Methods to Bicluster Validation and Comparison in Microarray Data

Rodrigo Santamaría, Luis Quintales, and Roberto Therón

University of Salamanca

Abstract. There are lots of validation indexes and techniques to study clustering results. Biclustering algorithms have been applied in Systems Biology, principally in DNA Microarray analysis, for the last years, with great success. Nowadays, there is a big set of biclustering algorithms each one based in different concepts, but there are few intercomparisons that measure their performance. We review and present here some numerical measures, new and evolved from traditional clustering validation techniques, to allow comparisons and validation of biclustering algorithms.

1 Introduction

Biclustering is one of the main options to find structure in gene microarray data. In the last years, lots of biclustering methods have been proposed [10]. Authors apply different procedures to individually validate them. Also, with the growing number of algorithms, its comparison is now being addressed [12]. Though not an optimal algorithm exists, these comparisons help to understand biclustering behavior and make easier the choice of the bests algorithms in each context.

Several measures for validation exist in clustering area, but they are usually not applied for biclustering methods. The authors that have treated more in deepness comparison methodologies for biclustering are Prelic et al. [12] and Turner et al. [14]. Validation and comparison are made by external indices. Non-biological indices as sensitivity and specificity are used when information of clustering is known, usually in synthetic data where biclusters are embedded. Only constant and additive biclusters are treated, as they are the most extended. Biological indices are used when no information intrinsic to the data is known. Internal and relative indices are seldom used because biclustering concepts are hard to adapt to clustering indices.

In this paper, we review these validation and comparison techniques, explaining the adaptations done in literature and proposing some other adaptations to biclustering characteristics. Specially, internal and relative index application to optimize input parameters and coherence measures have been developed. In Section 2, we discuss the different kinds of biclusters offering measures to determine each type. Section 3 covers the use of internal, external and relative indices, reviewing the most used and extending some of them to biclustering context. Section 4 makes a brief application of measures discussed in Section 2 and 3 on two biclustering algorithms. Finally, Section 5 presents the conclusions and future work.

H. Yin et al. (Eds.): IDEAL 2007, LNCS 4881, pp. 780–789, 2007.

2 Bicluster Structure

2.1 Bicluster Classification

A bicluster can be defined as 'a subset of objects (rows or columns) that jointly respond across a subset of other objects (columns or rows)'. In bioinformatics, rows usually refer to genes and columns to experiments or organism conditions. Madeira and Oliveira [10] classify biclusters depending on what is considered for 'jointly responds':

- Constant value bicluster (C): all elements have exactly the same value (μ). Elements of constant bicluster $B = [b_{ij}]$ with n rows and m columns are defined as

$$b_{ij} = \mu \tag{1}$$

- Coherent value bicluster (H): row and/or column variations are somehow related. This relationship may be additive (H^+), multiplicative ($H^\times$) or by sign ($H^\pm$). In case of H^+ and $H^\times$, each row and/or column differs from others in an additive or multiplicative factor (eqs. 2 and 3, respectively). In case of $H^\pm$, it is just a qualitative rule of change in tendency (α and β are binary vectors representing increasing or decreasing respect to another row or column –such as 1 or -1–, but it's not imposed any quantitative restriction on r_{ij}, c_{ij} variations)

$$b_{ij} = \mu + \alpha_i + \beta_j \tag{2}$$

$$b_{ij} = \mu \alpha_i \beta_j \tag{3}$$

$$b_{ij} = (b_{(i-1,j)} + \alpha_i r_{ij}) + (b_{(i,j-1)} + c_{ij}\beta_j) \tag{4}$$

- Coherent evolution bicluster (E): expression levels are first mapped to labels under certain criteria, such as order or proximity.

The above definitions can be applied to rows, columns or both, but measures are usually used in both dimesions. C biclusters are almost ideal, so algorithms searching for C biclusters usually treats 'constant' as a range of near values by a mapping with coherence evolution.

This bicluster classification presents overlaps. For example, C biclusters on rows and columns (C_{rc}) are included in C biclusters on rows (C_r)and C biclusters on columns (C_c). C biclusters of any type are included in H^+ biclusters and overlap with $H^\times$ biclusters. $H^\pm$ includes them all (Fig. 1). This will be important when comparing biclustering algorithms that search for different kinds of biclusters.

C is the most used group because of direct interpretation in biological data. H^+ biclusters, representing more subtle relations in data are the second group in references. $H^\times$ and $H^\pm$ are rarely used, being their biological relevance difficult to justify or interpret.

2.2 Coherence Measures

Having in mind the different groups of biclusters, we can define measures that determine how constant or how (additive, multiplicative, sign) coherent is our bicluster.

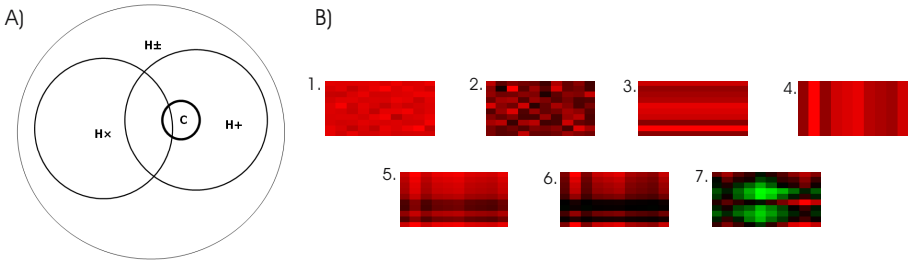

Fig. 1. A) Bicluster sets. Each of the sets is internally divided in row, column and both dimensions biclusters of the corresponding type. B) Heatmaps of different biclusters: 1) C_{rc} bicluster, 2) C_{rc} bicluster with high noise, 3) C_r bicluster, 4) C_c column constant, 5) H^+ bicluster, 6) $H^\times$ bicluster and 7) $H^\pm$ bicluster. 5),6) and 7) become, after row/column transformation, C_r and/or C_c biclusters 3) and 4).

Biclustering algorithms define internally what is considered coherent, but not always under an specific measure or value. Coherence measures can be used to define synthetic biclusters for testing or to check if the results over real data fits the bicluster definition of the algorithm. Constancy by rows of bicluster B ($C_r(B)$) and by columns ($C_c(B)$) are easy to measure by means of Euclidean distance

$$C_r(B) = \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sqrt{\sum_{k=1}^{m}(b_{ik}-b_{jk})^2} \tag{5}$$

$$C_c(B) = \frac{1}{m}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\sqrt{\sum_{k=1}^{n}(b_{ki}-b_{kj})^2} \tag{6}$$

Overall constancy $C_{rc}(B)$ can be derived from $C_r(B)$ and $C_c(B)$:

$$C_{rc}(B) = \frac{nC_r(B)+mC_c(B)}{n+m} \tag{7}$$

The average measure for all the biclusters found by an algorithm is the weighted mean of the measure for each bicluster. These measures, traditionally used to determine cluster compactness will give bad scores for coherent biclusters. To measure coherency, an incremental treatment of the data can be applied to make them 'constant', then applying above formulas to the transformed bicluster $B' = [b'_{ij}]$. In case of H^+:

$$b'_{ij} = b_{ij} - b_{(i-1)j}\ (b_{0j} = 0) \tag{8}$$

$$b'_{ij} = b_{ij} - b_{i(j-1)}\ (b_{i0} = 0) \tag{9}$$

That way, as seen in Fig. 1b, H^+ bicluster becomes C_r and/or C_c bicluster, and can be measured by eqs. 5, 6 and 7. A similar transform can be done with $H^\times$ using division instead of substraction, but now there is necessary to include an exception to avoid divisions by zero:

$$b'_{ij} = b_{ij}/b_{(i-1)j}\ (b_{0j} = 1) \tag{10}$$

$$b'_{ij} = b_{ij}/b_{i(j-1)}\ (b_{i0} = 1) \tag{11}$$

Finally, $H^{\pm}$ has a similar treatment:

$$b'_{ij} = 1 \Leftrightarrow b_{ij} > b_{(i-1)j},\ b'_{ij} = -1\ otherwise.\ (b_{0j} = 1) \tag{12}$$

$$b'_{ij} = 1 \Leftrightarrow b_{ij} > b_{i(j-1)},\ b'_{ij} = -1\ otherwise.\ (b_{i0} = 1) \tag{13}$$

Proximity to zero on all these measures points that the bicluster has the corresponding coherence property. There is no limit in the value they can take, but values above 1.5 usually tells us that coherency is lost (see Section 4 for some practical cases).

3 Validation Indices

Clustering validation indices are divided into three categories [7]: external, internal and relative. External indices measure the similarity between clustering results and a priori knowledge. Internal indices compare the intrinsic structure of data with cluster results. Internal indices are much harder to apply to biclustering than external indices because much of the internal concepts (such as compactness or separation) are not applying to biclusters, where overlapping and coherent variations are usual. Finally relative indices compare different configurations of input parameters and cluster results, trying to find optimal or stable parameters for a given input data.

In the context of biclustering, external validation is mainly used, preferring biological indices to traditional ones. Internal and relative indices are seldom used, because of the non trivial task of adapting biclustering concepts as overlapping and bi-dimensionality to clustering indices.

3.1 Biological External Indices

Biological knowledge used in validations are usually gene annotations as those of Gene Ontology (GO) [2] or KEGG [8]. We will call them external indices because imply information external to the data. Given a bicluster B, we get all (in example) GO terms annotated to any of the genes in B and then apply a statistical significance test to determine if each term appearance is relevant.

Biclustering algorithms presented in [12,3] use GO and/or KEGG enrichment. Other biological knowledge applied in the same way than annotations is related with Transcription Regulatory Networks (TRNs). A TRN is a directed acyclic graph where nodes are genes, and an edge between gene A and gene B means that gene A encodes for a transcription factor protein that transcriptionally regulates (activate or repress) gene B. In this case it is considered the number of genes connected in our bicluster or the average distance between genes in it [12]. It's expected that the number of genes connected will be greater and the average distance lower than in random biclusters, which is checked with a significance test. Another interesting characteristic to check is the number of network motifs (substructures that appear in TRNs [11]) that are included in a bicluster, but it is seldom used in bibliography.

Although useful for the objective of knowledge discovery, biological significance has a major disadvantage as a validation method: biological knowledge is not complete.

When a bicluster does not group known GO/KEGG annotations, or connected genes in a TRN, it may be because it's a bad bicluster, but also because information about TRN connectiveness or GO annotations are not complete. Just as an example, E. coli TRN grew from 424 genes and 577 interactions in 2002 [13] to 1278 genes and 2724 interactions in 2004 [9]. Also statistical significance tests are controversial [6,1].

3.2 Non-biological External Indices

Non-biological external indices are used to check if bicluster results match with previous knowledge of biclusters in the data. They also can be used in comparing biclusters of two different biclustering methods. There are two main techniques to generate external indices: two-matrix and single-matrix techniques.

In case of two-matrix technique, two binary matrices are built, P and R, of size $n \times n$, where n is the number of objects (genes or conditions) of our data. P represents the grouping of objects in the a priori partition and R the grouping in our results. Frow those two matrices, indices are defined, as Rand index, Jaccard coefficient, Minkowski measure or Folkes and Mallows measure [5]. Though the adaptation of two-matrix technique to bi-dimensionality is not very difficult, the concept of overlapping is harder to express with this method, so single matrix is preferred.

Single-matrix technique builds a unique bicluster matrix M of order $p \times r$ where p is the number of biclusters in P and r is the number of biclusters in R. m_{ij} will determine the similarity between the bicluster i of P and the bicluster j of R. A measure of this similarity is F_1 index proposed by Getz et al. [4] and adapted to biclusters by Turner et al. [15]. F_1 is based in the proportion of bicluster i present in bicluster j (sensitivity or module recovery of bicluster i) and the proportion of bicluster j present in bicluster i (specificity or relevance of bicluster i). Note that the sensitivity of bicluster i for j is the specificity of bicluster j for i, and the same with the specificity of i for j, that is the sensitivity of j for i. If g_x is the number of genes in X, c_x the number of conditions in X and $n_x = g_x c_x$; sensitivity, specificity and F_1 are defined as:

$$sensitivity = \frac{(g_{A\cap B})(c_{A\cap B})}{n_B} \tag{14}$$

$$specificity = \frac{(g_{A\cap B})(c_{A\cap B})}{n_A} \tag{15}$$

$$F_1(A,B) = \frac{2(g_{A\cap B})(c_{A\cap B})}{n_A + n_B} \tag{16}$$

When results in R reveal exactly a priori partition P, M will be (if computed with Eq. 16) a square ($p \times p$), symmetric matrix with $m_{ij} = 1$ if $i = j$ and $m_{ij} = m_{ji} < 1$ otherwise. From M we can get two measures of the overall matching between R and P.

$$S(R,P) = \frac{1}{r}\sum_{i=1}^{r} max_{j=1}^{p}(m_{ij}) \tag{17}$$

$$S(P,R) = \frac{1}{p}\sum_{j=1}^{p} max_{i=1}^{r}(m_{ij}) \tag{18}$$

$S(R,P)$ gives overall bicluster relevance of biclustering R, while $S(R,P)$ gives the module recovery capacity of biclustering R.

3.3 Internal Indices

Internal indices compare intrinsic information about data with the biclustering results. In this case, no a priori information further than the raw data is available. Internal indices are not as precise as external indices, but they are important when a priori information is not available. To avoid the use of internal indices, synthetic data with known structure are built to validate biclustering methods. When applied to real biological data where no a priori information is known, biological tests are used.

An internal index is computed from two matrices just as non-biological external indices. In this case, matrix P contains information about proximity between expression levels of genes or conditions. Now, $P_{ij} = P_{ji} = distance(o_i, o_j)$. Again two pairs of matrices are needed for biclustering, one where o_i are genes and another for conditions. P_{ij} is greater when o_i and o_j are different. R can be built as described for external indices, but inversed so higher values correspond to objects not grouped together. For example $C_{ij} = 1/(1+k)$, where k is the number of times that objects i and j are grouped together. C_{ij} will be in (0,1], being 1 if never grouped together and downing to near 0 if usually grouped. This two matrices can be compared with normalized Hubert statistic:

$$\bar{\Gamma}(C,P) = \frac{\frac{1}{m}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(P_{ij}-\mu_p)(C_{ij}-\mu_c)}{\sigma_p \sigma_c} \tag{19}$$

where n is the number of objects in the matrix, and $m = n(n-1)/2$. μ_p, μ_c are the mean of the matrices and σ_p, σ_p its variances. As with other measures, $\bar{\Gamma}$ index must be computed for the two pair of matrices, then combining as in Eq. 7.

$\bar{\Gamma}$ index and other similar indices, as cophenetic coefficient are less precise than external indices. For example, Jain and Dubes [7] survey different drawbacks of cophenetic coefficient, estimating than even a value of 0.9 will not be enough to assert that there is a good correlation between P and R.

3.4 Relative Indices

Relative indices try to determine the best choice of our algorithm parameters on each particular data set. If we want to compare two algorithms against the same data set, we want to compare its best parametrization for this data set.

However this is a difficult task because of the heterogeneity of the biclustering algorithms and its input parameters. Relative indices use to be external or internal indices, depending on the availability of a priori information from the data. Independently of the index, the procedure is to run the algorithm with different parameter configurations, and compute the index for each one. The parameter configuration with best index is selected as optimal for the data set. Selection of the different parameter configurations is up to the user and is key for the optimal search, so it must represent all the range of possibilities, avoiding deviations.

In clustering, another approach to find the best configuration is to find an stable number of clusters, retrieved by a great number of configurations. From them, we take

the one in the middle of the range, or the one with the best value for a given index. This method is also used in some biclustering validations, usually to find stability when the algorithm has pseudo-random behaviour [3], but not to find optimal initial parameters.

4 Application

4.1 Algorithms

We have applied some of the performance measures discussed to two biclustering algorithms, Bimax [12] and improved Plaid Model of Turner et al [15]. Bimax is one of the most compared biclustering methods, by means of non-biological and biological validation. For example, in [12], non-biological measures are used, but only based in gene dimension because hierarchical clustering was one of the methods compared. Also, in the mentioned comparison only default parameters are used for each algorithm, no parameter optimization is done. Turner plaid model was tested by their authors with different synthetic data sets with three to ten (overlapped in different proportions) biclusters. Turner and Bimax algorithms have never been compared in bibliography.

Both methods have been implemented in R according to the specifications in the corresponding bibliography. Bimax density of 1s against 0s is proved in a range from 1% to 10% (steps of 1%). Turner's t_1 and t_2 parameters are proved as $t_1 = t_2$ in a range from 0.4 to 0.8, with steps of 0.1.

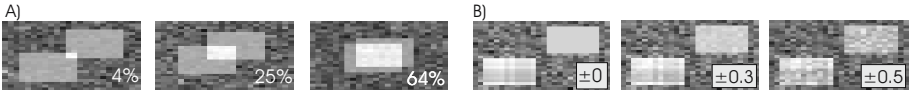

Fig. 2. A) Overlapped constant overexpression biclusters. A low noise has been added to biclusters. Overlapping degree is the same in rows and columns. B) Constant and coherent overexpression biclusters with random noise. Note how noise affects the structure of biclusters, being constancy undistinguishable from coherency with high noise.

4.2 Data Sets

Two sets of synthetic data matrices 100x50 are built. First set of matrices will contain two constant biclusters with overlapping degrees from 0% to 100%, with 10% increments. Second set of matrices have two non-overlapping biclusters, one constant and the other one additive coherent, with normal distribution random noise. Distribution deviation increases from 0 (no noise) to 1, with 0.1 increments. All matrices have a random noise background (see Fig. 2).

4.3 Methods

The proposed test will briefly apply the techniques discussed. First, we will try to find the best parameter choice for each biclustering algorithm in each data set, by means of

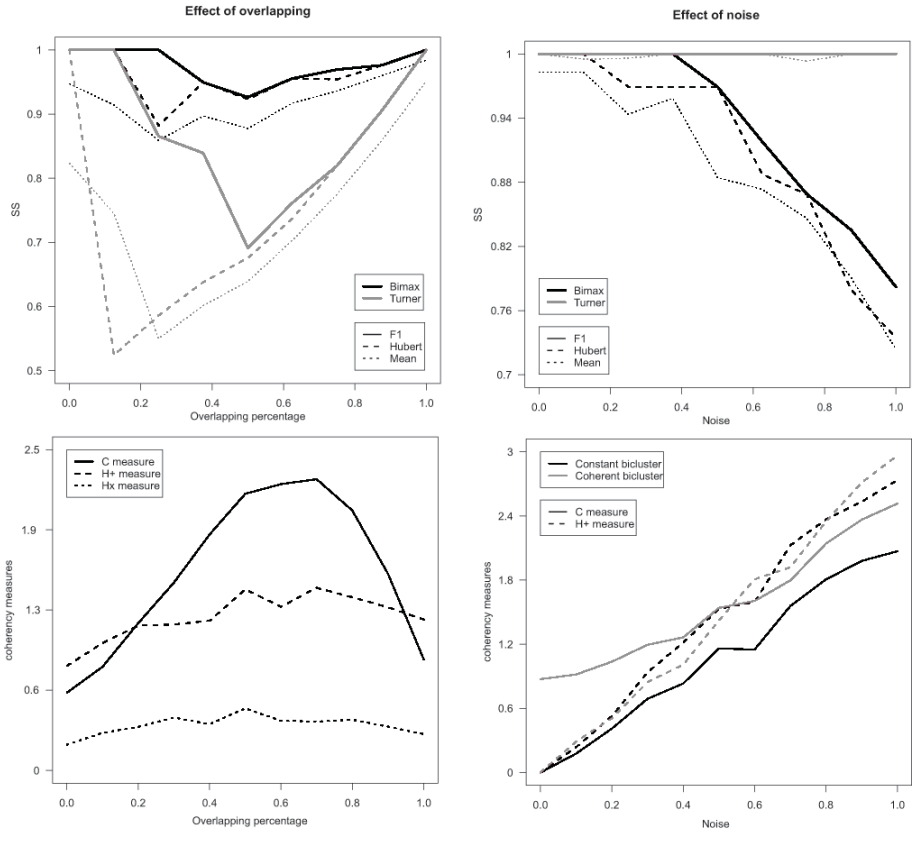

Fig. 3. a) Effect of overlapping in the algorithm and the biclusters. 1) Best *SS* measure achieved by using F_1 and $\bar{\Gamma}$ statistics along with the mean of *SS* for all the proven configurations. 2) Variation in the measures of constancy and coherency with changes in the overlap degree. b1) and b2) As a1) and a2), but representing the effect of the noise in the algorithms and biclusters, respectively.

F_1 measure (comparing against known biclusters) and of $\bar{\Gamma}$ (comparing against proximity matrix). That way, we can compare the performance of $\bar{\Gamma}$ as relative index against an a priori knowledge technique (F_1). Biological significance tests has left out of the scope of this discussion because studies with them are more extended and do not use the measures reviewed here. For known biclusters, constant and coherence measures will be also computed, analyzing its consistency against noise and overlap.

4.4 Results

Fig. 3a-1 presents the mean of sensitivity and specificity (*SS*) of the results of the best configuration given by F_1 and $\bar{\Gamma}$ (or Hubert statistic). F_1 will give the best configuration

at all, while $\bar{\Gamma}$ gives the best configuration supposing a priori information is not available. Also, the mean *SS* for all the tested parametrizations is given. With the appropriate parameter choice, Bimax finds a high percentage of row and columns present in biclusters embedded, even (sometimes) finding the exact biclusters without finding spurious biclusters ($SS = 1$). Performance is lower when overlapping is around 50%, being higher when biclusters are nearly separated or are almost the same. *SS* value of parameter configuration chosen by $\bar{\Gamma}$ measure is obviously worse, but still have better configurations than average. Turner algorithm has lower performance than Bimax. The pruning phase included to improve plaid model fails when trying to prune overlapped parts of the biclusters.

Overlapping effect on biclusters measures is represented in Fig. 3a-2. Because of additive overlapping, intersecting expression levels are higher than non-intersecting, so constant structure is lost with overlapping, in favor of coherent structure.

In Fig. 3b-1 we can see how Bimax performance is sensible to noise when it exceeds 0.4 deviations. Bimax discretization threshold is the responsible of this downgrading. On the other hand, Turner algorithm is not affected by noise, recovering data even in the most noisy cases. Again, $\bar{\Gamma}$ statistic does not give the best configuration in each case, but is better than average. About constancy and coherence measures (Fig. 3 b-2), the measures increase with noise, revealing how structure is eventually lost. Additive coherent bicluster has lower (better) H^+ measure than *C* measure, as expected. Note how H^+ measures increase with noise until, eventually, surpassing *C* measure and coinciding with Bimax performance downgrade.

5 Conclusions and Future Work

Due to the variation and drawbacks of validation indices, the best way to analyze biclustering performance is to use them exhaustively, generating a framework that will define bicluster specific measures (relative, internal and external indices), data type definitions (constant, coherent), benchmark algorithms and example (real and synthetic) data sets.

Though external indices use is extended, our approach to relative and internal index application is new. That helps in automatic optimization of biclustering input paremeters, a task seldom considered and critical for obtaining the highest performance. Data type definition exists as discussed, but only constant biclusters have been mathematically measured. We present an approach to measure coherence biclusters by using constant measures and transformation of data matrices.

External and internal indices used as relative indices have been applied to two biclustering algorithms to prove their consistency and capability to generate information about performance and bicluster behavior against noise and overlap, main problems of biclustering on microarrays. The search of the optimal input parameters for biclustering algorithms through $\bar{\Gamma}$ internal index outperforms the static use of recommended values.

Coherence measures have been also proposed and applied, proving helpful in typifying biclusters. Normalization of these measures must be done to help in comparisons between them. We expect to exhaustively prove all these measures (analyzing and comparing existing biclustering algorithms) and present newer ones in future works.

References

1. Anderson, D.R., Burnham, K.P., Thompson, W.L.: Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64(4), 912–913 (2000)
2. Ashburner, M., Ball, C.A., Blake, J.A., Bolsteing, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Geneontology: tool for the unification of biology the gene ontology consortium. Nature Genetics 25, 25–29 (2000)
3. Carmona-Saez, P., Pascual-Marqui, R.D., Tirado, F., Carazo, J.M., Pascual-Montano, A.: Biclustering of gene expression data by non-smooth non-negative matrix factorization. BMC Bioinformatics 7(78) (2006)
4. Getz, G., Levine, E., Domany, E.: Coupled two-way clustering analysis of gene microarray data. Proc. Natural Academy of Sciences 97(22), 12079–12084 (2000)
5. Halkidi, M., Batisfakis, Y., Vazirgiannis, M.: On clustering validation techniques. Journal of Intelligent Information Systems 17(2/3), 107–145 (2001)
6. Hubbard, R.: Why we don't really know what "statistical significance" means: a mayor educational failure. Journal of Marketing Education 28, 114–120 (2006)
7. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs (1988)
8. Kanehisa, M., Goto, S.: Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28(1), 27–30 (2000)
9. Ma, H.-W., Kumar, B., Ditges, U., Gunzer, F., Buer, J., Zeng, A.-P.: An extended transcriptional regulatory network of escherichia coli and analysis of its hierarchical structure and network motifs. Nucleic Acids Research 32(22), 6643–6649 (2004)
10. Madeira, S., Oliveira, A.: Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Transactions of Computational Biology and Bioinformatics 1(1), 24–45 (2004)
11. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: simple building blocks of complex networks. Science 298, 824–827 (2002)
12. Prelic, A., Bleuer, S., Zimmermann, P., Wille, A., Bhlmann, P., Gruissem, W., Hennig, L., Thiele, L., Zitzler, E.: A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics 22(9), 1122–1129 (2006)
13. Shen-Orr, S.S., Milo, R., Mangan, S., Alon, U.: Network motifs in the transcriptional regulation network of escherichia coli. Nature Genetics 31, 64–68 (2002)
14. Turner, H., Bailey, T., Krzanowski, W.: Improved biclustering of microarray data demonstrated through systematic performance tests. Computational Statistics and Data Analysis 48, 235–254 (2003)
15. Turner, H.L., Bailey, T.C., Krzanowski, W.J., Hemingway, C.A.: Biclustering models for structured microarray data. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2(4), 316–329 (2005)