A stability-based algorithm to validate hierarchical clusters of genes

Roberto Avogadri
Matteo Re
Giorgio Valentini

DSI - Dip. Scienze dell' Informazione, Università degli Studi di Milano, Italy. E-mail:{avogadri,re,valentini}@dsi.unimi.it

Matteo Brioschi
Alessandro Beghini

DBioGen - Dip. Biologia e Genetica per le Scienze Mediche, Università degli Studi di Milano, Italy.
E-mail:{matteo.brioschi,alessandro.beghini}@unimi.it

Fulvia Ferrazzi

Dip. Informatica e Sistemistica, Università degli Studi di Pavia, Italy. E-mail: fulvia.ferrazzi@unipv.it

Abstract: Stability-based methods have been successfully applied in functional genomics to the analysis of the reliability of clusterings characterized by a relatively low number of examples and clusters. The application of these methods to the validation of gene clusters discovered in bio-molecular data may lead to computational problems due to the large amount of possible clusters involved. To address this problem, we present a stability-based algorithm to discover significant clusters in hierarchical clusterings with a large number of examples and clusters. The reliability of clusters of genes discovered in gene expression data of patients affected by Human Myeloid Leukemia is analyzed through the proposed algorithm, and their relationships with specific biological processes are tested by means of Gene Ontology-based functional enrichment methods.

Keywords: hierarchical clustering; stability-based methods; cluster validation; DNA microarray

Reference to this paper should be made as follows: R. Avogadri, M. Brioschi, F. Ferrazzi, M. Re, A. Beghini, and G. Valentini 'A stability-based algorithm to validate hierarchical clusters of genes', *International Journal of Knowledge Engineering and Soft Data Paradigms (IJKESDP), Vol. x, No. x, pp.xxx–xxx.*

Biographical Notes: Roberto Avogadri received the “laurea” degree in Computer Science from the University of Milan. He is currently Ph.D. student in Computer Science in the same university. His research interests focus on bioinformatics and machine learning.

Matteo Brioschi received the ”laurea” degree in Industrial Biotechnology from the University of Milan-Bicocca. He is supported by an investigator fellowship from Niguarda Hospital of Milan. He is currently involved in several investigations on genome-wide approach at Department of Biology and Genetics for Medical Science of the University of Milan. His research interests focus on genomics and genetics.

Fulvia Ferrazzi received the Laurea (Master’s degree) in Computer Science and Engineering in 2003 with honors and the PhD in Bioengineering and Bioinformatics in 2007, both from the University of Pavia, Italy. Dr. Ferrazzi is currently a post-doctoral fellow at the Biomedical Informatics Laboratory, University of Pavia. Her research interests are probabilistic graphical models, microarray data analysis and Bayesian methods for genomic data analysis. She is author of 30 publications in international peer-reviewed journals, books and conference proceedings. She won several academic awards, has acted as a reviewer for international journals and conferences, and collaborates with several research groups in Italy and abroad.

Matteo Re received the “laurea” degree in Biological Sciences from the University of Milan, and the Ph.D. in Cellular and Molecular Biology from the DSBB, Biomolecular Sciences and Biotechnology Department of the same university. He is currently a post-doctoral fellow at DSI, Computer Science Department of the University of Milan. His research interests focus on bioinformatics, machine learning, functional genomics and comparative genomics.

Alessandro Beghini received the ”laurea” degree in Biological Sciences at the University of Milan, and the PhD in Medical Genetics at the University of Genova. He is currently Assistant Professor to the Department of Biology and Genetics for Medical Sciences at Medical Faculty of the University of Milan. His research focuses on medical genetics and oncogenomics. He is a member of the Italian Society of Human Genetics and Editorial board member of the online Atlas of Genetics and Cytogenetics in Oncology and Haematology. He is author of more than 30 full papers published in international peer-reviewed journals.

Giorgio Valentini received the “laurea” degree in Biological Science and in Computer Science from the University of Genova, and the Ph.D. in Computer Science from the DISI, Computer Science Department of the same university. He is currently assistant professor at DSI, Computer Science Department of the University of Milano, where he attends to both teaching and research. His research interests focus on bioinformatics and machine learning. He is author of about 70 papers published in international peer-reviewed journals, books and conference proceedings. He is member of the International Neural Network Society and of the International Society of Computational Biology.

1 Introduction

The unsupervised analysis of clusters in complex biomolecular data plays a central role in bioinformatics (Dopazo, 2006; Jiang *et al.*, 2004), and raises important issues ranging from the proper visualization of high-dimensional clustering results (Napolitano *et al.*, 2008), to the discovery of multiple structures underlying the data (Bertoni and Valentini, 2008), and to the validation and the assessment of the reliability of the discovered clusters (Datta and Datta, 2003).

In this context, different clustering validation techniques (see (Handl *et al.*, 2005) for a recent review), and software tools implementing classical validity indices (such as the *Dunn's index* and the *Silhouette index*) have been proposed (Bolshakova *et al.*, 2005).

Several recent methods to estimate the validity of the discovered clusterings are based on the concept of stability: multiple clusterings are obtained by introducing perturbations into the original data, and a clustering is considered reliable if it is approximately maintained across multiple perturbations (Kerr and Curchill, 2001; Monti *et al.*, 2003; Ben-Hur *et al.*, 2002; McShane *et al.*, 2002). Despite their successful application in several bioinformatics domains, these methods are well-suited to unsupervised problems characterized by a relatively low number of clusters and/or examples (Smolkin and Gosh, 2003; Bertoni and Valentini, 2006). Indeed if we try to apply them to the analysis of a very high number of clusters, computational problems may arise. For instance, to assess the reliability of clusters of N genes using DNA microarray data, we usually deal with thousands of examples (genes) and with an exponential ($O(2^N)$) number of potential clusters.

Considering that clusters of genes may show a hierarchical multi-level organization (Bertoni and Valentini, 2007), we could reduce the computational complexity by examining a linear number of clusters, computed by a hierarchical clustering algorithm.

The main idea of this work consists in the assessment of the reliability of the clusters discovered by a hierarchical clustering algorithm, using a stability based measure borrowed from our previous work (Bertoni and Valentini, 2006). Differently from our previous approach, we do not need to know in advance the correct or approximate number of clusters, but we can directly apply a stability measure that estimates the reliability of each individual cluster of the dendrogram computed by a hierarchical algorithm, thus reducing the complexity to a linear number of clusters with respect to the number of available examples.

In Section 2 we describe the proposed algorithm. In Sect. 3 we introduce an application of the algorithm to the discovery of significant gene clusters in patients affected by Human Myeloid Leukemia, by using DNA microarray gene expression data prepared and analyzed by our research group using the Affymetrix Human Genome U133 Plus 2.0 arrays. Functional enrichment of the most stable clusters was performed relying on Biological Processes represented in the Gene Ontology (The Gene Ontology Consortium, 2000). In Sect. 4 we discuss the advantages and the limitations of the proposed method and we propose some research lines for future work.

2 The algorithm

Our algorithm relies on a stability based approach to discover the significant clusters identified by a hierarchical clustering algorithm. The main logical steps of the algorithm are the following:

1. **Hierarchical clustering of the original data.** A hierarchical clustering algorithm is applied to the original data to discover the clusters whose reliability will be evaluated through the steps listed below.

2. **Multiple perturbation of the original data.** The original data are perturbed by randomized projections (Achlioptas, 2003), by subsampling or bootstrapping procedures (Efron and Tibshirani, 1993), or by controlled noise injection.

3. **Multiple hierarchical clustering of the perturbed data.** Multiple clusterings are obtained by applying the same hierarchical clustering algorithm as in step (1) to the perturbed data.

4. **Construction of the similarity matrix.** A similarity matrix that stores the frequency by which each pair of examples falls into the same cluster in the "perturbed" clustering is built (Dudoit and Fridlyand, 2003).

5. **Computation of the stability indices.** For each cluster obtained through the hierarchical clustering of the original data (step 1), a stability index (Bertoni and Valentini, 2006) is computed using the similarity matrix constructed at step 4.

6. **Selection of the most reliable clusters.** Using the stability indices computed in the previous step, the most reliable clusters are selected. Several approaches can be used; the easiest one consists in the selection of the clusters whose stability is above a given threshold.

More precisely, given a data set $D = \{\boldsymbol{x}_i \in \mathbb{R}^r, 1 \leq i \leq N\}$, a *clustering algorithm* $\mathcal{C}(D, k)$ is a procedure that, having as input a data set D and an integer k, outputs a k-clustering $C =< A_1, A_2, \ldots, A_k >$ on the basis of the distances $||\boldsymbol{x}_i - \boldsymbol{x}_j||$, $(1 \leq i, j \leq N)$. According to (Bertoni and Valentini, 2006) we can can associate a $N \times N$ similarity matrix M to a k-clustering; the elements $M(i,j)$ of M are defined as:

$$M(i,j) = \sum_{s=1}^{k} \chi_{A_s}[i] \cdot \chi_{A_s}[j] \tag{1}$$

where $i, j \in \{1, 2, \ldots, N\}$, and $\chi_{A_s} \in \{0,1\}^N$ is the characteristic vector of A_s, i.e. $\chi_{A_s}[i] = 1$ if $\boldsymbol{x}_i \in A_s$, otherwise $\chi_{A_s}[i] = 0$.

By applying multiple perturbations to the data through a randomized map $\mu : \mathbb{R}^r \rightarrow \mathbb{R}^m, m < r$, and by averaging the similarity matrices obtained from the application of a clustering algorithm $\mathcal{C}$ to the resulting projected data, we can compute the the following *stability index* s for a cluster A (Bertoni and Valentini, 2006):

$$s(A) = \frac{1}{|A|(|A|-1)} \sum_{\{(i,j)|\boldsymbol{x}_i \in A \wedge \boldsymbol{x}_j \in A, i \neq j\}} M(i,j) \tag{2}$$

The index $s(A)$ estimates the stability of a cluster A by measuring how much the projections of the pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j) \in A$ occur together in the same cluster in the projected subspaces.

An example of randomized map that realizes a dimensionality reduction from a d to a d'-dimensional space, $d' < d$, is the *Bernoulli* random projection $\mu(\boldsymbol{x}) = 1/\sqrt{d'}R * \boldsymbol{x}$ (Achlioptas, 2003). It is a randomized linear map represented through a $d' \times d$ random matrix R, whose elements $R_{ij} \in \{-1, 1\}$, are instances of Bernoulli random variables such that $Prob(R_{i,j} = 1) = Prob(R_{i,j} = -1) = 1/2$.

Using the stability index defined in eq. 2, the pseudo-code of the stability-based algorithm for finding reliable clusters in a given hierarchical clustering is the following:

```
Cluster stability algorithm:
Input:
- A data set $D = \{\boldsymbol{x}_i \in \mathbb{R}^r, 1 \leq i \leq N\}$.
- A hierarchical clustering algorithm $\mathcal{C}$.
- A number $n$ of perturbations of the data.
- A procedure that realizes a randomized map $\mu : \mathbb{R}^r \rightarrow \mathbb{R}^m, m < r$.
Begin algorithm
    (1) $\{A_1, \ldots, A_{2N-1}\} := \mathcal{C}(D)$;
    (2) $C := \{A_i | A_i$ is not a leaf or the root$\}$;
    (3) $M := 0$ ;
    (4) $d := 0$;
    Repeat for $j = 1$ to $n$
        (5) $D^j := \mu(D)$;
        (6) $\{B_1^j, \ldots, B_{2N-1}^j\} := \mathcal{C}(D^j)$;
        (7) $C^j := \{B_i^j | B_i^j$ is not a leaf or the root $\}$;
        (8) $d := d +$ depth $(\mathcal{C}(D^j)) - 1$;
        For each $B_k^j \in C^j$
            For each $(\boldsymbol{x}_t, \boldsymbol{x}_v) \in (B_k^j \times B_k^j)$
                (9) $M(t, v) := M(t, v) + 1$;
            end For
        end For
    end Repeat
    (10) $M := \frac{M}{d}$;
    For each $A_k \in C$
        (11) $s(A_k) := \frac{1}{|A_k|(|A_k|-1)} \sum_{(\boldsymbol{x}_t, \boldsymbol{x}_v) \in A_k \times A_k} M(t, v)$;
    end For
end algorithm.
Output:
- $S = \{s(A_i) | A_i \in C\}$.
```

Note that with abuse of notation we represent clusters and nodes with the same symbols, as well as dendrograms and corresponding clusterings. At line (2), from the original hierarchical clustering composed by $2N - 1$ clusters (line (1)), only the internal $N - 2$ nodes are selected. Indeed it is easy to see that all the singleton clusters (the leaves of the dendrogram) and the "root" cluster are always present in any hierarchical clustering and as a consequence their stability is always 1 (maximum stability).

The core of the algorithm is represented by the `Repeat` loop. At each iteration we obtain an instance of the perturbed (projected) data (step 5); then a hierarchical clustering algorithm is applied to the perturbed data, considering only the internal nodes (steps $6-7$). After updating the cumulative depth of the n dendrograms (8), the two nested iterative loops update the similarity matrix M, by adding 1 to the entry $M(t,v)$ if the examples $\boldsymbol{x}_t$ and $\boldsymbol{x}_v$ are both present in the cluster B_k^j (step 9). To maintain the value of each entry of the matrix M between 0 and 1 we need to normalize it by d (step 10). Indeed each pair of examples may belong to a number of clusters equal at most to the depth minus one of the corresponding tree (step 8). The output of the algorithm consists in the set of stability indices computed for each node of the hierarchical clustering C.

3 Results and discussion

We present the results obtained by applying the proposed algorithm to gene expression data collected during a study on leukemia. In particular sixteen samples were available, including fourteen patients affected by Human Myeloid Leukemia at diagnosis and two healthy donors as control. Samples were analyzed using Affymetrix Human Genome U133 Plus 2.0 arrays. Each gene on this chip is represented by 11 oligonucleotides, termed a “probe set”. This type of array contains 54675 probe sets and it analyzes the expression level of 47400 transcripts and variants including 38500 UniGene clusters at the time of array design.

During the laboratory procedures biotin-labeled RNA fragments are hybridized to the probe array. The hybridized probe array is stained with streptavidin phycoerythrin conjugated and scanned by the GeneChip Scanner 3000. From the image files, .cel files containing a single intensity value for each probe cell delineated by the grid are obtained. We used Bioconductor (Gentleman *et al.*, 2004) packages to assess data quality, using standard Affymetrix tests, as well as other quality check tests such as the Relative Log Expression (RLE) plot and Normalization Unscaled Standard Error (NUSE) (Irizarry *et al.*, 2003). All checks assured the high quality of the gene expression data.

Background correction, normalization and summarization were performed using the Robust Multi-array Average (RMA) procedure that summarizes probe level data to obtain gene expression levels (Irizarry *et al.*, 2003).

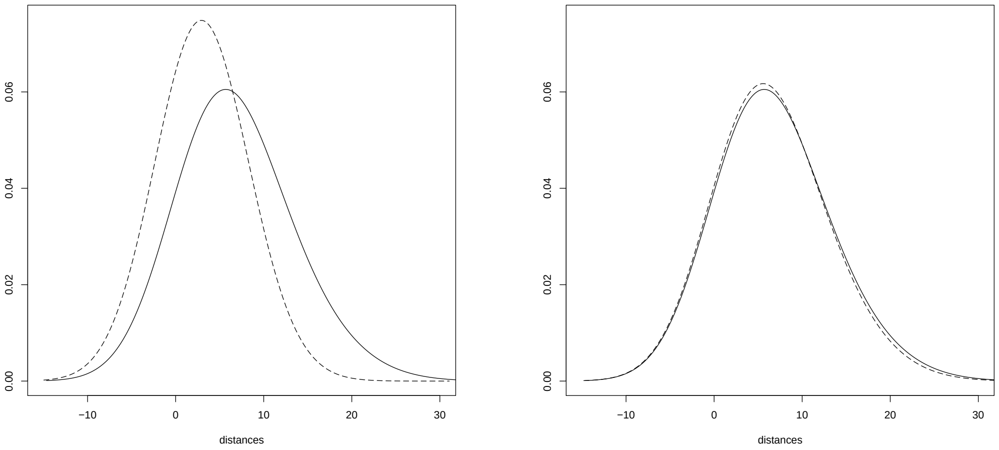
To reduce the high number of probe sets (54613 probe sets with the exclusion of the Affymetrix chip control probes), we used a t-test to select differentially expressed probe sets in patients with respect to controls. At a 0.01 significance level we selected 1038 probe sets. For clustering analysis we considered relative expression levels in the 14 patients with respect to the average value in the two controls. As we are dealing with logarithmic scale values, this corresponds to subtracting from the expression level of each probe set in a certain patient the mean expression level of the same gene in the two controls.

Using the algorithm described in Sect. 2 and the standard average-linkage algorithm with Euclidean distance to perform the hierarchical clusterings, we iterated 50 random projections from the original 14-dimensional space to a lower 10-dimensional space, using Bernoulli random projections (Bertoni and Valentini, 2007).

In this experimental setting we cannot apply the Johnson-Lindestrauss lemma (Johnson and Lindenstrauss, 1984) to directly compute the dimension m of the projected subspace:

$$m = c \ \log N/\epsilon^2 \tag{3}$$

where c is a suitable constant, N the cardinality of the available data and ϵ the desired upper bound to the distortion induced by the randomized projection. Indeed in our experiments $N = 1038$ and by setting $c = 4$ and a 20% distortion ($\epsilon = 0.2$) we should project to a 302-dimensional subspace, even larger than the original 14-dimensional space. Considering that in this experimental setting the theoretical bounds provided by the Johnson-Lindestrauss lemma are in practice unuseful, we empirically estimated the distortion induced by the Bernoulli random projections into the analyzed gene expression data. We chose 10-dimensional Bernoulli random mappings, because the distributions of the pairwise distances between genes in the original and in the projected 10-dimensional space are very similar, while projections into lower dimensional subspaces may induce relevant metric distortions (Fig. 1).



(a) (b)

Figure 1 Distribution of the pairwise euclidean distances between gene expression levels in the original 14-dimensional space (continuous line) and distribution of the pairwise distances in the projected subspace (dashed line). a) Bernoulli projection into a 3-dimensional subspace b) Bernoulli projection into a 10-dimensional subspace

Results are shown in Table 1. Different thresholds $0 < \alpha < 1$ were considered, in order to select the set R_α of reliable clusters, among those belonging to the clustering C in the original space:

$$R_\alpha = \{A_i \in C | s(A_i) > \alpha\}$$

The last column represents the ratio values with respect to the total number of clusters (1036), obtained excluding the singleton and the “root” clusters. From these results we may observe that 180 clusters show a stability larger than 0.5 and only 29 larger than 0.6. Functional enrichment was performed on the 48 most

Table 1 Number of clusters of the original hierarchical classification with a stability larger than α. The last column represents the ratio of the number of the selected clusters with respect to the total number of clusters.

α	**Number of clusters**	**Ratio**
0.1	1036	1
0.2	1018	0.983
0.3	889	0.858
0.4	536	0.517
0.5	180	0.174
0.6	29	0.028
0.7	3	0.003
0.8	0	0
0.9	0	0

stable clusters (we considered a threshold value slightly lower than 0.6 to work with a reasonable number of clusters) by using the Bioconductor package *GOStats* (rel. 2.8.0) (Falcon and Gentleman, 2007). This package relies on a hypergeometric test to find Gene Ontology biological processes that are over-represented in a given cluster with respect to a chosen background. In our case we employed the entire set of genes assayed on hgu133plus2.0 (based on the hgu133plus2.db annotation package rel. 2.2.5 (Carlson *et al.*, 2008a)) as background and we set the significance level to 0.01 to identify enriched biological processes in a given cluster.

Functional enrichment allows finding whether one or more functional classes (e.g. Gene Ontology terms or KEGG pathways) are significantly over-represented among the relevant genes selected in the experiment (Khatri and Draghici, 2005; Dopazo, 2006). Through functional enrichment it is possible to assign a putative function to unknown genes contained in a cluster, which can be confirmed with further extended biological validation.

We considered only genes with at least one GO annotation in the "Biological Process" ontology. 10 out of 48 clusters were enriched for a GO term represented by at least two genes in the cluster (Table 2). We used the Bioconductor package "org.Hs.eg.db" release 2.2.6 (see (Carlson *et al.*, 2008b)) to perform mappings from GeneIds to the related gene names and gene symbols. All the genes belonging to the 10 clusters were underexpressed with respect to the mean values of the controls (results not shown).

A further characterization of the clusters can be obtained by evaluating the overall "variability" of the probe sets contained in the clusters, represented by the median standard deviation of the profiles in a cluster (Table 3). This measurement allows us to distinguish between clusters formed by probe sets whose behavior does not vary across the different patients and clusters with probe sets behaving differently in the various patient samples. The biological processes enriched in the former set of clusters might be associated with Myeloid Leukemic development, irrespective of different tumor subclasses and might thus give us insights on the most important dysregulated processes in disease.

A preliminary biological analysis performed on the results showed in Table 2 indicates that the gene clustering proposed didn't show any specific "molecular"

Table 2 GO terms of the "Biological Process" ontology overrepresented in the discovered 48 most stable clusters. The first column reports the genes of the discovered stable clusters that belong to the enriched GO classes. The second column reports the enriched GO identifiers and the corresponding p-values.

Genes	**GOID (p-value)**
GJA4,LAMA4	GO:0007275 (0.0398)
CLEC7A,CD163	GO:0006950 (0.0342), GO:0006952 (0.0117), GO:0006954 (0.0057), GO:0009605 (0.022), GO:0009611 (0.0113)
GNAZ,IGF1	GO:0007166 (0.0226)
GJA4,LAMA4,SNIP	GO:0051179 (0.0193)
GRHL1,RGMA	GO:0007275 (0.0398)
SLC0B1,IGHM	GO:0006810 (0.0451), GO:0051234 (0.0474)
FPR3,IGF1	GO:0007166 (0.0226), GO:0006928 (0.0051), GO:0051674 (0.0051), GO:0009605 (0.022)
FABP4,APOE	GO:0006139 (0.0474), GO:0055088 (6.6477e-05), GO:0031323 (0.0413), GO:0048878 (0.0018), GO:0050790 (0.0021), GO:0019222 (0.0413), GO:0051338 (0.0007), GO:0042632 (3.9886e-05), GO:0048519 (0.0146), GO:0031347 (9.9716e-05), GO:0033673 (6.6477e-05), GO:0043549 (0.0007), GO:0060255 (0.0342), GO:0043086 (9.9716e-05), GO:0065009 (0.0025), GO:0051234 (0.0474), GO:0009889 (0.029), GO:0048523 (0.0125), GO:0006954 (0.0057), GO:0045859 (0.0007), GO:0019219 (0.026), GO:0006810 (0.0451), GO:0048583 (0.0005), GO:0006950 (0.0342), GO:0048518 (0.0155), GO:0009059 (0.0405), GO:0055092 (3.9886e-05), GO:0032101 (3.9886e-05), GO:0009611 (0.0113), GO:0006952 (0.0117), GO:0042592 (0.0026), GO:0050727 (1.9943e-05), GO:0051348 (6.6477e-05), GO:0006469 (6.6477e-05), GO:0065008 (0.0142), GO:0009605 (0.022)
NLRP3,NR4A3	GO:0006139 (0.0474), GO:0010468 (0.0278), GO:0019219 (0.026), GO:0031323 (0.0413), GO:0019222 (0.0413), GO:0043284 (0.0284), GO:0009059 (0.0405), GO:0060255 (0.0342), GO:0006350 (0.0254),GO:0009889 (0.029), GO:0045449 (0.0226), GO:0010467 (0.0405), GO:0010556 (0.0278)
APOE, PLA2G7	GO:0009056 (0.0021), GO:0006954 (0.0057), GO:0016042 (0.0002), GO:0006950 (0.0342), GO:0009611 (0.0113), GO:0006952 (0.0117), GO:0006629 (0.0044), GO:0009605 (0.022)

association, but the overall gene selection evidenced a deregulation of extracellular matrix interactions and adhesion. Of particular interest the LAMA4 gene that encodes the alpha chain isoform laminin, alpha 4. Laminin, alpha 4 contains the C-terminal G domain which distinguishes all alpha chains from the beta and gamma chains. RNA analysis from adult and fetal tissues revealed developmental regulation of expression, however, the exact function of laminin, alpha 4 is not known (Jaluria *et al.*, 2007). The results of this study are consistent with the role LAMA4 plays in adhesion processes in vivo and indicate that modifying the expression of the

Table 3 "Variability" of the analyzed clusters as represented by the median standard deviation of the profiles contained in each cluster.

GO enriched genes in the clusters	Median st. dev.
GJA4,LAMA4	0.1449598
CLEC7A,CD163	0.1903084
GNAZ,IGF1	0.1650183
GJA4,LAMA4,SNIP	0.1326144
GRHL1,RGMA	0.1947557
SLC0B1,IGHM	0.1970877
FPR3,IGF1	0.3379482
FABP4,APOE	0.5194621
NLRP3,NR4A3	0.4548647
APOE, PLA2G7	0.3309572

gene can influence adhesion of AC113+ cells. By reducing the expression of LAMA4 in a cell model, a reduction in cellular adhesion was observed. Thus, changes of the expression levels of LAMA4 are consistent with the evolution of different adhesion properties for the cells evaluated in the current study. The association of LAMA4 with GJA4 is of interest as the human gene encoding connexin37 (encoded by GJA4, also known as CX37), is also involved in monocyte adhesion regulation in bone marrow (Wong *et al.*, 2006). Moreover, the observation of IGF1 downregulation in the patient samples is of particular interest for Imatinib-related treatment implications. Imatinib (imatinib mesylate, STI-571, Gleevec) is a selective tyrosine kinase inhibitor that has been successfully used to treat chronic myeloid leukemia (CML). However, relapse after the initial hematologic and cytogenetic response frequently occurred in late-stage disease. Heterogeneous mechanisms might be responsible for imatinib-resistance. It has been demonstrated that IGF1 showed consistent downregulation after the acquisition of imatinib-resistance (Chung *et al.*, 2006).

Despite the biological insights obtained from this analysis, the proposed approach shows some limitations that need to be considered for future work. For instance, the algorithm has a bias versus very low sized and very large sized clusters. Indeed it is easy to see that singleton clusters and the cluster that contains all the examples are always present in every hierarchical clustering algorithm, thus resulting in a stability equal to 1. All the other clusters lie somewhere in between: hence it is necessary to include a proper correction with respect to the cluster size. Another relevant problem, related to the previous one, is the choice of the threshold α to select the significant clusters. From a general standpoint, larger values of α assure a high precision in identifying stable and significant clusters, even if at a cost of a likely lower sensitivity, while the opposite is true with lower values of α. By varying α we could tune the trade-off between sensitivity and precision, but a weakness of the proposed approach is the lack of a fully automated and principled method to set an "optimal" value for α to discover the significant set of clusters. Finally, the choice of classical hierarchical algorithms to discover the clusters of genes may represent another limitation. Even if clusters of genes may show a hierarchical structure, a gene may belong to multiple nodes in different non-nested subtrees of the hierarchical structure, and classical hierarchical clustering algorithms cannot capture these characteristics of the data. To this end a possibly more consistent approach could be a fuzzy or probabilistic hierarchical clustering approach, in order to address the problem of "not-hierarchically-related" clusters.

4 Conclusions and future work

We presented an algorithm to discover reliable clusters in hierarchical clusterings characterized by a large number of examples and clusters, a situation in which classical stability-based methods are not applicable for computational complexity reasons.

The method proposes a stability-based approach that uses multiple randomized projections of the original data and a stability measure constructed through a similarity matrix that summarizes multiple clusterings on the perturbed data. A preliminary application to patients affected by Human Myeloid Leukemia discovered a small number of gene clusters that were analyzed by means of Gene Ontology based functional enrichment.

In future works we will address the problem of the bias of the stability measure with respect to the cardinality of the clusters, and we will also define a principled method to choose the threshold to select the set of significant clusters. To this end, we are working on a non-parametric statistical test to solve both these open problems.

From a biological standpoint, we will extend the analysis to a larger number of patients and healthy samples, which are currently being collected. This will allow us to perform a more reliable analysis. The selection of the differentially expressed genes is significantly impaired when the number of samples is low and especially when the number of samples in the two groups being compared is unbalanced, as in our case. Moreover, a larger number of samples can lead to more robust and reliable gene clusters.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments and suggestions, and gratefully acknowledge partial support by the PASCAL2 Network of Excellence under EC grant no. 216886. This publication only reflects the authors' views. Fulvia Ferrazzi was supported by an Investigator Fellowship from Collegio Ghislieri, Pavia, Italy.

References

Achlioptas, D. (2003). Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of Comp. & Sys. Sci.*, **66**(4), 671–687.

Ben-Hur, A., Ellisseeff, A., and Guyon, I. (2002). A stability based method for discovering structure in clustered data. In R. Altman, A. Dunker, L. Hunter,

T. Klein, and K. Lauderdale, editors, *Pacific Symposium on Biocomputing*, volume 7, pages 6–17, Lihue, Hawaii, USA. World Scientific.

Bertoni, A. and Valentini, G. (2006). Randomized maps for assessing the reliability of patients clusters in DNA microarray data analyses. *Artificial Intelligence in Medicine*, **37**(2), 85–109.

Bertoni, A. and Valentini, G. (2007). Model order selection for bio-molecular data clustering. *BMC Bioinformatics*, **8**(S3).

Bertoni, A. and Valentini, G. (2008). Discovering multi-level structures in bio-molecular data through the Bernstein inequality. *BMC Bioinformatics*, **9**(S2).

Bolshakova, N., Azuaje, F., and Cunningham, P. (2005). An integrated tool for microarray data clustering and cluster validity assessment. *Bioinformatics*, **21**(4), 451–455.

Carlson, M., Falcon, S., Pages, H., and Li, N. (2008a). Technical report, Bioconductor. http://www.bioconductor.org/packages/release/data/annotation/html/hgu133plus2.db.html.

Carlson, M., Falcon, S., Pages, H., and Li, N. (2008b). Technical report, Bioconductor. http://inn.weizmann.ac.il/bioconductor/packages/2.2/data/annotation/html/org.Hs.eg.db.html.

Chung, Y.-J., Kim, T.-M., Kim, D.-W., Namkoong, H., Kim, H., Ha, S.-A., Kim, S., Shin, S., Kim, J.-H., Lee, Y.-J., Kang, H.-M., and Kim, J. (2006). Gene expression signatures associated with the resistance to imatinib. *Leukemia*, **20**, 1542–1550.

Datta, S. and Datta, S. (2003). Comparison and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, **19**, 459–466.

Dopazo, J. (2006). Functional interpretation of microarray experiments. *OMICS*, **3**(10).

Dudoit, S. and Fridlyand, J. (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, **19**(9), 1090–1099.

Efron, B. and Tibshirani, R. (1993). *An introduction to the Bootstrap*. Chapman and Hall, New York.

Falcon, S. and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**.

Gentleman, R. *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, **5**(10).

Handl, J., Knowles, J., and Kell, D. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, **21**(15), 3201–3215.

Irizarry, R., B., H., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **2**, 249–264.

Jaluria, P., Betenbaugh, M., Konstantopoulos, K., Frank, B., and Shiloach, J. (2007). Application of microarrays to identify and characterize genes involved in attachment dependence in HeLa cells. *Metabolic Engineering*, **9**, 241–251.

Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*, **16**(11), 1370–1386.

Johnson, W. and Lindenstrauss, J. (1984). Extensions of Lipshitz mapping into Hilbert space. In *Conference in modern analysis and probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. Amer. Math. Soc.

Kerr, M. and Curchill, G. (2001). Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *PNAS*, **98**, 8961–8965.

Khatri, P. and Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

McShane, L., Radmacher, D., Freidlin, B., Yu, R., Li, M., and Simon, R. (2002). Method for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, **18**(11), 1462–1469.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus Clustering: A Resampling-based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*, **52**, 91–118.

Napolitano, F., Raiconi, G., Tagliaferri, R., Ciaramella, A., Staiano, A., and Miele, G. (2008). Clustering and visualization approaches for human cell cycle gene expression data analysis. *Int. J. Approx. Reasoning*, **47**(1), 70–84.

Smolkin, M. and Gosh, D. (2003). Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics*, **36**(4).

The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.

Wong, C., Christen, T., Roth, I., Chadjichristos, C., J.C., D., Foglia, B., Chanson, M., Goodenough, D., and Kwak, B. (2006). Connexin37 protects against atherosclerosis by regulating monocyte adhesion. *Natural Medicine*, **12**, 950–954.