Published online 30 May 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W523–W528***
doi:10.1093/nar/gkn335

bioNMF: a web-based tool for nonnegative matrix factorization in biology

E. Mejía-Roa[1], P. Carmona-Saez[2], R. Nogales[1], C. Vicente[1], M. Vázquez[3], X. Y. Yang[1], C. García[1], F. Tirado[1] and A. Pascual-Montano[1,*]

[1]Computer Architecture Department, Complutense University, 28040, [2]Integromics, S.L. Calle Darwin 3. 28049 and [3]Software Engineering Department, Complutense University, 28040 Madrid, Spain

Received January 30, 2008; Revised April 18, 2008; Accepted May 10, 2008

ABSTRACT

In the last few years, advances in high-throughput technologies are generating large amounts of biological data that require analysis and interpretation. Nonnegative matrix factorization (NMF) has been established as a very effective method to reveal information about the complex latent relationships in experimental data sets. Using this method as part of the exploratory data analysis, workflow would certainly help in the process of interpreting and understanding the complex biology mechanisms that are underlying experimental data. We have developed bioNMF, a web-based tool that implements the NMF methodology in different analysis contexts to support some of the most important reported applications in biology. This online tool provides a user-friendly interface, combined with a computational 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 website as a public web services interface, enabling users with more computer expertise to launch jobs into bioNMF server from their own scripts and workflows. bioNMF application is freely available at http://bionmf.dacya.ucm.es.

INTRODUCTION

The analysis of complex data sets generated by *–omics* technologies requires the use of statistical and data mining techniques able to find natural group structures in the data. Different data mining methods have shown to be very useful in providing significant information for hypothesis formulation and discovery of biological patterns. Clustering algorithms or matrix factorization techniques, such as PCA or SVD, are among the most popular tools for the exploratory analysis of high-dimensional biological datasets.

Nonnegative matrix factorization (NMF) (1) is one of such techniques that, although relatively new, is increasingly used in biomedical sciences. It has gained a lot of popularity in the scientific community due to its capability of providing new insights and relevant information about the complex latent relationships in high-dimensional biological data sets. In the particular case of biomedical sciences, several successful studies have been conducted using this method and some of its variants. For example, NMF has been successfully applied to gene-expression analysis (2–5), scientific literature mining (6,7), proteomics, metabolomics (8,9), sequence analysis (10) or neurosciences (11), among others.

Due to the increasing interest on this technique by the bioinformatics community, several standalone applications and code in different programming languages have been developed to support NMF analysis and related alternatives, in special for the biomedical field. In 2006, we introduce one of such standalone applications, *bioNMF* (12), which implements the NMF methodology in different analysis contexts to support some of the most popular applications of this new methodology. This includes clustering and biclustering of gene-expression data and sample classification.

Even if standalone applications play their role in the research process, online web tools are clearly the preferred option for most of the users, because no resources and computational expertise are required to run a scientific analysis.

In this work, we propose a user-friendly, web-based tool that implements the same methodologies present on the previous standalone application (12). 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.

*To whom correspondence should be addressed. Tel: +34 913944420; Fax: +34 913944687; Email: pascual@fis.ucm.es

To the best of our knowledge, this is the first web-based dedicated application for NMF. This new tool is freely available at http://bionmf.dacya.ucm.es/.

FEATURES AND FUNCTIONALITY

Experimental biological information, like for example gene expression, is usually represented and stored as a numerical data matrix, where observations or genes are stored in rows and conditions, experiments or samples are represented in columns. In this case, each cell corresponds to the expression value of a gene in a specific experimental condition.

Formally, the nonnegative matrix decomposition can be described as $\mathbf{V} \approx \mathbf{WH}$, where $\mathbf{V} \in \mathbb{R}^{m \times n}$ is a positive data matrix with m variables and n objects, $\mathbf{W} \in \mathbb{R}^{m \times k}$ are the reduced k basis vectors or factors and $\mathbf{H} \in \mathbb{R}^{k \times n}$ contains the coefficients of the linear combinations of the basis vectors needed to reconstruct the original data. The number of factors (k) is generally chosen so that it takes a value less than n and m. The distinctive attributes of NMF with respect to other factorization models are the nonnegativity constraints imposed on **V**, **W** and **H**. In this way, only additive combinations of **W** and **H** are possible, which induces not only an effective dimensionality reduction but also a more interpretable information (1). Figure 1 shows a graphic representation of the model in the case of gene-expression data.

The bioNMF online tool provides a functionality to cover some of the most important applications of the NMF algorithm. More particularly in biology (2,3,6,7,10,13). This is achieved through three different modules: *Sample Classification*, *Standard NMF* and *Biclustering Analysis*.

The *Sample Classification* module implements the method proposed by Brunet *et al.* (2) to determine the most suitable number of sample clusters in a given data set and to group the data samples into k clusters, being k the best factorization rank within a given input range. This method is probably one of the most used methods in the field to estimate the best factorization rank.

Standard NMF

This module performs the classical NMF factorization using the algorithm proposed by Lee and Seung in 1999 (1). This wide-ranging module is not specifically focused to any particular analysis, but more generally oriented to any potential application that might use this factorization method. As a new feature with respect to the previous standalone application (12), this module now integrates the consensus clustering methodology described above to determine the best rank of factorization in a given range. This saves the need of launching the *Standard NMF* analysis (and therefore, uploading the data matrix) several times, or running the *Sample Classification* process as a previous step.

Finally, the *Biclustering Analysis* module implements a two-way clustering method to identify gene–experiment relationships. bioNMF estimates biclusters using a method based on a modified variant of the NMF algorithm, which produces a suitable decomposition as a product of three matrices that are constrained to have nonnegative elements. This variant, denoted as *Nonsmooth Nonnegative Matrix Factorization* (nsNMF) (14), produces a sparse representation of the gene-expression data matrix, making possible the simultaneous clustering of genes and conditions that are highly related in subportions of the data (3). This module also incorporates the consensus clustering methodology to determine the best rank of factorization.

SOFTWARE USAGE

bioNMF has been designed as a web-based tool that mimics its standalone predecessor application (12). In all analysis methods, the original matrix is decomposed in two new nonnegative matrices that encode the latent information embedded in the original input data. In addition, visualizations used in this application also help in the interpretation of the results.

The full process is carried out in three very simple steps:

Data set selection

The input data is a single standard tab-delimited text file that contains the data matrix with, or without labels, for example, a gene-expression matrix. In addition, if an email address is provided, the user will be notified when the analysis is finished. This feature is very useful when submitting large data sets or analysis that might take long time to process.

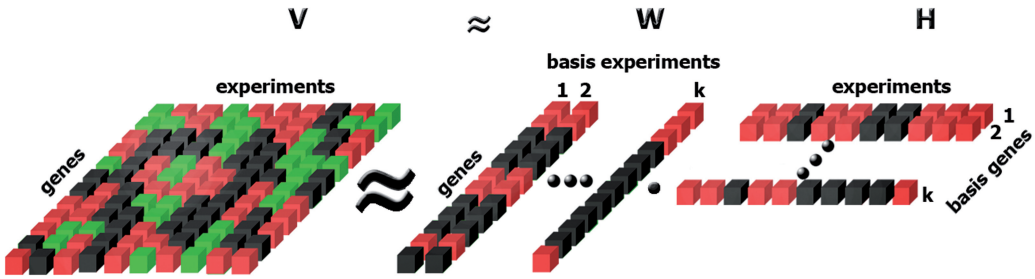

Figure 1. Schematic representation of the NMF model applied to gene-expression data. Input data matrix **V** is represented as a gene–experiment matrix and it is decomposed by the product of two new nonnegative matrices **W** and **H**. The k columns of **W**, therefore, will have the dimension of a single array (genes) and are known as basis experiments or factors. The columns of **H** are known as encoding vectors and are in one-to-one correspondence with a single experiment of the gene-expression matrix. Consequently, each row of **H** has the dimension of a single gene and it is denoted as basis gene.

Data preprocessing

Before the analysis, the data matrix can be transposed, normalized and/or transformed by several methods to satisfy the nonnegative constraints required by the NMF algorithm. Normalization methods include data centering, standardization of rows and columns (independently or simultaneously), mean subtraction by rows and columns and the normalization method proposed by Getz *et al.* (15) that first divides each column by its mean and then normalizes each row. On the other hand, transformation methods to make data positive include subtracting the absolute minimum, the exponential scaling and two data folding methods proposed by Kim and Tidor (13). These folding methods duplicate each row or column in which the first occurrence indicates positive expressions and the second indicates negative values.

Data analysis

As described in the previous section, three different types of analysis are provided in *bioNMF* to cover some of the most important applications of this methodology: (i) *Sample Classification*; (ii) *Standard NMF* and (iii) *Biclustering Analysis*.

The sample classification module implements the methodology described by Brunet *et al.* (2). This methodology uses NMF and a model selection algorithm to determine the most suitable number of sample clusters in a given data set. This sample classification model is based on a reduced set of metagenes, and it has been proved to provide a more accurate and robust classification with respect to the classification based on the high-dimensional gene space. Results will be an estimation of the best number of clusters in the data set and the cluster assignments of each experimental condition. According to (2) the proper factorization, rank should be selected where the magnitude of the cophenetic correlation coefficient begins to fall. A graphical representation of the cophenetic correlation coefficient and the ordered consensus matrix as described in ref. (2) are also provided. Figure 2 provides a snapshot of the results of this step when applied to the acute myelogenous leukemia (AML) and acute lymphoblastic leukemia (ALL) data set (17).

The ordered consensus matrix, in conjunction with the cophenetic correlation coefficient provided in this step,

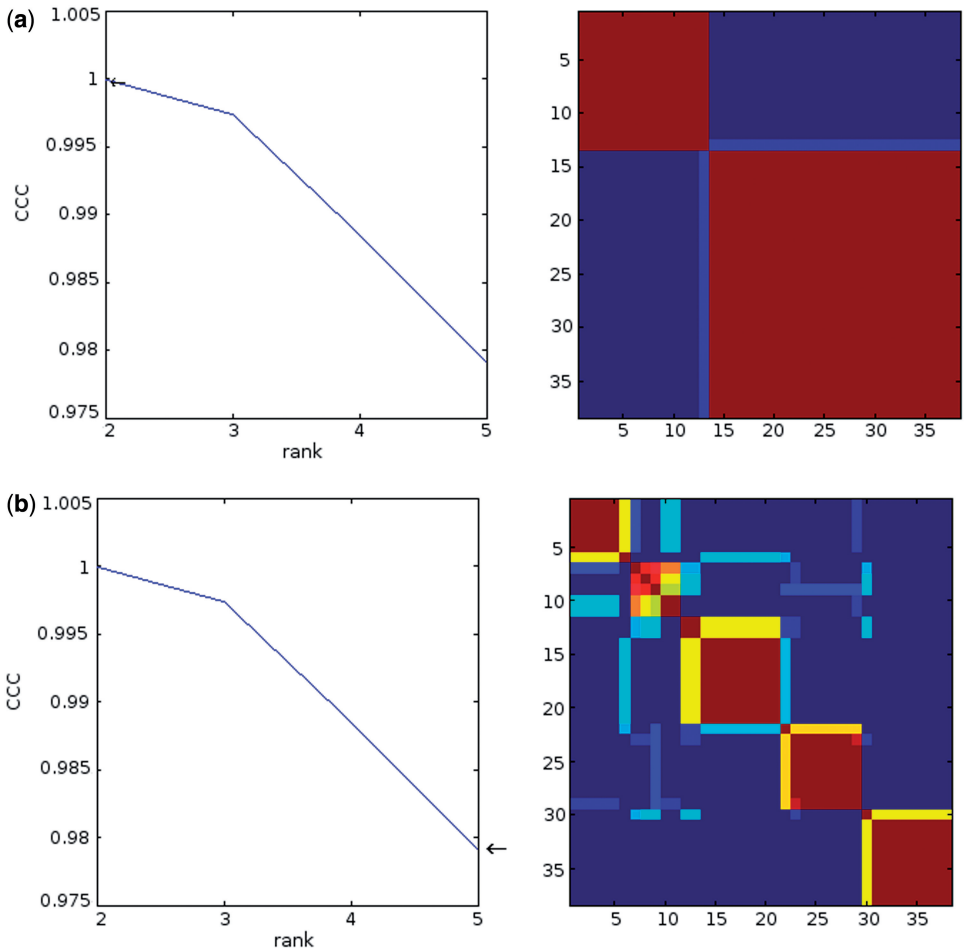

Figure 2. Snapshot of the output of the Sample Classification module. Results show the cophenetic correlation coefficient (left) for different values of k and the reordered consensus matrices (right) calculated for the AML–ALL data set. (**A**) The consensus matrix pattern for $k = 2$ indicates a stable classification into two samples (most of the values are either 0 or 1 represented in red and blue colors in the picture). This is the expected clustering pattern in this two-class data set. (**B**) Consensus matrix for $k = 5$ showing a scattered pattern that indicates a more unstable classification in five classes.

gives an appropriate idea of the stability of the factorization for a given k. Since the NMF algorithm is nondeterministic, its solutions might vary from run to run when executed with different random initial values for **W** and **H**. The rationale behind the model selection approach proposed in ref. (2) is based on the fact that if the factorization is stable for a given value of k, we would expect that column's assignments to those k factors would vary little from run to run.

For each run, the column assignment is defined by a connectivity matrix **C** of size $n \times n$ (where n is the number of columns). Each entry Cij in this matrix equals 1, if column i and j have their maximum for the same factor, and $Cij = 0$ if they do not. Consensus matrix is then defined as the average connectivity matrix over many factorization runs with different initial random conditions. The entries range from 0 to 1 and reflect the level of reproducibility of the columns' assignments. If a factorization is stable **C** will tend not to vary among runs, and thus entries will be close to 0 or 1. Dispersion between 0 and 1 will indicate a lack of reproducibility of the columns' assignments along the different runs. **C** matrix is then reordered to reflect the column's similarity. The more scattered the reordered matrix is, the less stable solution it reflects, since it will indicate that columns that were assigned to the same factor in one run, are probably assigned to different factors in another.

The standard NMF module performs the classical NMF factorization to the input data matrix. The tool returns the **W** and **H** matrices resulting in the factorization. It is also possible to run NMF different times using random initial conditions each time. Results can be independently saved or combined for further analysis as described in ref. (6). NMF is nondeterministic and therefore it may or may not converge to the same solution on each run depending on the random initial conditions. Therefore, executing the algorithm several times with different random initializations is a good approach for selecting the **W** and **H** that best approximates the input matrix **V**. Depending on the problem, less or more runs will be necessary to achieve an optimum solution. However, considering that the computational cost of this algorithm is very high a limited number of runs is recommended. On our own experience a value of 100 runs is normally enough to achieve reasonable results (3). This flexibility makes this unit a general instrument for any potential application in life sciences.

The biclustering algorithm module implements the methodology described in ref. (3). It is intended mainly for gene-expression analysis, although its applications can be extended to other type of data. Taking gene expression as a case study, this analysis group genes and samples based on local features generating sets of samples and genes that are locally related. Results are a set of biclusters (submatrices) encoding modular patterns. Each bicluster matrix contains the set of genes that are highly associated to a local pattern and samples sorted by its importance in this pattern. An image of the heatmap of each bicluster is generated. As an example, Figure 3 depicts a bicluster obtained from a data set containing the expression profiles of 46 soft-tissue tumor samples reported in (16).

WEB SERVICES

In addition to the online access, bioNMF provides the same functionality included in the website as web services. Web services are a public programmatic API that enables users with more computer expertise to launch jobs into bioNMF server from their own programs, scripts and workflows.

The web services provided in bioNMF are built on open standards such as SOAP (*Simple Object Access Protocol*, a messaging protocol for transporting information; see http://www.w3.org/TR/soap/) and WSDL (*Web Services Description Language*, an XML format for describing web service capabilities and provided methods; see http://www.w3.org/TR/wsdl). The WDSL file describing bioNMF methods can be accessed at http://bionmf.dacya.ucm.es/WebService/BioNMFWS.wsdl.

The system allows the upload of matrices, and performs any of the three analyses described in the previous section. The web service works in a nonblocking way. The user launches the analysis and gets a job identifier as results. By using this job identifier, it is possible to poll the status

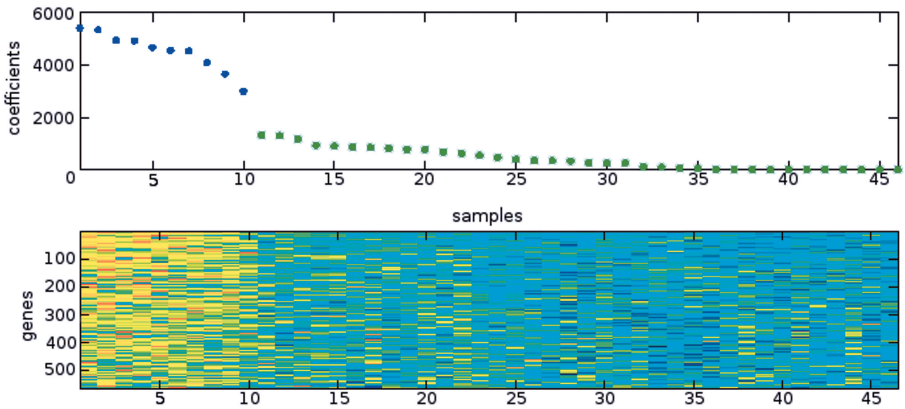

Figure 3. A Heatmap showing the subset of genes and samples in the bicluster. All samples are shown sorted by it association to the bicluster (local pattern). The plot on the upper part of the image represents the coefficients of all samples in the corresponding row of **H**. In blue are the samples that show the largest coefficient for that factor while in green are those samples associated to others.

of the job. Once the job is finished, the results can be retrieved using another web service function. The results for each operation include the essential information; however, all the analysis methods produce a set of files that provide other information, such as visualization images.

Information about all supported web services can be accessed from the web page at: http://bionmf.dacya.ucm.es/webservices.html. Full examples of a client program, implemented in Perl and Ruby are also provided.

IMPLEMENTATION

This online tool has been designed to process big data sets in a multiprocessor environment. A batch-queuing manager controls most systems of this kind. In order to make this tool easy to use, bioNMF consists of two main components: the user interface (web or web services), that handles the batch-queuing system, and the analysis software executed by that system in the multiprocessor environment.

The web interface is implemented in PHP (*PHP: Hypertext Preprocessor*; see http://php.net/), a programming language for generating dynamic web pages. When an analysis is started, the web server submits a job to the batch-queuing system and shows a status page. This page is refreshed periodically until an independent process produces the results page when the job is finished. This system of two processes allows the user to close the browser at the status page stage without losing the results. An email with an URL to the results is optionally sent to the user when the process is finalized.

The analysis software is currently implemented in two layers. The external layer makes use of Matlab software (www.mathworks.com) to apply the preprocessing methods to the input data and to generate the graphical visualizations. On the other hand, the core of the system, explained below, is implemented in C language.

NMF-parallelization

NMF is a very computing intensive technique. With this web application, we also provide an extremely efficient implementation of the NMF algorithm. All of the methods implemented in bioNMF use *parallel computing*. This technique is based on the principle that large problems can be divided into smaller ones, which may be solved in parallel (i.e. simultaneously) in multiple processors. This permits taking advantage of multicore CPUs and computing clusters environments.

The parallelization is focused on the most demanding part of the *bioNMF* analysis methods: the NMF algorithm. As it is based on matrix–matrix operations, our approach divides each matrix into a set of sub-matrices. Operations between these sub-matrices are computed in parallel. This is the case of **W** and **H** matrices, which are broken into smaller pieces and distributed among the processors. Each sub-matrix operation is then performed simultaneously. When necessary, the results are gathered in order to synchronize the updated matrices. This is done using MPI (*Message Passing Interface*; see http://www.mpi-forum.org/) that provides a low overhead communication mechanism.

This implementation represents a very cost-effective alternative to improve the throughput of this web-based application where simultaneous user's requests are going to be handled. Currently, a dedicated eight-node cluster is supporting this application.

APPLICATION PERFORMANCE

As an example of performance, as well as of validation, a test of *Sample Classification* method was made by comparing the Matlab algorithm reported by Brunet *et al.* (2), available at www.broad.mit.edu/cancer/pub/nmf/with the online bioNMF analysis module. Both algorithms were tested with the AML and ALL data set (17), which is a 5000-gene by 38-sample data matrix. bioNMF's results are very close to those obtained in. Brunet *et al.*'s (2) algorithm took 2102 s (about 35 min) to complete the analysis in a single AMD Opteron processor, while bioNMF finished in 310 s (5 min, 10 s) in an Opteron eight-processor system. This effective implementation using parallel computing represents a suitable approach to reduce the long computing times required by bigger data sets. Better speedups can also be obtained if computer clusters with a larger number of nodes are used. The current implementation allows this upgrade in a transparent manner.

DISCUSSION AND CONCLUSIONS

In the era of *–omics* technologies, the use of sophisticated statistical and data mining methods has become an essential task in many molecular biology laboratories. Web-based tools offer the opportunity of use complex data analysis methods in a friendly environment that quickly bring to many potential users new methodological developments. These types of applications are helping researchers in the exploration and analysis of large volume of data.

In this work, we present a web-based implementation of the NMF algorithm. This technique is a matrix factorization method that is increasingly used in many fields, such as image analysis, proteomics, metabolomics or genomics. This tool provides an efficient implementation of the NMF and different analysis pipelines based on this algorithm: standard NMF, biclustering analysis and sample classification. The architecture we used to implement this tool also allows the insertion of new variants of NMF or related methods in a straightforward manner. This is particularly important since many new factorization approaches are developing (18). A good example of this is the projected gradient NMF algorithms like ALS and HALS (19,20) which outperforms in many aspects the standard NMF models (see NMFLAB toolbox at http://www.bsp.brain.riken.jp/ICALAB/nmflab.html).

The design and implementation of bioNMF permit nonexpert users in exploring their data with NMF algorithm in an easy and transparent manner or even insert this analysis in their workflows using the provided

web services. Therefore, it is our hope that bioNMF will become an important tool to assist life-sciences researches in the exploratory data analysis cycle.

ACKNOWLEDGEMENTS

This work has been partially funded by the Spanish grants BIO2007-67150-C03-02, S-Gen-0166/2006, CYTED-505 PI0058, CSD00C-07-20811 and TIN2005-5619. E.M.R. is supported by the grant *FPU* from the Spanish Ministry of Education. A.P.M. acknowledges the support of the Spanish *Ramón y Cajal* program. Funding to pay the Open Access publication charges for this article was provided by Spanish Grant. BIO2007-67150-C03-02.

Conflict of interest statement. None declared.

REFERENCES

1. Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.
2. 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. USA*, **101**, 4164–4169.
3. 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 Bioinform.*, **7**, 78.
4. Tamayo,P., Scanfeld,D., Ebert,B.L., Gillette,M.A., Roberts,C.W. and Mesirov,J.P. (2007) Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *Proc. Natl Acad. Sci. USA*, **104**, 5959–5964.
5. Inamura,K., Fujiwara,T., Hoshida,Y., Isagawa,T., Jones,M.H., Virtanen,C., Shimane,M., Satoh,Y., Okumura,S., Nakagawa,K. *et al.* (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**, 7105–7113.
6. 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 Bioinform.*, **7**, 41.
7. Pehkonen,P., Wong,G. and Toronen,P. (2005) Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC Bioinform.*, **6**, 162.
8. Dueck,D., Morris,Q.D. and Frey,B.J. (2005) Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics*, **21 (Suppl. 1)**, i144–i151.
9. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
10. Heger,A. and Holm,L. (2003) Sensitive pattern discovery with 'fuzzy' alignments of distantly related proteins. *Bioinformatics*, **19 (Suppl. 1)**, i130–i137.
11. Lohmann,G., Volz,K.G. and Ullsperger,M. (2007) Using non-negative matrix factorization for single-trial analysis of fMRI data. *Neuroimage*, **37**, 1148–1160.
12. Pascual-Montano,A., Carmona-Saez,P., Chagoyen,M., Tirado,F., Carazo,J.M. and Pascual-Marqui,R.D. (2006) bioNMF: a versatile tool for non-negative matrix factorization in biology. *BMC Bioinform.*, **7**, 366.
13. Kim,P.M. and Tidor,B. (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.*, **13**, 1706–1718.
14. 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**, 403–415.
15. Getz,G., Levine,E. and Domany,E. (2000) Coupled two-way clustering analysis of gene microarray data. *Proc. Natl Acad. Sci. USA*, **97**, 12079–12084.
16. Nielsen,T.O., West,R.B., Linn,S.C., Alter,O., Knowling,M.A., O'Connell,J.X., Zhu,S., Fero,M., Sherlock,G., Pollack,J.R. *et al.* (2002) Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet*, **359**, 1301–1307.
17. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
18. Cichocki,A., Zdunek,R. and Amari,S. (2008) Nonnegative matrix and tensor factorization. *IEEE Signal Processing Magazine*, **25**, 142–145.
19. Cichocki,A. and Zdunek,R. (2007) Regularized alternating least squares algorithms for non-negative matrix/tensor factorizations. *Lect. Notes Comput. Sci.*, **4493**, 793–802.
20. Cichocki,A., Zdunek,R. and Amari,S. (2007) Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization. *Lect. Notes Comput. Sci.*, **4666**, 169–176.