An algorithm to assess the reliability of hierarchical clusters in gene expression data

Roberto Avogadri[1], Matteo Brioschi[2], Francesca Ruffino[1], Fulvia Ferrazzi[3], Alessandro Beghini[2] and Giorgio Valentini[1]

[1] DSI - Dip. Scienze dell' Informazione, Università degli Studi di Milano, Italy
{avogadri,ruffino,valentini}@dsi.unimi.it
[2] DBioGen - Dip. Biologia e Genetica per le Scienze Mediche, Università degli Studi di Milano, Italy
{matteo.brioschi,alessandro.beghini}@unimi.it
[3] Dip. Informatica e Sistemistica, Università degli Studi di Pavia, Italy
fulvia.ferrazzi@unipv.it

Abstract. The validation of clusters discovered in bio-molecular data is a central issue in bioinformatics. Recently, stability-based methods have been successfully applied to the analysis of the reliability of clusterings characterized by a relatively low number of examples and clusters. Nevertheless, several problems in functional genomics are characterized by a very large number of examples and clusters. We present a stability-based algorithm to discover significant clusters in hierarchical clusterings with a large number of examples and clusters. Preliminary results on gene expression data of patients affected by Human Myeloid Leukemia, show how to apply the proposed method when thousands of gene clusters are involved.

1 Introduction

The unsupervised discovery and validation of clusters underlying data is a central issue in several branches of bioinformatics [1], as well as the proper visualization of clustering results [2]. Different clustering validation techniques (see [3] for a recent review), and software tools implementing classical validity indices (such as the *Dunn's index* and the *Silhouette index*) have been proposed [4].

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 [5, 6, 7, 8]. Despite their successful application in several bioinformatics domains, they are well-suited to unsupervised problems characterized by a relatively low number of clusters and/or examples [9, 10]. 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 (2^N) number of potential clusters.

Considering that clusters of genes may show a hierarchical multi-level organization [11], 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 mutuated from our previous work [10]. Differently from our previous approach, we do not need to know in advance the correct or the 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 the next section 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 hgu133plus2 GeneChip. Then we discuss the advantages and the limitations of the proposed method. In the conclusions we propose some research lines for future work.

2 The algorithm

Our algorithm is founded 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 [12], by subsampling or bootstrapping procedures [13], 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 the step (1) to the perturbated 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 [14].
5. **Computation of the stability indices.** For each cluster obtained through the hierarchical clustering of the original data (step 1), a stability index [10] is computed using the similarity matrix constructed at step 4. The stability index S (see line (11) of the pseudo-code of the algorithm) has values between 0 (low stability) and 1 (high stability).
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.

The pseudo-code of the algorithm is reported below:

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 on perturbations of the data.
- A procedure that realizes a randomized map $\mu : \mathbb{R}^r \to \mathbb{R}^m, m < r$.

`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`:
- $S = \{s(A_i) | A_i \in C\}$.

In this algorithm a randomized map is applied to perturb the data. 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 following 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. 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

As an example of application of the proposed algorithm, we analyzed gene expression data of eight samples, including seven patients affected by Human Myeloid Leukemia at diagnosis and one healthy donor as control. Samples were analyzed using the Affymetrix hgu133plus2 GeneChip. Each gene on this chip is represented by 11 oligonucleotides, termed a "probeset". The hgu133plus2 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 $Affymetrix$. From the image files .cel files containing a single intensity value for each probe cell delineated by the grid are obtained. We used Bioconductor [15] packages to assess the quality level of the data, 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) [16]. Fig. 1 shows the MA plots of the expression levels of the seven samples using the healthy donor as reference. All quality checks assured the high quality of the gene expression data. Background correction, normalization and summarization have been performed using the Robust Multi-array Average (RMA) procedure that summarizes the probe level data to obtain gene expression levels [16] and the "expresso" method from the *Affy* Bioconductor package [17].

To reduce the high number of examples (54613 probe sets with the exclusion of the Affymetrix chip control probes), we used a z-test to select the genes whose gene expression levels significantly differ from the healthy donor control patient. At a 0.1 significance level we selected 1007 genes. Using the algorithm described in Sect. 2 and the classical average-linkage algorithm to perform the hierarchical clusterings, we iterated 50 random projections from the original 7-dimensional space to a lower 5-dimensional space, using *Bernoulli* random projections [11].

The results are showed in Table 1. Different thresholds $0 < \alpha < 1$ have been 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 (1005), obtained excluding the singleton and the "root" clusters. From

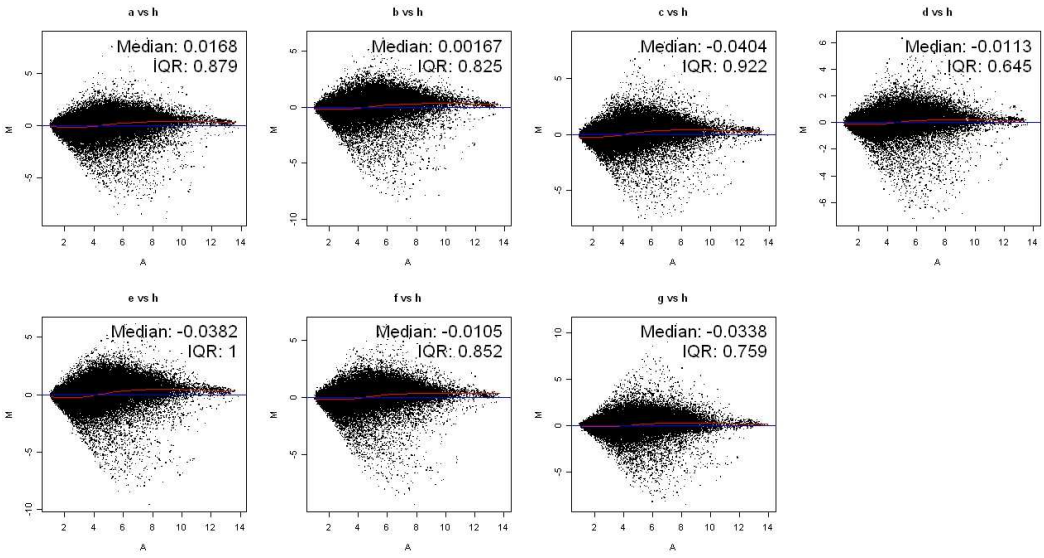

Fig. 1. MA plots of the seven patients affected by Human Myeloid Leukemia using the healthy donor as reference.

these results we may observe that 43 clusters show a stability larger than 0.8, and only 5 clusters show a very high reliability (stability larger than 0.9).

The proposed approach shows several limitations that need to be carefully 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 clusters 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 in between: hence it is necessary to include a proper correction with respect to the cluster size.

Another relevant problem is the choice of the threshold α to select the significant clusters. In the proposed algorithm the choice is somehow arbitrary: we considered very reliable the 5 clusters selected with a threshold equal to 0.9, but there is no reason to consider this threshold as a warranty of reliability. Moreover this problem is related to the previous one, because the threshold should be related to the cardinality of the clusters.

The choice of classical hierarchical algorithms to discover the clusters of genes may represents another limitation. Even if clusters of genes may show a hierarchical structure [18], 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 possible more consistent approach could be a fuzzy or probabilistic hierarchical cluster-

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

α	**Number of clusters**	**Ratio**
0.1	1004	0.999
0.2	919	0.914
0.3	680	0.677
0.4	392	0.390
0.5	227	0.226
0.6	138	0.137
0.7	76	0.076
0.8	43	0.043
0.9	5	0.005

ing approach, in order to address the problem of "not-hierarchically-related" clusters.

From a bioinformatics standpoint we need also to biologically validate the clusters discovered as reliable by the proposed method. To this end we need a careful biological and bio-medical analysis of the clusters of genes individuated as significant. To support this bio-medical task functional enrichment methods are often used to find if one or more of gene modules (e.g. Gene Ontology classes or KEGG pathways) are significantly over-represented among the relevant genes selected in the experiment [19, 20]. Over-representation of a given gene module means that genes with a particular property have been activated or deactivated in the experiment.

4 Conclusions

We presented an algorithm to discover reliable clusters in hierarchical clusterings characterized by a large number of examples and clusters. 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 relatively small number of gene clusters that need to be biologically validated. The aim of this preliminary work consists in showing the applicability of a stability-based method to discover significant clusters when their number is relatively high and classical stability-based methods are not applicable for computational complexity reasons. Nevertheless in future works we need to address the problem of the bias of the stability measure and we need also a principled method to choose the threshold to select the set of significant clusters. We are working on a non-parametric statistical test to solve both these open problems.

References

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

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

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

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

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

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

[7] Ben-Hur, A., Ellisseeff, A., Guyon, I.: A stability based method for discovering structure in clustered data. In Altman, R., Dunker, A., Hunter, L., Klein, T., Lauderdale, K., eds.: Pacific Symposium on Biocomputing. Volume 7., Lihue, Hawaii, USA, World Scientific (2002) 6–17

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

[9] Smolkin, M., Gosh, D.: Cluster stability scores for microarray data in cancer studies. BMC Bioinformatics **36** (2003)

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

[11] Bertoni, A., Valentini, G.: Model order selection for bio-molecular data clustering. BMC Bioinformatics **8** (2007)

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

[13] Efron, B., Tibshirani, R.: An introduction to the Bootstrap. Chapman and Hall, New York (1993)

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

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

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

[17] Gautier, L., Cope, L., Bolstad, B., Irizarry, R.: Affy–analysis of affymetrix genechip data at the probe level. Bioinformatics **20** (2004) 307–315

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

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

[20] Dopazo, J.: Functional interpretation of microarray experiments. OMICS **3** (2006)