilent, Codelink and Illumina. Regarding annotations, GeneCodis integrates functional and regulatory information by supporting Gene Ontology (GO), KEGG Pathways, micro RNA, transcription factors and InterPro motifs.

This tool can be freely accessed through its main site at http://genecodis.cnb.csic.es. A mirror has also been recently set up at the China Node for EMBnet at the Center for Bioinformatics in Peking University, available at http://genecodis.cbi.pku.edu.cn.

bioNMF

Non-negative Matrix Factorization in Biology

Matrix factorization techniques have become well established methods for the analysis of numerical data and signals. These methods can be applied to the analysis of multidimensional datasets in order to reduce the dimensionality, discover patterns and aid in the interpretation of the data. Among the most popular, Principal Component Analysis (PCA), Singular Value Decomposition (SVD) or Independent Component Analysis (ICA) have been successfully used in a broad range of contexts.

In 1999, Lee and Seung, developed a new matrix factorization technique named Non-Negative Matrix Factorization (NMF) [22] to decompose images into recognizable features. The main difference between NMF and other classical factorization methods relies on the non-negativity constraints imposed by the model. These constraints tend to lead to a parts-based representation of the data, because they allow only additive, not subtractive, combinations of data items. In this way, the factors produced by this method can be interpreted as parts of the data or, in other words, as subsets of elements that tend to occur together in sub-portions of the dataset. By contrast, classical factorization techniques decompose the data matrix into a new set of matrices of any sign, involving complex cancellations of positive and negative elements to reconstruct the original dataset. Therefore, the interpretation of the factors becomes non-intuitive and difficult. The comprehensible properties of the NMF method and the intuitivism of the results it provides have centered the attention of many researches in different fields of science and, in particular, in the bioinformatics field, where NMF has been applied in different contexts.

Formally, the non-negative matrix decomposition can be described as follow:

$$V \approx WH$$

where $V \in \mathbb{R}^{m \times n}$ is a positive data matrix with n variables and m objects, $W \in \mathbb{R}^{m \times k}$ are the reduced k basis vectors or factors, and $H \in \mathbb{R}^{k \times n}$ contains the coefficients of the linear combinations of the basis vectors needed to reconstruct the original data (also known as encoding vectors). The main difference between NMF and other classical factorization models relies on the non-negativity constraints imposed on both the basis (W) and encoding vectors (H). In this way, only additive combinations are possible:

$$(V)_{mn} \approx (WH)_{mn} = \sum_{l=1}^{k} W_{ml} H_{ln}$$

Increasing interest in this factorization technique is due to the intuitive nature of the method, which has the ability to extract additive parts of data sets that are highly interpretable, while reducing the dimensionality of the data at the same time. In the biomedical field, NMF-related methods have recently gained a lot of popularity. For example, gene expression analysis [23-31], analysis of protein sequences [32], functional categorization of genes [33] or biological text mining [34].

We have developed a user-friendly, web-based tool that implements this factorization technique and its application in data biclustering (clustering by rows and columns simultaneously) and sample classification [6,35]. In addition, this web tool offers new improvements to process big data-sets in a distributed computing environment without the usage complexity present on this kind of systems. It also provides an automated access to external applications through a web-services interface.

This online tool provides a user-friendly interface, combined with a computationally efficient parallel implementation of the NMF methods to explore the data in different analysis scenarios. In addition to the online access, bioNMF also provides the same functionality included in the web

site as a public web-services interface, enabling users with more computer expertise to launch jobs on bioNMF server from their own scripts and workflows. The bioNMF application is freely available at http://bionmf.cnb.csic.es.

SENT

Semantic Features in Texts

Applications like GeneCodis described above are examples of initiatives to help in the biological interpretation of lists of genes and proteins in the context of certain experimental conditions. Annotations-based approaches provide a fast, easy, and statistically sound interpretation of a list of genes or their products. Although this information is extremely useful, 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 no limit to the types of information that may be recovered through careful and exhaustive mining [36]. Many researchers have focused their attention on the use of text mining, with methodologies that go from determining protein-protein interactions from biomedical texts [37-40], to providing summary descriptions for genes or determining their similarities [41-45]. Even though lots of work in this area has been reported, its practical use by the scientific community is hindered by the lack of efficient and easy-to-use software.

SENT is a usable implementation of a methodology based on [34]: for a given set of genes or proteins, the associated abstracts from PubMed are extracted and processed. Then the NMF technique (like the one implemented in the bioNMF application described above) is used to cluster genes together with relevant terms. In this way, a set of semantic features (topics or collection of semantically related words) are associated to a set of genes. In other words, genes are clustered according to the content of their associated literature and, at the same time, those semantic topics provide insights into their functional implications.

The input of the system is a set of gene or protein identifiers and the number of semantic features (factors or topics) to extract in the analysis. Titles and abstracts from articles associated with 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. Coupled with the Gene Ontology enrichment analysis (using GeneCodis), also provided in the web site, SENT serves as a guide in the examination of the literature.

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 on a small set of semantic topics. Second, SENT is able to categorize the list of genes or proteins according to these topics, and also to associate them with Biological Process terms in the Gene Ontology. Finally this functionality, and the way it is implemented using web-service technology, allows researchers to easily include this analysis into their workflows, providing their research with one more pieces of information to be taken into account. SENT is available at http://sent.cnb.csic.es.

MARQ

A tool to mine GEO by content

DNA microarrays have become an extensively used technique to analyze global gene-expression profiles. Its widespread usage quickly promoted the creation of gene-expression data repositories such as the NCBI Gene Expression Omnibus (GEO) [7], which currently holds more than 10,000 experiments for over 500 organisms. Despite the huge amounts of information available in GEO, researchers usually focus their analysis on individual experiments without exploiting such valuable resources. Gene-expression analyses are usually performed from a gene-centered perspective, with the goal of identifying relevant sets of genes that are then used to determine the biological mechanisms relevant in that experimental setup. However, besides this gene-

to-phenotype approach, gene signatures can also be used to establish connections among different physiological conditions [46]. For example, we can discover connections among diseases, drugs or pathways by similarities in their gene-expression signatures; conditions inducing similar signatures might be modulating the same pathways, while conditions with opposite signatures might be involved in reversing the original phenotype.

MARQ compiles a gene-expression signature database from all datasets in GEO for five model organisms (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*). For each dataset, a number of signatures are extracted, each representing a potentially relevant comparison between the dataset samples. The query consists of two sets of genes, up-regulated and down-regulated (optionally one may be empty). A rank statistic is used to give a score and significance p-value to the relationship of each signature to the query based on how the input genes rank in differential expression for that signature. Signatures with a direct relationship receive a positive score and signatures with an inverse relationship receive a negative score. In this way, datasets, similar in direct or opposite relationship with the content of the query gene set are successfully extracted by MARQ.

There are two types of possible queries: platform based, and organism based (cross-platform) depending on the preference for searching. Additionally, each signature is annotated with additional meta-information, such as words in the dataset description or GO terms over-represented in the list of genes deregulated in the comparison. An annotation-mining tool can be used to find patterns in these annotations to highlight relevant features common to the most significant signatures.

Finally, a set of signatures of interest can be analyzed using a meta-analysis methodology based in Rank Products [47] which highlight genes commonly deregulated in those signatures. The application is freely accessible at http://marq.cnb.csic.es.

Conclusions

Biology is becoming more and more a science that necessarily needs to include information technology as a basic discipline in order to understand the extremely complex biological processes of living organisms. Tools and methodologies able to efficiently process the huge amounts of data that are generated in this field are still demanded by the molecular biology community. In this report, we have described a set of bioinformatics tools publicly available online through web browsers or web-service interfaces. The applications are oriented to the efficient and robust analysis of gene-expression information and the functional characterization of lists of genes or proteins from different perspectives. Integrative analysis of all available data and biological information, as the ones described here, is probably one of the best ways to move towards a complete study of biology from a global perspective.

Acknowledgments

This work has been partially funded by the Spanish grants BIO2007-67150-C03-02, S-Gen-0166/2006, PS-010000-2008-1 and European Union Grant FP7-HEALTH-F4-2008-202047.

References

1. Mardis, E.R. (2009) New strategies and emerging technologies for massively parallel sequencing: applications in medical research. Genome Med, 1, 40.
2. Ansorge, W.J. (2009) Next-generation DNA sequencing techniques. N Biotechnol, 25, 195-203.
3. Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K. and Wang, J. (2009) SNP detection for massively parallel whole-genome resequencing. Genome Res, 19, 1124-32.
4. Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q., Huang, T., Dong, G., Sang, T. and Han, B. (2009) High-throughput genotyping by whole-genome resequencing. Genome Res, 19, 1068-76.
5. Garcia de la Nava, J., Santaella, D.F., Cuenca Alba, J., Maria Carazo, J., Trelles, O. and Pascual-Montano, A. (2003) Engene: the processing and exploratory analysis of gene expression data. Bioinformatics, 19, 657-8.
6. 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 nonnegative matrix factorization in biology. Nucleic Acids Res, 36, W523-8.

7. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Muertter, R.N. and Edgar, R. (2009) NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res, 37, D885-90.
8. Nogales-Cadenas, R., Carmona-Saez, P., Vazquez, M., Vicente, C., Yang, X., Tirado, F., Carazo, J.M. and Pascual-Montano, A. (2009) GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. Nucleic Acids Res, 37, W317-22.
9. Abascal, F., Carmona-Saez, P., Carazo, J.M. and Pascual-Montano, A. (2008) ChIPCodis: mining complex regulatory systems in yeast by concurrent enrichment analysis of chip-on-chip data. Bioinformatics, 24, 1208-9.
10. Vazquez, M., Carmona-Saez, P., Nogales-Cadenas, R., Chagoyen, M., Tirado, F., Carazo, J.M. and Pascual-Montano, A. (2009) SENT: semantic features in text. Nucleic Acids Res, 37, W153-9.
11. Neves, M.L., Carazo, J.M. and Pascual-Montano, A. (2010) Moara: a Java library for extracting and normalizing gene and protein mentions. BMC Bioinformatics, 11, 157.
12. Nogales-Cadenas, R., Abascal, F., Diez-Perez, J., Carazo, J.M. and Pascual-Montano, A. (2009) CentrosomeDB: a human centrosomal proteins database. Nucleic Acids Res, 37, D175-80.
13. Vialas, V., Nogales-Cadenas, R., Nombela, C., Pascual-Montano, A. and Gil, C. (2009) Proteopathogen, a protein database for studying Candida albicans--host interaction. Proteomics, 9, 4664-8.
14. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 25, 25-9.
15. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y. (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res, 36, D480-4.
16. Khatri, P. and Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics, 21, 3587-95.
17. Dopazo, J. (2006) Functional interpretation of microarray experiments. OMICS, 10, 398-410.
18. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 102, 15545-50.
19. 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.
20. 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.
21. Carmona-Saez, P., Chagoyen, M., Rodriguez, A., Trelles, O., Carazo, J.M. and Pascual-Montano, A. (2006) Integrated analysis of gene expression by Association Rules Discovery. BMC Bioinformatics, 7, 54.
22. Lee, D.D. and Seung, H.S. (1999) Learning the parts of objects by non-negative matrix factorization. Nature, 401, 788-91.
23. Brunet, J.P., Tamayo, P., Golub, T.R. and Mesirov, J.P. (2004) Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A, 101, 4164-9.
24. Carmona-Saez, P., Pascual-Marqui, R.D., Tirado, F., Carazo, J.M. and Pascual-Montano, A. (2006) Biclustering of gene expression data by non-smooth non-negative matrix factorization. BMC Bioinformatics, 7, 78.
25. Carrasco, D.R., Tonon, G., Huang, Y., Zhang, Y., Sinha, R., Feng, B., Stewart, J.P., Zhan, F., Khatry, D., Protopopova, M., Protopopov, A., Sukhdeo, K., Hanamura, I., Stephens, O., Barlogie, B., Anderson, K.C., Chin, L., Shaughnessy, J.D., Jr., Brennan, C. and Depinho, R.A. (2006) High-resolution genomic profiles define distinct clinico-pathogenetic subgroups of multiple myeloma patients. Cancer Cell, 9, 313-25.
26. Wang, G., Kossenkov, A.V. and Ochs, M.F. (2006) LS-NMF: a modified non-negative matrix factorization algorithm utilizing uncertainty estimates. BMC Bioinformatics, 7, 175.
27. Kim, P.M. and Tidor, B. (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res, 13, 1706-18.

28. Gao, Y. and Church, G. (2005) Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics, 21, 3970-5.
29. Inamura, K., Fujiwara, T., Hoshida, Y., Isagawa, T., Jones, M.H., Virtanen, C., Shimane, M., Satoh, Y., Okumura, S., Nakagawa, K., Tsuchiya, E., Ishikawa, S., Aburatani, H., Nomura, H. and Ishikawa, Y. (2005) Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization. Oncogene, 24, 7105-13.
30. Kim, H. and Park, H. (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics, 23, 1495-502.
31. Han, X. (2007) Cancer molecular pattern discovery by subspace consensus kernel classification. Comput Syst Bioinformatics Conf, 6, 55-65.
32. Heger, A. and Holm, L. (2003) Sensitive pattern discovery with 'fuzzy' alignments of distantly related proteins. Bioinformatics, 19 Suppl 1, i130-7.
33. Pehkonen, P., Wong, G. and Toronen, P. (2005) Theme discovery from gene lists for identification and viewing of multiple functional groups. BMC Bioinformatics, 6, 162.
34. 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.
35. Pascual-Montano, A., Carazo, J.M., Kochi, K., Lehmann, D. and Pascual-Marqui, R.D. (2006) Nonsmooth nonnegative matrix factorization (nsNMF). IEEE Trans Pattern Anal Mach Intell, 28, 403-15.
36. Shatkay, H. and Feldman, R. (2003) Mining the biomedical literature in the genomic era: An overview. Journal of Computational Biology, 10, 821-855.
37. 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.
38. Jenssen, T.K., L\aegreid, A., Komorowski, J. and Hovig, E. (2001) A literature network of human genes for high-throughput analysis of gene expression. Nature Genetics, 28, 21-28.
39. 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.
40. Hoffmann, R. and Valencia, A. (2005) Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics, 21.
41. Chaussabel, D. and Sher, A. (2002) Mining microarray expression data by literature profiling. Genome Biology, 3, 1-0055.
42. Jelier, R., Jenster, G., Dorssers, L.C.J., van der Eijk, C.C., van Mulligen, E.M., Mons, B. and Kors, J.A. (2005) Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. Bioinformatics, 21, 2049-2058.
43. Raychaudhuri, S., Schütze, H. and Altman, R.B. (2002) Using Text Analysis to Identify Functionally Coherent Gene Groups. Genome Research, 12, 1582-1590.
44. Huang, W. and Marth, G. (2008) EagleView: a genome assembly viewer for next-generation sequencing technologies. Genome Res, 18, 1538-43.
45. 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 Research, 36, W406.
46. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.P., Subramanian, A., Ross, K.N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S.A., Haggarty, S.J., Clemons, P.A., Wei, R., Carr, S.A., Lander, E.S. and Golub, T.R. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 313, 1929-35.
47. Girolami, M. and Breitling, R. (2004) Biologically valid linear factor models of gene expression. Bioinformatics, 20, 3021-33.