Poster L-18

Non-negative matrix factorization of biological data

Authors:
Alberto Pascual-Montano (*Computer Architecture Department.. Complutense University of Madrid. Spain*)
Pedro Carmona-Saez (*BioComputing Unit. National Center of Biotechnology. Madrid. Spain*)
Monica Chagoyen (*BioComputing Unit. National Center of Biotechnology. Madrid. Spain*)
Francisco Tirado (*Computer Architecture Department.. Complutense University of Madrid. Spain*)
Jose M. Carazo (*BioComputing Unit. National Center of Biotechnology. Madrid. Spain*)
Roberto D. Pascual-Marqui (*The KEY Institute for Brain-Mind Research. Zurich. Switzerland*)

Short Abstract: We describe a versatile tool for using non-negative matrix factorization technique (NMF) in solving different biological problems like biclustering and sample classification using gene expression, as well as other applications in protein sequences and text mining of biomedical literature. This tool is publicly available to the bioinformatics community.

Long Abstract:

In the Bioinformatics field, a great deal of interest has been given to Non-negative matrix factorization technique [1], due to its capability of providing new insights and important information about the complex latent relationships in experimental data sets. This method has been successfully applied to gene expression [2-4], biclustering [5], sequence analysis [6], functional characterization of gene lists [7] and text mining [8].

Even if the interest on this technique has been increased during the last few years, there are not available tools to perform all these types of analysis in an integrated environment. In this work we are proposing a versatile and user-friendly tool that comprises most of the reported applications of this new methodology.

The application we propose has been implemented as a single standalone application and it does not require any special installation or libraries. The tool takes as input a numeric data set in raw text format. If the data is not positive, this tool offers a variety of methodologies to make it positive. Once the factorization has been done, results can be explored using a graphical user interface and can also be saved in graphical and textual format.

The functionality of this tool has been divided in three main modules depending on the application:

Standard NMF: This module performs classical NMF algorithm for gene expression analysis. For example for clustering applications based on local patterns, where NMF is able to represent sets of genes that behaves in a strongly correlated fashion in sub-portions of the data [4]. Similarly, this functionality can also be applied to textual information [8] or protein sequence analysis [6].

Gene expression Biclustering: This module implements a methodology to estimate biclusters using a method based on a modified variant of the NMF algorithm which produces a suitable decomposition as product of two sparse matrices. The methodology implemented in this case is denoted as Non-smooth Non-negative Matrix Factorization (nsNMF) [9], and its application in biclustering gene expression patterns has been reported by [5].
Sample Classification: This module implements the approach proposed in [3] where NMF was used to classify tumor samples. This option implements a statistical approach to help in determining the optimum number of classes.

We hope this application is a useful tool for those researches that need to use of classical NMF algorithm in an easy and efficient way.

Acknowledgements
This work has been partially funded by the Spanish grants CICYT BFU2004-00217/BMC, GEN2003-20235-c05-05, TIN2005-5619, PR27/05-13964-BSCH and a collaborative grant between the Spanish CSIC and the Canadian NRC (CSIC-050402040003). PCS is recipient of a grant from CAM. APM acknowledges the support of the Spanish Ramón y Cajal program.

References:

1. DD Lee, HS Seung: Learning the parts of objects by non-negative matrix factorization. Nature 1999, 401:788-91.
2. Y Gao, G Church: Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics 2005, 21:3970-5.
3. JP Brunet, P Tamayo, TR Golub, JP Mesirov: Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A 2004, 101:4164-9.
4. PM Kim, B Tidor: Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res 2003, 13:1706-18.
5. P Carmona-Saez, RD Pascual-Marqui, F Tirado, JM Carazo, A Pascual-Montano: Biclustering of gene expression data by non-smooth non-negative matrix factorization. BMC Bioinformatics 2006, 7:78.
6. A Heger, L Holm: Sensitive pattern discovery with 'fuzzy' alignments of distantly related proteins. Bioinformatics 2003, 19 Suppl 1:i130-7.
7. P Pehkonen, G Wong, P Toronen: Theme discovery from gene lists for identification and viewing of multiple functional groups. BMC Bioinformatics 2005, 6:162.
8. M Chagoyen, P Carmona-Saez, H Shatkay, JM Carazo, A Pascual-Montano: Discovering semantic features in the literature: a foundation for building functional associations. BMC Bioinformatics 2006, 7:41.
9. A Pascual-Montano, JM Carazo, K Kochi, D Lehmann, RD Pascual-Marqui: Non-smooth Non-Negative Matrix Factorization (nsNMF). IEEE Transactions on Pattern Analysis and Machine Intelligence 2006, 28:403-415.