Generalized Clustergrams for Overlapping Biclusters

Liviu Badea
AI Lab, National Institute for Research in Informatics
badea@ici.ro

Abstract

Many real-life datasets, such as those produced by gene expression studies, exhibit complex substructures at various levels of granularity and thus *do not have unique well-defined numbers of clusters*. In such cases, it is important to be able to trace the evolution of the individual clusters as the number of dimensions of the clustering is varied. While the *dendrograms* produced by bottom-up clustering methods such as hierarchical clustering are very useful for this purpose, the approach is known to produce unreliable clusters due to its instability w.r.t. resampling. Moreover, hierarchical clustering does not apply to *overlapping* (*bi*)clusters, such as those obtained in gene expression studies. On the other hand, the instability w.r.t. the initialization of top-down methods, such as k-means, prevents the comparison between clusters obtained at different dimensionalities. In this paper, we present a method for constructing *generalized dendrograms for overlapping biclusters*, which depict the evolution of the biclusters as their number is varied. An essential ingredient is a stable biclustering method based on positive tensor factorization of a number of nonnegative matrix factorization runs. We apply our approach to a large colon cancer dataset, which shows several distinct subclasses whose dimensional evolution must be carefully analyzed to enable a more meaningful biological interpretation and sub-classification.

1 Introduction

Biological processes are extremely complex, showing a hierarchical organization at various levels of granularity. On the other hand, many clustering methods require the number of clusters to be given as input. But in real gene expression data [Eisen et al., 1998] one cannot unequivocally determine a well-defined *number of clusters*, as coarser-grained clusters may exhibit progressively finer-grained structure. For example, Figure 1 shows the evolution of the error of nonnegative decompositions (biclusterings) as the number of clusters is increased in the case of two datasets: a synthetic dataset with 5 biclusters (Figure 1a) and a large colon cancer dataset (Figure 1b).

Note that in the synthetic dataset, a well-defined number of clusters ($n_c = 5$) can be discerned, while the colon cancer dataset shows a steep drop in error until $n_c = 3$, followed by a series of progressively smaller ones. The latter could represent either overfitting or finer-grained substructure and only a more in-depth biological analysis can settle the issue. Anyhow, it is not enough from a biological point of view to determine a unique number of clusters and to analyze the resulting clusters at that fixed dimensionality. Instead, we need to perform some sort of multi-scale analysis of the clusters and their evolution as the dimensionality (n_c) of the clustering is varied.

For this, it is essential to be able to determine the relationships between the clusters generated at different n_c[1], which is possible only if we can guarantee the *stability* of the clusters and if we have a mathematically sound method of *comparing biclusters*. (We deal with *biclusters* [Cheng and Church, 2000] since in the case of gene expression data, as in many other domains, objects/genes tend to be correlated only for certain subsets of samples, i.e. specific biological contexts.)

We achieve biclustering *stability* using the meta-clustering approach of [Badea, 2005; Badea and Tilivea, 2007], which is based on a *positive tensor factorization* (PTF) of the biclusters obtained in various repeated clustering runs. *Bicluster comparisons* are "built into" this approach in an elegant manner.

These two key features allow us to construct a series of decompositions at varying n_c and to trace the evolution of the individual clusters as n_c changes. This generalizes hierarchical clustering dendrograms to a more complicated setting where items that have been grouped (at a certain n_c) can be separated later on at a lower dimensionality.

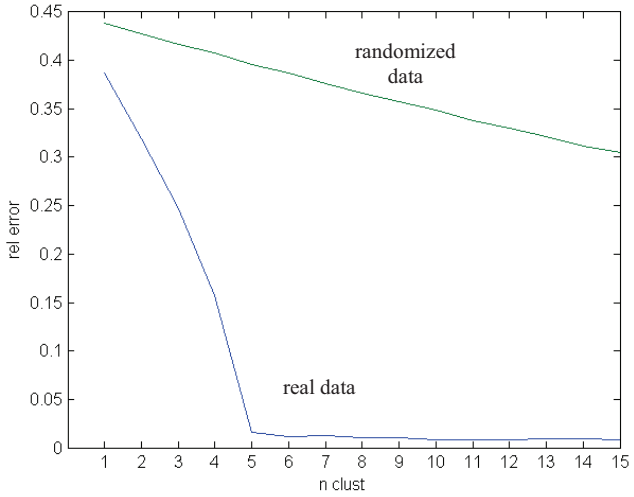
Hierarchical clustering is probably the most frequently used clustering method in the domain of gene expression data analysis [Eisen et al., 1998]. This is due not only to its simplicity, but also to the very intuitive nature of the clustering *dendrograms* it produces, which graphically depict the evolution of clusters at various dimensionalities. Since

[1] For example, we may want to analyze in more detail the cluster merges and splits as n_c is varied.

choosing a unique well-defined number of clusters is usually problematic in the case of gene expression data, being able to trace the dimensional evolution of clusters is essential in this domain.

However, extensive experimental evaluations [Thalamuthu et al., 2006] have demonstrated that hierarchical clustering shows mediocre performance especially for noisy data, with many "unrelated" (or, so called "scattered") genes. Thus, it might seem that *performance* has been sacrificed for better *visualization* and that an inherent trade-off exists between the two. In the following we show that we can trace the dimensional evolution of clusters in a more complicated setting involving biclustering based on PTF by developing a generalization of dendrograms for the case of overlapping biclusters.

a. synthetic data (5 biclusters)



b. colon adenocarcinoma

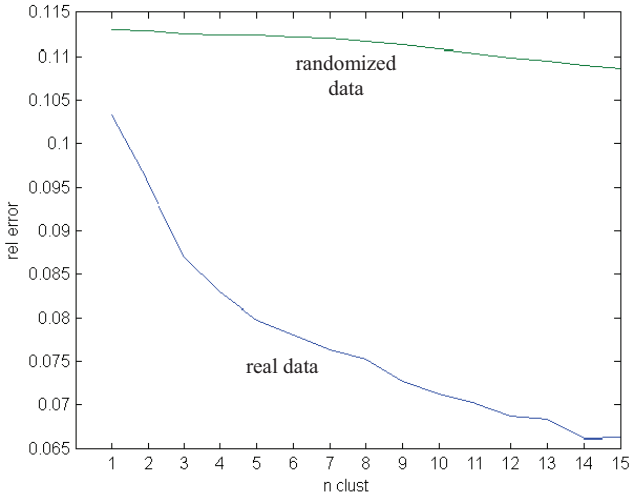


Figure 1. The dimensional evolution of NMF decomposition errors for a synthetic (a) and a real-life dataset (b).

2 Stable meta-clustering with PTF

As already mentioned in the Introduction, the ability of clustering genes and samples simultaneously ('biclustering') is essential for analyzing gene expression data, where genes tend to be co-expressed only for certain *subsets* of samples, corresponding to specific biological contexts. In the following, we use the biclustering approach based on positive tensor factorizations (PTF) of [Badea, 2005; Badea and Tilivea, 2007], which we briefly review here. The increased stability of the meta-clustering approach is essential for being able to follow the dimensional evolution of biclusters.

2.1 Biclustering using Nonnegative Matrix Factorizations with offset

An elegant method of biclustering consists in factorizing the gene expression matrix X as a product of an $n_s \times n_c$ (samples × clusters) matrix A and an $n_c \times n_g$ (clusters × genes) matrix S [2]

$$X_{sg} \approx \sum_c A_{sc} \cdot S_{cg} + So_g \quad (1)$$

subject to additional nonnegativity constraints:

$$A_{sc} \geq 0, S_{cg} \geq 0, So_g \geq 0 \quad (2)$$

which express the obvious fact that expression levels and cluster membership degrees cannot be negative.

Factorization (1) differs from the standard *Nonnegatuive Matrix Factorization* (NMF) [Lee and Seung, 1999; 2000] by the additional "*gene offset*" *So*, whose main role consists in absorbing the constant expression levels of genes, thereby making the cluster samples S_{cg} "cleaner".

The factorization (1-2) can be regarded more formally as a constrained optimization problem:

$$\min f(A,S,So) = \frac{1}{2}||X - A \cdot S - e \cdot So||_F^2 = \frac{1}{2}\sum_{s,g}(X - A \cdot S - e \cdot So)_{sg}^2 \quad (3)$$

subject to the nonnegativity constraints (2). This problem can be solved using an iterative algorithm with the following multiplicative update rules (which can be easily derived using the method of Lee and Seung [Lee and Seung, 2000]):

$$\begin{aligned} A_{sc} &\leftarrow A_{sc} \frac{(X \cdot S^T)_{sc}}{((A \cdot S + e \cdot So) \cdot S^T)_{sc} + \varepsilon} \\ S_{cg} &\leftarrow S_{cg} \frac{(A^T \cdot X)_{cg}}{(A^T \cdot (A \cdot S + e \cdot So))_{cg} + \varepsilon} \\ So_g &\leftarrow So_g \frac{(e^T \cdot X)_g}{(e^T \cdot (A \cdot S + e \cdot So))_g + \varepsilon} \end{aligned} \quad (4)$$

where e is a column vector of 1 of size equal to the number of samples and ε is a regularization parameter (a very small positive number).

[2] X_{sg} represents the gene expression level of gene g in sample s, S_{cg} the *membership degree* of gene g in cluster c and A_{sc} the mean *expression level of cluster* (biological process) c in sample s.

The algorithm initializes A, S and So with random entries, so that (slightly) different solutions may be obtained in different runs. (This is due to the non-convex nature of the optimization problem (3), which in general has many different local minima.)

We can view the different solutions obtained by the generalized NMF_O algorithm as *overfitted* solutions, whose *consensus* we'll need to construct.

We have observed experimentally that adding offsets to standard *NMF* leads to significant improvements in the quality of the recovered clusters.

More precisely, the genes with little variation are reconstructed by the standard *NMF* algorithm from combinations of clusters, while NMF_O uses the additional degrees of freedom So to produce null cluster membership degrees S_{cg} for these genes. Moreover, NMF_O recovers with much more accuracy than standard *NMF* the original sample clusters, the standard *NMF* algorithm being confused by the cluster overlaps. This improvement in recovery of the original clusters is very important in our application, where we aim at a correct sub-classification of samples.

2.2 Meta-clustering with PTF

Unfortunately, virtually all clustering methods that are currently used for gene expression data analysis tend to produce highly unstable clusters, especially when clustering genes. (The instability manifests itself either w.r.t. the initialization of the algorithm, as in the case of k-means and NMF, or w.r.t. small perturbations of the dataset in the case of deterministic algorithms, such as hierarchical clustering.)

A frequently used method to obtain more stable clusters consists in building a *consensus* of several individual clusterings constructed from different NMF_O initializations.

More precisely, starting with a number of NMF_O runs

$$X \approx A^{(i)} \cdot S^{(i)} + e \cdot So^{(i)} \qquad i = 1,...,r \tag{5}$$

a *consensus biclustering* is constructed using a *Positive Tensor Factorization* (PTF) [Welling and Weber, 2001] of the biclusters[3], which simultaneously determines the bicluster correpondence α and the consensus biclustering (β,γ) [Badea, 2005; Badea and Tilivea, 2007]:

$$A_{s(ic)} \cdot S_{(ic)g} \approx \sum_{k=1}^{n_c} \alpha_{(ic)k} \cdot \beta_{sk} \cdot \gamma_{kg} \tag{6}$$

where s are samples, g, genes, c clusters and k metaclusters (or "consensus clusters").[4] β and γ represent the *consensus* of $A^{(i)}$ and $S^{(i)}$ respectively. More precisely, the columns β_k of β and the corresponding rows $\gamma_{k\cdot}$ of γ make up a *base set of bicluster prototypes* $\beta_k \cdot \gamma_{k\cdot}$ out of which all biclusters of all individual runs can be recomposed, while α encodes the *(bi)cluster-metacluster correspondence*. The factorization (6) can be computed using the multiplicative update rules from [Badea, 2005; Badea and Tilivea, 2007]:

$$\begin{aligned}
\alpha &\leftarrow \alpha * \frac{(A^T \cdot \beta) * (S \cdot \gamma^T)}{\alpha \cdot [(\beta^T \cdot \beta) * (\gamma \cdot \gamma^T)]} \\
\beta &\leftarrow \beta * \frac{A \cdot [\alpha * (S \cdot \gamma^T)]}{\beta \cdot [(\alpha^T \cdot \alpha) * (\gamma \cdot \gamma^T)]} \\
\gamma &\leftarrow \gamma * \frac{[\alpha * (A^T \cdot \beta)]^T \cdot S}{[(\alpha^T \cdot \alpha) * (\beta^T \cdot \beta)]^T \cdot \gamma}
\end{aligned} \tag{7}$$

where '$*$' and '$--$' represent element-wise multiplication and division of matrices, while '$\cdot$' is ordinary matrix multiplication. After convergence of the PTF update rules, the rows of γ are normalized to unit norm to make the gene clusters directly comparable to each other, whereas the columns of α are normalized such that $\sum_{i,c} \alpha_{(ic)k} = r$ (r is the number of runs). Then, NMF_O initialized with (β,γ,γ_0) is run[5] to produce the final factorization $X \approx A \cdot S + e \cdot So$.

The nonnegativity constraints of PTF meta-clustering are essential both for allowing the interpretation of $\beta_k \cdot \gamma_{k\cdot}$ as consensus biclusters, as well as for obtaining sparse factorizations. In practice, the rows of the correspondence matrix α tend to contain typically one or only very few significant entries.

3 PTF for tracing the dimensional evolution of biclusters

Given a stable set of biclusters ($A^{(i)}$, $S^{(i)}$, $So^{(i)}$) generated for progressively larger numbers of clusters $n_c = i \in \{2,...,n_{c_{\max}}\}$:

$$X_{sg} = \sum_{c=1}^{i} A_{sc}^{(i)} S_{cg}^{(i)} + So_g^{(i)}, \tag{8}$$

we aim at determining the relationships between the biclusters at $n_c = i$ and those at larger n_c (e.g. $n_c = i + 1$). (For example, bicluster c_1 at $n_c = i$ may be very similar to bicluster c_2 at $n_c = i + 1$. Or, bicluster c_3 at $n_c = i$ may result from merging biclusters c_4 and c_5 at $n_c = i + 1$.)

To achieve this, we start with a "*reference*" factorization $(A^{(ref)}, S^{(ref)}, So^{(ref)})$ that is *fine-grained* enough to approximate any lower-dimensional factorization (for example with $n_c = n_{c_{\max}}$ [6]).

We then express the biclusters ($A^{(i)}$, $S^{(i)}$, $So^{(i)}$) obtained at various dimensionalities in terms of the fine-grained biclusters of the reference factorization as follows:

[3] A tensor factorization is needed instead of a matrix factorization since biclusters are matrices.

[4] To simplify the notation, the indices i and c were merged into a single index (ic).

[5] γ_0 is obtained from the 1-dimensional NMF decomposition $So_g^{(i)} = \alpha_{00}^{(i)} \gamma_{0g}$ with the normalization $\sum_i \alpha_{00}^{(i)} = r$.

[6] We can estimate $n_{c_{\max}}$ using the error curve as in Figure 1.

$$A^{(i)}_{sc} \cdot S^{(i)}_{cg} \cong \sum_{k=0}^{n_{c_{\max}}} \lambda^{(i)}_{ck} \cdot A^{(ref)}_{sk} \cdot S^{(ref)}_{kg} \quad (9)$$

where $A^{(i)}_{s0} = 1$, $A^{(ref)}_{s0} = 1$, $\lambda^{(i)}_{c0} = 0$ for $c \neq 0$ and $\lambda^{(i)}_{0k} = 0$ for $k \neq 0$.

More precisely, (9) can be expressed as an optimization problem

$$\min p(\lambda) = \frac{1}{2} \sum_{i,c,s,g} (A^{(i)}_{sc} S^{(i)}_{cg} - \sum_{k} \lambda^{(i)}_{ck} A^{(ref)}_{sk} S^{(ref)}_{kg})^2 \quad (10)$$

subject to the constraints $\lambda^{(i)}_{ck} \geq 0$.

Note that (10) is similar to the positive tensor factorization (6), except that in (10) only $\lambda^{(i)}_{ck}$ are free variables. Moreover, since in (10) the factorizations (i) and the reference factorization have different dimensionalities, the consensus tensor $\lambda^{(i)}_{ck}$ may contain several significant entries $\lambda^{(i)}_{ck_1}$, $\lambda^{(i)}_{ck_2}$,… on a given row c (corresponding to a cluster c formed by superposing (merging) reference clusters k_1, k_2, …).

It can be shown that the decomposition (10) is computable using the following multiplicative update rule:

$$\lambda \leftarrow \lambda * \frac{(A^T \cdot A^{(ref)}) * (S \cdot S^{(ref)T})}{\lambda \cdot \left[(A^{(ref)T} \cdot A^{(ref)}) * (S^{(ref)} \cdot S^{(ref)T}) \right]} \quad (11)$$

where we have merged indices i and c as (ic):

$$\lambda_{(ic)k} = \lambda^{(i)}_{ck}, \quad A_{s(ic)} = A^{(i)}_{sc}, \quad S_{(ic)g} = S^{(i)}_{cg} \quad (12)$$

The cluster correspondence matrix λ generalizes hierarchical clustering dendrograms and will be called in the following "*generalized clustergram*". For example, Figure 2.c shows λ for the clustering dendrogram from Figure 2.a. λ can thus be used to trace the evolution of individual clusters as n_c is varied. In the case of hierarchical clustering, once two items have been grouped (at a given n_c), they remain grouped for all smaller n_c. The clustergram from Figure 2.b and the generalized clustergram (correspondence matrix) from Figure 2.c can be used to describe more complex cluster evolutions such as that shown in Figure 3, which presents the evolution of PTF decompositions for n_c ranging from 2 to 8 using a 15-dimensional reference factorization of a synthetic dataset with 5 partially overlapping biclusters (for reasons of layout, Fig. 3 presents the transpose of λ).

For visualization purposes, the rows c of $\lambda_{(ic)k}$ were permuted such that the cluster c from decomposition i matches cluster c of decomposition $i-1$ for all i and $c = 1, \ldots, i-1$. This was achieved in a greedy manner as follows:

for $i = 2 \ldots n_{c_{\max}}$
  for $c = 1, \ldots, i-1$
    let $k_{\max} = \arg\max_k \lambda^{(i-1)}_{ck}$
    and $c_{\max} = \arg\max_{c'} \lambda^{(i)}_{c'k_{\max}}$ s.t. c' is not assigned
    assign row $c_{\max}$ of decomposition i to position c
  assign remaining row of decomposition i to position i

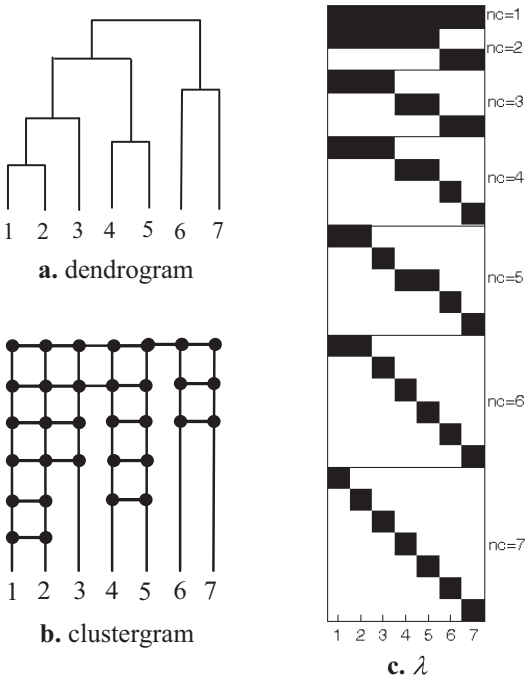


Figure 2. λ generalizes hierarchical clustering dendrograms. A hierarchical clustering dendrogram (a.) with its associated clustergram (b.) and generalized clustergram λ (c.) While the dendrogram is tree-like, the clustergram and the generalized clustergram λ may be more general.

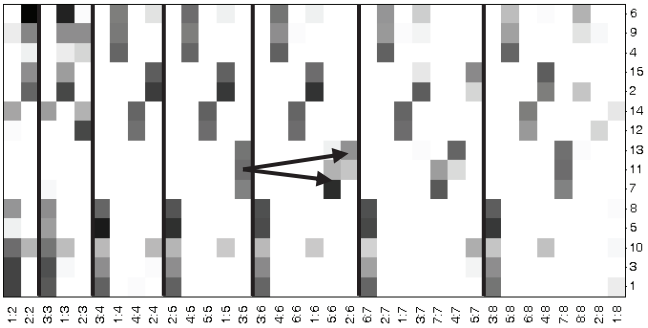


Figure 3. The generalized clustergram (correspondence matrix) λ^T for a synthetic dataset with 5 partially overlapping biclusters using a 15-dimensional reference factorization (only n_c=2,...,8 are shown)

The tests of the algorithm on synthetic datasets showed that it is quite effective at tracing the *incremental* evolution of the individual biclusters as the number of clusters is increased, despite the *variability* of the individual NMF clustering runs. (As far as we know, this is the first algorithm to achieve this.) For example, Figure 3 shows that the clusters for n_c=5 are essentially preserved when going to n_c=6, except for cluster 3:5 (i.e. cluster 3 for n_c=5), which splits into two overlapping clusters 2:6 and 5:6.

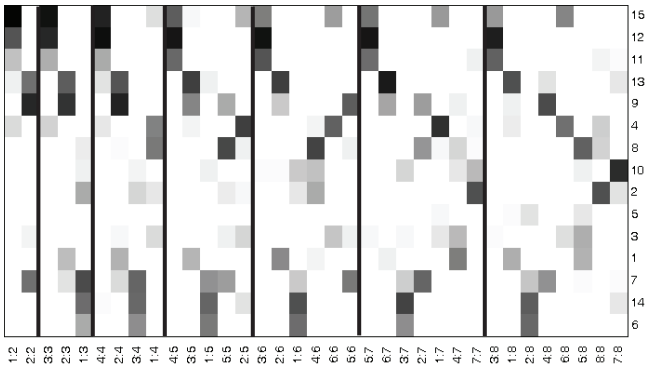


Figure 4. The generalized clustergram (correspondence matrix) λ^T for the colon adenocarcinoma dataset using a 15-dimensional reference factorization (only n_c=2,...,8 are shown)

4 Multi-scale analysis of a colon cancer dataset

We have applied our approach to a large gene expression dataset of colon adenocarcinoma. Our preliminary investigations of this dataset have shown that the disease has several distinct subclasses, but no unique well-defined dimensionality could be determined (as discussed in the Introduction – see also Figure 1.b).

The most frequent colon cancer type, *sporadic colon adenocarcinoma*, is very heterogeneous and its best current classification based on the presence or absence of microsatellite instabilities (MSI-L, MSI-H and MSS) [Jass et al., 1999] is far from ideal from the point of view of gene expression. A more refined analysis of the biclusters generated at different dimensionalities is therefore needed in order to be able to determine the *biologically relevant subclasses* of this disease.

To obtain a more accurate subclassification based on gene expression profiles, we have applied our approach to a large dataset (204 samples) containing 182 colon adenocarcinoma samples from the expO database [expO] and 22 control ("normal") samples from [Hong et al., 2007]. (All of these had been measured on Affymetrix U133 Plus 2.0 chips.)

The combined raw scanning data was preprocessed with the RMA normalization and summarization algorithm [Irizarry et al., 2003]. (The logarithmic form of the gene expression matrix was subsequently used, since gene expression values are approximately log-normally distributed.) After eliminating the probe-sets (genes) with relatively low expression as well as those with a nearly constant expression value[7], we were left with 3666 probe-sets. Finally, the Euclidean norms of the expression levels for the individual genes were normalized to 1 to disallow genes with higher absolute expression values to overshadow the other genes in the factorization.

To estimate the number of clusters (n_c), we compared the dimensional evolution of the decomposition error $\varepsilon_{rel} = \|X - A \cdot S - e \cdot So\|_F / \|X\|_F$ of PTF meta-clustering of the real dataset with that corresponding to a randomized dataset [8] (similar to [Kim and Tidor, 2003], see Figure 1.b).

Figure 4 depicts the evolution of PTF decompositions of the colon adenocarcinoma dataset for n_c ranging from 2 to 8 using a 15-dimensional reference factorization. Note that as we go from n_c = 2 to n_c = 3, a new cluster appears, involving fine-grained subclusters (rows) k = 7,14,6,2. This cluster is largely conserved at n_c = 4, where another new cluster (involving k = 4,8) is introduced. At n_c = 5, the latter suffers a complex transition: it essentially splits in two, but each of the two subclusters picks up additional contributions, so that they end up as clusters 5:5 (corresponding to k = 8,7,9) and respectively 2:5 (k = 4,15). The fact that additional subclusters are involved at n_c = 5 as compared to n_c = 4 suggests that there are at least 5 biologically significant clusters.

At n_c = 6, the main change w.r.t. n_c = 5 consists in the appearance of cluster 4:6 (k = 8,2,10).[9] It may be interesting to note that subcluster k = 10 is formed virtually only by probesets associated to the *XIST* ('X inactive-specific transcript') gene, which is the major effector of X chromosome inactivation (normally expressed only in females by inactivated X chromosomes). Actually, cluster 4:6 is generated, together with 5:6, from 5:5. At n_c = 7, the two clusters 4:6 and 5:6 are merged and split again to 2:7 (which is almost identical to 5:5) and 7:7 (which contains subclusters k = 2,10). Therefore, the transition from n_c = 5 to n_c = 7 amounts to adding two new clusters 4:7 and 7:7.

For a biological interpretation of the biclusters, we have used several high-throughput studies of microsatellite-instability in colon cancer (e.g. [Banerjea et al., 2004]). Clusters 1:5, 1:6 and 3:7 correspond to a very well defined *microsatellite stable* subtype (MSS as defined in [Jass et al., 1999]). More specifically, the keratin 23 gene, which we find specifically over-expressed in these clusters was previously known to be specific to MSS colon cancer [Birkenkamp et al., 2007]. Some of the most important genes involved in this bicluster are the following transcription factors: ASCL2, DACH1, FOXQ1, EREG, TNRC9, all previously related to colon cancer and which presumably control the various biological processed involved in microsatellite stable colon carcinoma.

Clusters 5:5, 5:6, 2:7 correspond to the microsatellite instability-high (MSI-H) subtype, which very interestingly over-expresses the developmental homeobox genes HOXC6, PRRX1, HOP. Moreover, the genes overexpressed in this cluster are typically overexpressed in MSI-H tumors

[7] Only genes with an average expression value over 100 and with a standard deviation above 150 were retained.

[8] The randomized dataset was obtained by randomly permuting for each gene its expression levels in the various samples. The original distribution of the gene expression levels is thereby preserved.

[9] Note that in our method, unlike in the case hierarchical clustering, subclusters that have been merged at a given dimension (for instance k=7,8,9 at n_c=5) can later on be separated if the this improves the overall clustering (e.g. at n_c=6, k=8 is separated from the other two subclusters and merged with k=2,10).

with BRAF (rather than KRAS) mutations (using [Kim et al., 2006]).

The normal colon samples are grouped in clusters 4:5, 3:6, 5:7, whose genes are *down-regulated* in colon cancer. For instance, the down-regulation of carcinoembryonic antigen CEACAM7 is an early event in colorectal tumorigenesis [Thompson et al., 1997], while the critical cell cycle gene CDKN2B is frequently inactivated in colon cancer [Ishiguro et al., 2006].

Our approach to tracing the dimensional evolution of biclusters has proved a useful tool for analyzing the complex subclassification of colon adenocarcinoma that seems to emerge from such gene expression studies.

5 Conclusions

Many real-life data mining datasets and most gene expression datasets exhibit complex substructures at various levels and thus do not have unique well-defined numbers of clusters. Therefore, it is essential to be able to trace the evolution of gene expression biclusters as the number of dimensions of the clustering is varied.

Hierarchical clustering dendrograms are visually very informative in this regard, but have certain essential drawbacks related to their rather mediocre performance for noisy data, as well as to their *unidimensional* nature (for gene expression data, *biclustering* methods are more appropriate, since genes tend to be co-expressed only for certain subsets of samples, in certain specific biological contexts). Also, hierarchical clustering produces *tree-like* cluster merging structures, which may not reflect the dimensional evolution of real-life (biological) processes.

In this paper, we present an original approach that enables tracing the evolution of biclusters as the clustering dimensionality is varied. The method heavily relies on the *stability* of a positive tensor factorization-based metaclustering of NMF decompositions. We also show that the cluster correspondence matrix λ of the decomposition of the factorizations produced at various n_c (w.r.t. a common reference factorization) can play the role of a generalization of hierarchical clustering dendrograms.

Finally, we applied our approach to a large colon adenocarcinoma dataset, which shows a complex, non-tree-like dimensional evolution of subclusters. Our method has already proven to be useful in the difficult task of determining the biologically relevant subclasses of colon adenocarcinoma.

References

[Badea, 2005] Badea Liviu. Clustering and Metaclustering with Nonnegative Matrix Decompositions. Proc. ECML-2005, Vol. 3720, pp. 10-20.

[Badea and Tilivea, 2007] Badea Liviu, D. Tilivea. Stable Biclustering of Gene Expression Data with Nonnegative Matrix Factorizations. Proc. IJCAI-07, pp. 2651-2656.

[Banerjea et al., 2004] Banerjea A, et al. Colorectal cancers with microsatellite instability display mRNA expression signatures characteristic of increased immunogenicity. Mol Cancer. 2004 Aug 6;3:21.

[Birkenkamp et al., 2007] Birkenkamp-Demtroder K. et al. Phosphoprotein Keratin 23 accumulates in MSS but not MSI colon cancers in vivo and impacts viability and proliferation in vitro. Mol Oncology 1(2007),181–195.

[Cheng and Church, 2000] Cheng Y., G. Church. Biclustering of expression data. Proc. ISMB-2000, 93-103.

[Eisen et al., 1998] Eisen M.B., P.T. Spellman, P.O. Brown, D. Botstein. Cluster analysis and display of genome-wide expression patterns, PNAS Vol.95, 14863-8, Dec.1998.

[expO] expO. Expression Project for Oncology http://expo.intgen.org/expo/geo/goHome.do

[Hong et al., 2007] Hong Y, K.S. Ho, K.W. Eu, P.Y. Cheah. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. Clin Cancer Res. 2007 Feb 15;13(4):1107-14.

[Irizarry et al., 2003] Irizarry R.A., B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, T.P. Speed. Summaries of Affimetrix GeneChip probe level data. Nucleic Acids Res 2003; 31(4):e15.

[Ishiguro et al., 2006] Ishiguro A, et al. J Gastroenterol Hepatol. 2006 Aug; 21(8):1334-9. (PMID: 16872319)

[Jass et al., 1999] Jass J.R., et al. Characterisation of a subtype of colorectal cancer combining features of the suppressor and mild mutator pathways. J.Clin.Pathol. 52: 455-460, 1999.

[Kim and Tidor, 2003] Kim P.M., B. Tidor. Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res. 2003 Jul;13(7):1706-18.

[Kim et al., 2006] Kim IJ, et al. Carcinogenesis. 2006 Mar; 27(3):392-404. (PMID:16219636)

[Lee and Seung, 1999] Lee D.D., H.S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, vol. 401, no. 6755, pp. 788-791, 1999.

[Lee and Seung, 2000] Lee D.D., H.S. Seung. Algorithms for non-negative matrix factorization. Proc. NIPS-2000, pp. 556-562, MIT Press.

[Thalamuthu et al., 2006] Thalamuthu A. et al. Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics, Vol. 22, no. 19 (2006), pp. 2405-12.

[Thompson et al., 1997] Thompson J, et al. Cancer Res. 1997 May 1;57(9):1776-84. (PMID:9135022)

[Welling and Weber, 2001] Welling M., M. Weber. Positive tensor factorization. Pattern Recognition Letters 22(12): 1255-1261 (2001).