Two-way clustering of gene expression profiles by sparse matrix factorization

Pascual-Montano, A.[1] ,Carmona-Sáez, P.[2], Pascual-Marqui, R.D.[3], Tirado, F.[1], Carazo, J.M.[2]
[1] Computer Architecture Department. Universidad Complutense de Madrid. 28040. Spain
[2] National Center of Biotechnology. CNB-CSIC. Universidad Autónoma de Madrid. 28049. Spain.
[3] The KEY Institute for Brain-Mind Research. Lenggstr. 31, CH-8029 Zurich, Switzerland
(pascual@fis.ucm.es, pcarmona@cnb.uam.es, pascualm@key.unizh.ch, ptirado@dacya.ucm.es, carazo@cnb.uam.es)

Abstract

We propose a new methodology for two-way cluster analysis of gene expression data using a novel sparse matrix factorization technique that produces a decomposition of a matrix in a set of sparse factors. This method produces a set of bases and coding matrices that are not only able to represent the original data, but they also extract important localized parts-based patterns. We applied the method to gene expression data sets in an attempt to uncover latent relationships between samples and genes in DNA microarray experiments.

1. Introduction

One of the main goals in the analysis of large and heterogeneous gene expression datasets is to identify groups of genes that are co-expressed in subsets of experimental conditions. The identification of these local structures plays a key role to understand the biological events associated to different physiological states as well as to identify gene expression signatures. Classical one-way clustering techniques, especially hierarchical clustering, have been frequently applied to cluster genes and samples separately in order to identify this type of local patterns. In the last few years, many authors have proposed the application of two-way clustering methods (also known as biclustering algorithms) to identify gene-experiment relationships [1].

In this paper we introduce a new biclustering method based on a modified variant of the Non-negative Matrix Factorization (NMF) algorithm [2] that produces a sparse representation of the gene expression data matrix, making possible in this way, its use as a biclustering algorithm. NMF has been introduced as a matrix factorization technique that produces a useful decomposition of data in a product of two matrices that are constrained by having non-negative elements. It can be interpreted as a parts-based representation of the data due to the fact that only additive, not subtractive, combinations are allowed. Here, we show that this matrix decomposition can be used to cluster genes and conditions that are highly related in sub-portions of the data.

2. Methods

Formally, the non-negative matrix decomposition can be described as $\mathbf{V} \approx \mathbf{WH}$ where $\mathbf{V} \in \mathbb{R}^{p \times n}$ is a positive data matrix with p variables (samples) and n objects (genes), $\mathbf{W} \in \mathbb{R}^{p \times q}$ are the reduced q basis vectors or factors ($q \le p$), and $\mathbf{H} \in \mathbb{R}^{q \times n}$ contains the coefficients of the linear combinations of the basis vectors (also known as encoding vectors). All matrices **V**, **W** and **H** are non-negative, and the columns of **W** are normalized (sum up to 1). The objective function, based on the Poisson likelihood, is:

$$D(\mathbf{V}, \mathbf{WH}) = \sum_{i=1}^{p} \sum_{j=1}^{n} \left(V_{ij} \ln \frac{V_{ij}}{(\mathbf{WH})_{ij}} - V_{ij} + (\mathbf{WH})_{ij} \right)$$

The relative magnitude of non-zero coefficients in matrix **W** reflects the relevance of each gene in each factor. In the same way, the coefficient values of the samples in each row of **H** reflect the importance of each factor to approximate the original experiment. Therefore, the set of genes and samples that show high coefficients for the same factor are highly related in a sub-portion of the data and constitute a gene expression bicluster.

Even if NMF has been presented as a method capable of finding the underlying component-based structure of complex data, there is no explicit guarantee in the method to support this property, other than the non-negativity constraints. In fact, in MMF there is a high degree of overlapping among basis vectors that contradict the intuitive nature of the "parts" [3]. Therefore in order to extract significant biclusters a real sparse NMF model capable of producing more localized feature representations is needed. To this end, we conducted a modification of the model as the means to achieve global sparseness. The new model denoted as "Non-Smooth Non-negative Matrix Factorization" (*ns*NMF), is defined as **V**=**WSH**, where the positive symmetric matrix $\mathbf{S} \in \mathbb{R}^{q \times q}$ is a smoothing matrix defined as $\mathbf{S} = (1-\theta)\mathbf{I} + \frac{\theta}{q}\mathbf{1}\mathbf{1}^T$, **I** is the identity matrix, **1** is a vector of ones, and the parameter θ controls the sparseness of the model, satisfying $0 \le \theta \le 1$.

The interpretation of **S** as a smoothing matrix can be explained as follows. Let **X** be a positive, non-zero, vector. Consider the transformed vector **Y**=**SX**. If $\theta = 0$, then **Y**=**X**, and no smoothing on **X** has occurred. However, as $\theta \rightarrow 1$, the vector **Y** tends to the constant vector with all elements almost equal to the average of the elements of **X**. This is the smoothest possible vector, in the sense of "non-sparseness", because all entries are equal to the same non-zero value, instead of having some values close to zero and others clearly non-zero. The new algorithm is derived by simply substituting the *ns*NMF model into the divergence functional defined above, and minimizing it for a given sparseness parameter.

3. Results

In order to evaluate the performance of the method we applied it to the analysis of several simulated datasets as well as the soft-tissue tumor dataset generated by Nielsen *et al*. [4]. In the case of synthetic data, our method was able to correctly identify different types of embedded biclusters, many of which could not be detected by standard clustering algorithms. In the case of the real gene expression dataset, we applied *ns*NMF with *q*=4 and the obtained results showed that the proposed methodology was able to cluster samples belonging to the same tumor type and, at the same time, the set of the most important genes that induced these partitions.

For example, using the coefficients in the first factor to order the genes and samples clustered together the set synovial sarcomas (Figure 1A) and those genes that were co-expressed in this group of tumors. Among the genes that showed high coefficients in this factor were, for example, EGFR and SALL2, which have been previously related to this type of tumors.

Similarly, the second factor revealed the partition of a group of 8 gastrointestinal stromal tumors as well as the set of genes that are relevant to induce this partition (Figure 1B). As in the above case, genes that have been reported as markers of gastrointestinal stromal tumors such as the KIT gene, FLJ10261 (DOG1) or PRKCQ, showed very high coefficients in this factor.

Using the values of the third factor a heterogeneous group of samples comprising liposarcomas, leiomyosarcomas, schwannomas and malignant fibrous histiocytosarcomas were clustered together (Figure 1C). This cluster is similar to the heterogeneous group of tumors discussed in the original paper. Additionally, our approach gave not only a clustering of samples and genes but also their internal ranking within this local structure.

The last structure defined by the fourth sorted factor revealed genes mainly over-expressed in a group containing six of the 11 leiomyiosarcomas and one liposarcoma (Figure 1D). Genes involved in muscle contraction and muscle development (for example CNN1, KCNMB1, MYH11, PLN, SNTA1) showed high coefficients for this factor. It is clear, therefore, the relationship among these biological processes and the tissue origin of the leiomyosarcomas samples.

4. Conclusions

In the Bioinformatics field, a great deal of interest has been given to Biclustering due to its capacity in providing new insights and important information about the complex relationship between genes and experimental conditions. We have shown in this study that Non-smooth Non-negative matrix factorization seems to be a good alternative for this analysis. Experimental results have proved that the *ns*NMF algorithm described here is able to achieve this goal. The obtained representation of the bases show clear localized features of the data due to the sparseness conditions imposed to the algorithm. We hope this new method actively helps in the data analysis and knowledge discovery process in gene expression experiments.

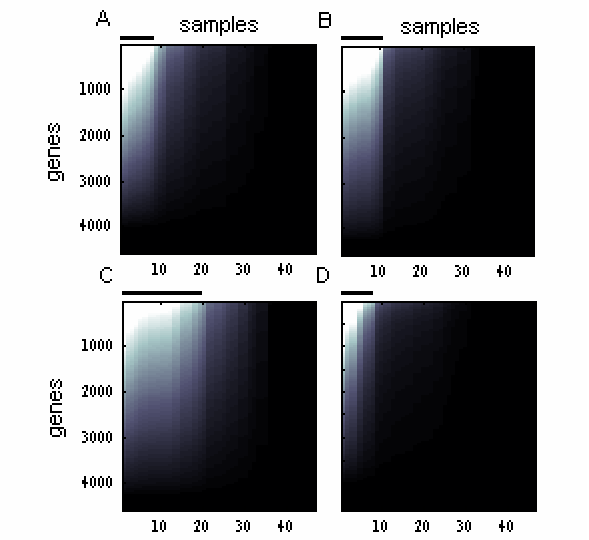

Figure 1. Reconstructed matrix using: A) The first sorted factor, both the basis and the encoding vector (synovial sarcomas are marked with a black line). B) The second sorted factor (gastrointestinal stromal tumors are marked with a black line). C) The third factor (the group of heterogeneous tumor cluster is marked with a black line). D) The fourth factor (six leiomyosarcomas and one liposarcoma samples are marked with a black line)

5. References

[1] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, pp. 24-45, 2004.

[2] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788-91, 1999.

[3] B. W. Mel, "Computational neuroscience. Think positive to find parts," *Nature*, vol. 401, pp. 759-60, 1999.

[4] T. O. Nielsen, *et al.*, "Molecular characterisation of soft tissue tumours: a gene expression study," *Lancet*, vol. 359, pp. 1301-7, 2002.

Acknowledgments: Spanish research grant GR/SAL/0653/2004 and Ramon y Cajal Program.