Proximity Measures and Results Validation in Biclustering – A Survey

Patryk Orzechowski

AGH University of Science and Technology,
Department of Automatics and Biomedical Engineering
30-059 Cracow, Mickiewicza Av. 30, Poland
patrick@agh.edu.pl
http://home.agh.edu.pl/~patrick

Abstract. The concept of biclustering evolved from traditional clustering techniques, which have proved to be inadequate for discovering local patterns in gene microarrays, in particular with shifting and scaling patterns. In this work we compare similarity measures applied in different biclustering algorithms and review validation methodologies described in literature. To our best knowledge, this is the first in-depth comparative analysis of proximity measures and validation techniques for biclustering. Current trends in design of similarity measures as well as a rich collection of state-of-the-art benchmark datasets are presented, supporting algorithm designers in classification of comparison and quality assessment criteria of emerging biclustering algorithms.

Keywords: biclustering, co-clustering, shifting and scaling patterns, pattern similarity, proximity measures, results validation, microarray gene expression data, state-of-the-art, survey.

1 Introduction

Gene expression data (or simply: *expression data*) is usually organized in form of a matrix with rows corresponding to different genes and columns to conditions. Data samples come from individual organs, tissues (healthy or affected by some disease) or organisms exposed to various environmental conditions or a single condition over certain amounts of time. Genes that react similarily to certain conditions are supposed to have corresponding functionalities and be involved in similar biological processes [38].

The concept of biclustering (formulated by Hartigan [24] and applied primarily to gene expression data by Cheng and Church [12]) emerged as clustering algorithms were inadequate to detect similar expressions of genes exhibited to certain set of (not all) conditions. Clustering managed to detect similar patterns only on a global scale and grouped either all genes by analysing they response to all conditions, or conversely grouped all conditions by taking into account whole expression profiles of the genes. Another problem was clustering complexity, as clustering process had to be applied separately to columns and rows of the data

L. Rutkowski et al. (Eds.): ICAISC 2013, Part II, LNAI 7895, pp. 206–217, 2013.

matrix. Afterwards, like in hierarchical clustering [15], typically all combinations of rows/columns had to be considered so that a subset of rows and columns satisfying criteria could be determined. Keeping in mind the fact that for a n element set we have 2^n combinations in total, biclustering approaches that approached both dimensions simultaneously were a natural consequence.

In this article we summarize developments in biclustering and analyse current trends in evolution of similarity measures that determine algorithm architecture.

2 Definitions

Given a n x m data matrix $A = \{a_{ij}\}$ with values obtained by exposing n objects $X = \{x_1, ..., x_n\}$ to m conditions $Y = \{y_1, ..., y_m\}$, the response of the *i-th* gene to the *j-th* condition can be described as $a_{ij} = (x_i, y_j)$, where $i \leq n$ and $j \leq m$.

Formally, biclustering problem may be formulated as follows. Given any n x m data matrix $A = \{a_{ij}\}$ with rows $X = \{x_{1,}..., x_n\}$ and columns $Y = \{y_{1,}..., y_m\}$, its values could be presented as $a_{ij} = (x_i, y_j)$. Subset of p rows $I = \{i_{1,...,}i_p\}$, where $I \subseteq X$ and $p \leq n$ and subset of q columns $J = \{j_{1,...,}j_q\}$, where $J \subseteq Y$ and $q \leq m$ is called a *bicluster* $B = (I, J) = \{a_{ij} \in A : i \in I, j \in J\}$ when its rows I are as similar as possible to each other across its columns J and vice-versa. *Biclustering* is task of identifying a series of *biclusters* $B_k = (I_k, J_k)$, such that each B_k would meet a specified homogeneity (or similarity) criterion [38].

The second definition of bicluster [1] defines bicluster as a triplet $B(I, J, h)$, in which $I \subseteq X$, $J \subseteq Y$ and $h : I \times J \to R$ is the level function of the bicluster, such that $\forall_{(x_i,y_j)\in I\times J} : h(x_i, y_j) = a_{ij}$. Basing on this definition, biclustering aims at identification of triplets.

Biclustering formulation varies across authors, though, with some wanting to obtain motifs [40], plaid models [34], perfect δ -biclusters [38] etc.

One of the main challenges of biclustering is to find biclusters with shifting and scaling patterns. A group of genes show a *shifting pattern* (when the values of genes vary with the addition of an additive constant β_i) or *scaling pattern* (when the values vary with multiplication by a multiplicative constant α_i).

Formally, a bicluster $B = \{a_{ij}\}$ has a shifting or scaling pattern when it follows the expressions (1) or (2), respectively:

$$a_{ij} = \pi_j + \beta_i \tag{1}$$

$$a_{ij} = \pi_j \times \alpha_i \tag{2}$$

where the values β_i and α_i are fixed for all genes and π_j is a fixed value for every condition. With random noise ε_{ij} included, elements of bicluster that follow both expressions (1) and (2) could be defined in general form as (3):

$$a_{ij} = \pi_j \times \alpha_i + \beta_i + \varepsilon_{ij} \tag{3}$$

Shifting and scaling patterns are of much interest, as genes may respond to conditions similarly, even though they may have started with different initial conditions or the level of their response may differ in strength [1]. Plaid model [34] and Bayesian Biclustering (BBC) [22] extend this concept.

3 Proximity Measures

A perfect biclustering algorithm needs to be in line with the following rules [33]:

- handle high dimensional data
- be insensitive to outliers, noise and order of data input
- have low complexity (time and space)
- require few input parameters
- incorporate meta-data knowledge
- produce biologically interpretable results

Many different measures have been tested for clustering or biclustering. Some of the commonly used measures have been categorized in this chapter.

3.1 Distance-Based Measures

Distance functions, including the so-called Minkowski measures (Euclidean, Manhattan, Chebyshev) are useful when extracting exact matches between two objects [33]. They are easily computable and yield global similarities between two vectors. Unfortunately, they are also very sensitive to noise and outliers [5][33]. Distance measures are excessively documented in [13]. More sophisticated conceptual measures need to be applied in order to overcome distance-based measure disadvantages [33].

3.2 Qualitative Measures

Qualitative measures often assume that positive and negative values in microarray data carry equal amount of information. Usually, these are numbers of ups/down/no changes or (in binary case) number of positive/negative values for conditions [5][13]. The most popular coefficients take into account the numbers of positive elements in both objects (denoted as a), negative elements in both objects (denoted as d), number of positives in $j-th$ objects and negatives in $k-th$ object (denoted as b) or conversely (denoted by c). Qualitative measures are also classified as part of non-correlation measures [13].

The most widely applied qualitative measures are the so-called simple coefficient measure, *Jaccard coefficient* and *Sorenson coefficient*. Simple matching coefficient is defined as (4):

$$C_{jk} = \frac{a+d}{a+b+c+d}, \quad 0 \leq C_{jk} \leq 1 \tag{4}$$

Jaccard coefficient is defined for two objects j and k as follows (5):

$$J_{jk} = \frac{a}{a+b+c}, \quad 0 \leq J_{jk} \leq 1 \tag{5}$$

Sorenson coefficient between objects j and k is defined as (6):

$$S_{jk} = \frac{2a}{2a+b+c}, \quad 0 \leq S_{jk} \leq 1 \tag{6}$$

Maximum similarity is achieved when coefficients are equal to one. All coefficients fall within the range of 0 to 1. It is worth noticing how the last two coefficients are sensitive to the so-called direction of coding (switching all 0's and 1's would result in different coefficients values) [46].

The QUBIC [35] concept of sorting microarray values, taking only extreme values into account and later representing similarity as the number of element co-occurrences, extends qualitative measures. Such measure is used by ISA as well [28].

3.3 Non-correlation-Based Measures

The group of non-correlation measures contains algorithms which do not rely on counting elements of the set (nor their occurrences) and specify a formula for determining object similarity instead. One group of algorithms is dominant in this category – numerous approaches based on the measure proposed by Cheng-Church, namely Mean Square Residue [12]. Mean square residue H of bicluster (I, J) is defined by equation (7):

$$H(I,J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \tag{7}$$

where $a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{ij}$ is a row mean, $a_{Ij} = \frac{1}{|I|} \sum_{i \in I} a_{ij}$ is a column mean and $a_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} a_{ij} = \frac{1}{|I|} \sum_{i \in I} a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{Ij}$ is the mean of submatrix (I, J) .

The second representative of non-correlation-based measures is HARP [57] that uses quality metric assessment and detects constant patterns only, which is not suitable for real data [7]. Many algorithms use Mean Square Residue straightforwardly or as a part of the determinant of similarity. These algorithms include Cheng-Church algorithm [12], its improvement – FLOC algorithm [56] and multiple recent publications, such as CPB [8], Particle Swarm Optimization [37], Simulated Annealing [9], greedy randomized adaptive search [14], localization and extraction with adaptive noise hiding LEB [17] or estimation of distribution algorithms [36].

It has been proven, however, that mean square residue manages shifting patterns (i.e. addition of a constant), but does not manage scaling patterns (i.e. multiplication by a constant might affect the score). This occurs when variance of gene values is high [1][5][47]. It suggests that the measure may have already become insufficient for biclustering purposes.

3.4 Correlation-Based Measures

There are two groups of measures using correlation: *parametric* (which estimate population parameters and assume bivariate normal distribution of data) and *non-parametric* (which allow less demanding assumptions and do not attempt to estimate population parameters) [11]. Correlation-based measures are scale-invariant, with distance between two objects usually calculated as $distance = 1 - correlation^2$ if sign is accurate [33].

Parametric Correlation-Based Measures. Two most commonly recognized parametric correlation measures are cosine similarity (also reckoned as distance measure, but classified in this category basing on its similarity to other representatives) and standard Pearson correlation. Cosine similarity between two vectors X and Y is defined as (8):

$$\cos(\theta) = \frac{X \cdot Y}{\|X\| \, \|Y\|} = \frac{\sum_{i=1}^{n} X_i \cdot Y_i}{\sqrt{\sum_{i=1}^{n} (X_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (Y_i)^2}} \tag{8}$$

Definition of traditional correlation includes means of both vectors (9):

$$\begin{aligned} cov(X,Y) = \sigma_{xy} = E((X - \overline{X})(Y - \overline{Y})) = \\ = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2} \cdot \sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}} \end{aligned} \tag{9}$$

Pearson measure of correlation is defined as (10) [54]:

$$\rho = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - (E(X))^2} \cdot \sqrt{E(X^2) - (E(X))^2}} \tag{10}$$

Parametric measures are said to be sensitive to outliers and noise [33][46] and to fail to capture true grouping [47] in opposition to non-parametric correlation based measures which base only on the ordinal position of elements. Some of recognized algorithms use modified Pearson correlation coefficient as the measure of similarity, BBC[22] and CPB [8] and Scatter Search [42] are examples.

Non-parametric Correlation Based Measures. Non-parametric correlation measures include i.a. Spearman's rank correlation and Kendall's τ. They are sometimes classified as distance measures [5].

Spearman's rank correlation coefficient (11), called also Spearman's rho coefficient, is considered to be equal to Pearson correlation between ranked variables. First, X and Y raw scores are ranked into x_i and y_i (taking into account average positions in the ascending order of the values), then Spearman's rank is calculated [41].

$$\rho = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot \sum_{i=1}^{n} (y_i - \overline{y})^2}} \tag{11}$$

Some recently published similarity functions base on Spearman's rank correlation, for example Average Spearman's Rho (ASR) defined as (12):

$$ASR(I,J) = 2 \; max \left\{ \frac{\sum_{i \in I} \sum_{j \in I, j>i} \rho_{ij}}{|I|(|I|-1)}, \frac{\sum_{k \in J} \sum_{l \in J, l>k} \rho_{ij}}{|J|(|J|-1)} \right\} \tag{12}$$

where ρ_{ij} and ρ_{kl} are the Spearman's rank correlantions associated respectively with row indices i and j, and column indices k and l of a bicluster (I, J) [5].

Another approach basing on Spearman's rho is Average Correlation Value (ACV) defined as (13) [53]:

$$ACV(I,J) = max\left\{\frac{\sum_{i\in I}\sum_{j\in I}|\rho_{ij}| - |I|}{|I|(|I|-1)}, \frac{\sum_{k\in J}\sum_{l\in J}|\rho_{kl}| - |J|}{|J|(|J|-1)}\right\} \quad (13)$$

where ρ_{ij} and ρ_{kl} are defined analogically. Notice that $ASR(I,J) \in [-1;1]$ and $AVR(I,J) \in [0;1]$.

Second most recognizable non-parametric covariation-based measure is Kendall's τ defined as (14), which is said to be generally insensitive to outliers (i.e. outliers are detectable, but can change the value of correlation):

$$\tau = \frac{2\sum_{i<j} K_{ij}}{n(n-1)} \quad (14)$$

where

$$K_{ij} = \begin{cases} 1 & \text{when when } x_i \text{ and } y_i \text{ are concordant} \\ -1 & \text{when } x_i \text{ and } y_i \text{ are disconcordant} \end{cases}$$

Concordance between two samples is understood as agreement in order, i.e. two points $P = (x_i, y_j)$ and $Q = (x_j, y_j)$ are concordant if $(x_i - x_j)(y_i - y_j) > 0$.

Non-paramteric covariation-based measures are represented by FABIA [25] and PDNS[5]. Two more correlation measures are promising for biclustering and worth mentioning: *Mahalanobis distance* (scale invariant, with ellipsoidal shapes of clusters, instead of spherical as in Euclidean distance calculation) and *adaptive distance norm* from Gustafson-Kessel method of fuzzy clustering (with covariances estimated in eigenvalue calculations, unique distance measure applied to each cluster) [29][33].

4 Result Validation Methodology

Different problem formulations used in biclustering schemes impede general comparison of biclustering algorithms [45]. As algorithms are designed to work with specific constraints, they may perform better in a specific scenario and worse in others. Choice of correct initial parameters is also crucial to obtain satisfactory biclustering results [16].

The majority of biclustering algorithms refer to biological data and start with identification of locally co-expressed genes. Classification of samples or inference of regulatory mechanism are also areas of biclustering application [6][12][17][28][35][51].

The results obtained from biclustering algorithms may be validated basing on three index categories: *internal*, *external* and *relative* [23][30][29][31][45]. Internal indices verify if the recovered structure is appropriate for the data, external indices compare the recovered structure to the structure known *a priori*, while relative indices assess which of the recovered structures is better according to some quality measure.

4.1 Internal Indices

The number of biclusters, as well as sizes (minimum, maximum, average) of biclusters returned by an algorithm are first choice criteria to classify biclustering algorithms [17]. Such measures are useful for quantity assessment, but they do not involve the quality of each bicluster nor completeness of biclustering. Two clustering measures could be adapted to present (dis-)simmilarity of biclusters: *homogeneity* – the degree of similarity of elements in the same cluster and *separation* – the value determining the similarity of different clusters (i.e. how much biclusters overlap between each other) [20].

Homogeneity and Separation. *Homogeneity* represents average distance between objects in the same bicluster, while separation is defined as the weighted average similarity between objects from different biclusters [20]. *Homogeneity H* of an object u belonging to the bicluster X can be determined according to (15):

$$H = \frac{1}{N} \cdot \sum_{u \in N} f(g(u), g(X)) \tag{15}$$

where f is a similarity function and g is the expression level of an object.

Separation S for biclusters $X_1, ..., X_n$ is defined as (17) [49]:

$$S = \frac{1}{\sum_{i \neq j} |X_i||X_j|} \cdot \sum_{i \neq j} |X_i||X_j| f(g(X_i), g(X_j)) \tag{16}$$

Superior algorithms provide high homogeneity and low separation [49].

Other Measures. *Significance* adapted from clustering may be considered a third type of internal index. Basing on Monte Carlo method or probability assessment, significance assesses the likelihood of obtaining randomly a bicluster of given quality [30]. *Average silhouette width* may also be considered as a criterion, determining the distance of each object from other objects in a bicluster [20][10]. Several other measures for clustering algorithms are available in [31].

4.2 External Indices

Biological Datasets. The information used for biological validation is concerned to be external to data. Gene annotation databases, such as Gene Ontology (GO) [4] or KEGG [43][32], are common choice for performing validation of biclustering algorithms with respect to biological knowledge.

Most biclustering algorithms carry out experiments on one (or more) of the following datasets:

- yeast cell-cycle data set of *Saccharomyces cerevisiae*(selection of 2993 genes with 173 different stress conditions) [19][50]
- cancer dataset: 4026 genes of over 96 human tissue samples with 9 types of lymphoma and control [2][51]

- 12533 probes from 72 patients with different subtypes of leukemia [3][35]
- gene expression datasets provided by Broad Institute as "Cancer Program Data Sets", analyzed in [26][25]
- bio-synthetic pathways of *Arabidopsis thaliana* 734 genes under 69 experimental conditions [55]
- M3D database (4217 *Escherichia coli* genes under 264 conditions) [18][35]
- metabolic pathways [28], promoter motifs [27] etc.

Gene enrichment is measured with p-values, which specify the probability of finding the number of genes from a particular GO category (function, process and component) within each bicluster. In order to calculate probability p of finding at least k-genes from a category within a bicluster of size n, cumulative hyper-geometric distribution is used[52]:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{m}{i}\binom{g-m}{n-i}}{\binom{g}{n}} \tag{17}$$

Correspondence with GO category requires calculation of p-values for each functional category in each bicluster [37]. There exists some controversies as to whether biological validation of an algorithm may be considered to constitute a true validation of its performance. One of the reasons is incompleteness of biological knowledge. Hence, any bicluster that has been correctly depicted by a biclustering algorithm may be still erroneous, as GO/KEGG annotations or connected genes in a TRN are increasingly expanded [48].

Synthetic Datasets. Synthetic datasets may be used as benchmark as well. Most recognizable datasets have been generated by Prelic eat al. [45], Li et al. [35] and Hochreiter et al. [25]. First two benchmark datasets are small (50-100 genes), contain biclusters of equal sizes and have only simultaneous row and column overlaps. The third one, designed to match gene expression data characteristic in terms of heavy tails, contains 100 datasets of size 1000x100 with 10 multiplicative biclusters in each and additive noise. This artificial scenario is claimed to be more realistic in terms of densities and moments of datasets compared to real data [25].

4.3 Relative Indices

The third type of indices measure which recovered structure is better according to a quality measure and how input parameters influence biclustering outcome. Comparison between two different sets of biclusters may be achieved with different consensus measures [22][45][35]. Various modifications of Jaccard index are commonly used consensus measures for computing distance between two sets. Jaccard index represents the fraction of row-column combinations in both biclusters from all row-column combinations in at least one of biclusters (18):

$$f_j(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{18}$$

Those techniques in general do not take into account overlapping biclusters or entirely consider different numbers of biclusters. Novel consensus score for biclustering has been defined by [25] and bases on computing pairwise comparison between two sets of biclusters with Munkres algorithm [39], quite similar to a technique applied to visualize clustering results for proteins [44].

Another index which is worth mentioning is the F_1 index proposed by [21]. Some consider relative indices as determination of the best parametrization set for the algorithm [45][48]. The importance of proper choice of input parameters and their impact on biclustering outcome for many biclustering algorithms have been repetitively emphasized [16][17].

4.4 Conclusions

In this article current trends in biclustering are summarized with respect to proximity measures of biclusters. With many approaches available, most promising area of development of biclustering algorithms seems to be correlation based measures, especially non-parametric ones.

Novel approaches may also agglomerate results obtained from different biclustering approaches. There is a tremendous perspective for developing ensemble approach that will combine different measures and automatically detect which scheme should be applied basing on data analysis.

In the second part of the article, we classify different state-of-the-art validation techniques of biclustering algorithms. Novel approaches need to perform well in synthetic datasets in order to become recognizable, followed by successes with real data examples. So-called *reference set* of biclustering algorithms needs to be updated to include at least the following algorithms: BBC [22], CPB[8], QUBIC[35], ISA[28], FABIA[25]. Each of those algorithms achieves best results its specific cateogory of biclustering [16]. Comparison against previous algorithms such as CC [12], OPSM [6] or xMotifs[40], SAMBA [51] or HCL[15] could be treated as optional [35]. To our best knowledge, this is the first in-depth analysis considering the proximity measures fundamental in design of biclustering algorithm architecture.

Acknowledgments. The author would like to express his gratitude to prof. Krzysztof Boryczko for his consultancy and support.

References

1. Aguilar-Ruiz, J.: Shifting and scaling patterns from gene expression data. Bioinformatics 21(20), 3840–3845 (2005)
2. Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., et al.: Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature 403(6769), 503–511 (2000)
3. Armstrong, S., Staunton, J., Silverman, L., Pieters, R., den Boer, M., Minden, M., Sallan, S., Lander, E., Golub, T., Korsmeyer, S., et al.: Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genetics 30(1), 41–47 (2002)

4. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., et al.: Gene ontology: tool for the unification of biology. Nature Genetics 25(1), 25 (2000)
5. Ayadi, W., Elloumi, M., Hao, J.: Pattern-driven neighborhood search for biclustering of microarray data. BMC bioinformatics 13(suppl. 7), S11 (2012)
6. Ben-Dor, A., Chor, B., Karp, R., Yakhini, Z.: Discovering local structure in gene expression data: the order-preserving submatrix problem. In: Proceedings of the Sixth Annual International Conference on Computational Biology, RECOMB 2002, pp. 49–57. ACM, New York (2002), http://doi.acm.org/10.1145/565196.565203
7. Bozdağ, D., Kumar, A.S., Catalyurek, U.V.: Comparative analysis of biclustering algorithms. In: Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology, BCB 2010, pp. 265–274. ACM, New York (2010), http://doi.acm.org/10.1145/1854776.1854814
8. Bozdağ, D., Parvin, J.D., Catalyurek, U.V.: A biclustering method to discover co-regulated genes using diverse gene expression datasets. In: Rajasekaran, S. (ed.) BICoB 2009. LNCS, vol. 5462, pp. 151–163. Springer, Heidelberg (2009), http://dx.doi.org/10.1007/978-3-642-00727-9_16
9. Bryan, K.: Biclustering of expression data using simulated annealing. In: Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems, CBMS 2005, pp. 383–388. IEEE Computer Society Press, Washington, DC (2005), http://dx.doi.org/10.1109/CBMS.2005.37
10. Chen, G., Jaradat, S., Banerjee, N., Tanaka, T., Ko, M., Zhang, M.: Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data. Statistica Sinica 12(1), 241–262 (2002)
11. Chen, P., Popovich, P.: Correlation: Parametric and nonparametric measures, pp. 137–139. Sage Publications, Incorporated (2002)
12. Cheng, Y., Church, G.: Biclustering of expression data. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, vol. 8, pp. 93–103 (2000)
13. Choi, S., Cha, S., Tappert, C.: A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics 8(1), 43–48 (2010)
14. Dharan, S., Nair, A.S.: Biclustering of gene expression data using reactive greedy randomized adaptive search procedure. BMC Bioinformatics 10(suppl. 1), S27 (2009)
15. Eisen, M., Spellman, P., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 95(25), 14863–14868 (1998)
16. Eren, K., Deveci, M., Küçüktunç, O., Çatalyürek, Ü.: A comparative analysis of biclustering algorithms for gene expression data. Briefings in Bioinformatics (2012)
17. Erten, C., Sözdinler, M.: Biclustering expression data based on expanding localized substructures. In: Rajasekaran, S. (ed.) BICoB 2009. LNCS, vol. 5462, pp. 224–235. Springer, Heidelberg (2009)
18. Faith, J., Driscoll, M., Fusaro, V., Cosgrove, E., Hayete, B., Juhn, F., Schneider, S., Gardner, T.: Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. Nucleic Acids Research 36(suppl. 1), D866–D870 (2008)
19. Gasch, A., Spellman, P., Kao, C., Carmel-Harel, O., Eisen, M., Storz, G., Botstein, D., Brown, P.: Genomic expression programs in the response of yeast cells to environmental changes. Science Signalling 11(12), 4241 (2000)
20. Gat-Viks, I., Sharan, R., Shamir, R.: Scoring clustering solutions by their biological relevance. Bioinformatics 19(18), 2381–2389 (2003)

21. Getz, G., Levine, E., Domany, E.: Coupled two-way clustering analysis of gene microarray data. Proceedings of the National Academy of Sciences 97(22), 12079–12084 (2000)
22. Gu, J., Liu, J.S.: Bayesian biclustering of gene expression data. BMC genomics 9(suppl. 1), 4 (2008)
23. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. Journal of Intelligent Information Systems 17(2), 107–145 (2001)
24. Hartigan, J.: Direct clustering of a data matrix. Journal of the American Statistical Association 67(337), 123–129 (1972)
25. Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., Khamiakova, T., Van Sanden, S., Lin, D., Talloen, W., et al.: Fabia: factor analysis for bicluster acquisition. Bioinformatics 26(12), 1520–1527 (2010)
26. Hoshida, Y., Brunet, J., Tamayo, P., Golub, T., Mesirov, J.: Subclass mapping: identifying common subtypes in independent disease data sets. PloS One 2(11), e1195 (2007)
27. Ihmels, J., Bergmann, S., Barkai, N.: Defining transcription modules using large-scale gene expression data. Bioinformatics 20(13), 1993–2003 (2004)
28. Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., Barkai, N., et al.: Revealing modular organization in the yeast transcriptional network. Nature Genetics 31(4), 370–378 (2002)
29. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. 31(3), 264–323 (1999), http://doi.acm.org/10.1145/331499.331504
30. Jain, A.K., Dubes, R.: Algorithms for clustering data. Prentice-Hall, Inc. (1988)
31. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recogn. Lett. 31(8), 651–666 (2010), http://dx.doi.org/10.1016/j.patrec.2009.09.011
32. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., et al: Kegg for linking genomes to life and the environment. Nucleic acids research 36(suppl. 1), D480–D484 (2008)
33. Kerr, G., Ruskin, H., Crane, M., Doolan, P.: Techniques for clustering gene expression data. Computers in Biology and Medicine 38(3), 283–293 (2008)
34. Lazzeroni, L., Owen, A., et al.: Plaid models for gene expression data. Statistica Sinica 12(1), 61–86 (2002)
35. Li, G., Ma, Q., Tang, H., Paterson, A., Xu, Y.: Qubic: a qualitative biclustering algorithm for analyses of gene expression data. Nucleic Acids Research 37(15), e101–e101 (2009)
36. Liu, F., Zhou, H., Liu, J., He, G.: Biclustering of gene expression data using eda-ga hybrid. In: IEEE Congress on Evolutionary Computation, CEC 2006, pp. 1598–1602. IEEE (2006)
37. Liu, J., Li, Z., Hu, X., Chen, Y.: Biclustering of microarray data with mospo based on crowding distance. BMC bioinformatics 10(suppl. 4), S9 (2009)
38. Madeira, S., Oliveira, A.: Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics 1(1), 24–45 (2004)
39. Munkres, J.: Algorithms for the assignment and transportation problems. Journal of the Society for Industrial & Applied Mathematics 5(1), 32–38 (1957)
40. Murali, T., Kasif, S.: Extracting conserved gene expression motifs from gene expression data. In: Proc. Pacific Symp. Biocomputing, vol. 3, pp. 77–88 (2003)
41. Myers, J., Well, A.: Research design and statistical analysis. Lawrence Erlbaum (2002)
42. Nepomuceno, J., Troncoso, A., Aguilar-Ruiz, J., et al.: Biclustering of gene expression data by correlation-based scatter search 4(3) (2011)

43. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M.: Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research 27(1), 29–34 (1999)
44. Orzechowski, P., Boryczko, K.: Parallel approach for visual clustering of protein databases. Computing and Informatics 29(6+), 1221–1231 (2010), http://www.cai.sk/ojs/index.php/cai/article/view/140
45. Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, 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)
46. Romesburg, C.: Cluster analysis for researchers. Lulu. com (2004)
47. Roy, S., Bhattacharyya, D., Kalita, J.: Deterministic approach for biclustering of co-regulated genes from gene expression data. Advances in Knowledge-Based and Intelligent Information and Engineering Systems 243, 490–499 (2012)
48. Santamaría, R., Quintales, L., Therón, R.: Methods to bicluster validation and comparison in microarray data. In: Yin, H., Tino, P., Corchado, E., Byrne, W., Yao, X. (eds.) IDEAL 2007. LNCS, vol. 4881, pp. 780–789. Springer, Heidelberg (2007)
49. Sharan, R., Elkon, R., Shamir, R.: et al.: Cluster analysis and its applications to gene expression data. In: Ernst Schering Res Found Workshop, vol. 38, pp. 83–108 (2002)
50. Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell 9(12), 3273–3297 (1998)
51. Tanay, A., Sharan, R., Shamir, R.: Discovering statistically significant biclusters in gene expression data. Bioinformatics 18(suppl. 1), S136–S144 (2002)
52. Tavazoie, S., Hughes, J., Campbell, M., Cho, R., Church, G., et al.: Systematic determination of genetic network architecture. Nature Genetics 22, 281–285 (1999)
53. Teng, L., Chan, L.: Discovering biclusters by iteratively sorting with weighted correlation coefficient in gene expression data. Journal of Signal Processing Systems 50(3), 267–280 (2008)
54. Wilcox, R.: Introduction to robust estimation and hypothesis testing. Academic Press (2005)
55. Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., Von Rohr, P., Thiele, L., et al: Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. Genome Biol. 5(11), R92 (2004)
56. Yang, J., Wang, H., Wang, W., Yu, P.: Enhanced biclustering on expression data. In: Proceedings of Third IEEE Symposium on Bioinformatics and Bioengineering, pp. 321–327 (March 2003)
57. Yip, K.Y., Cheung, D.W., Ng, M.K.: Harp: A practical projected clustering algorithm. IEEE Trans. on Knowl. and Data Eng. 16(11), 1387–1397 (2004), http://dx.doi.org/10.1109/TKDE.2004.74