**Biclustering algorithms :**

**BBC :**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | Nucleic Acids Research | No | others | Cited as Background | Subspace-clustering algorithms have been de- veloped to analyse 2D microarray data (also known as biclustering in microarray data analysis literature), associating subsets of genes whose expression are coherent under a subset of conditions (46–50). |
| 2 | Journal of Bone and Mineral Research | Yes | usage | Used as a basis for finding block clusters in GWAS data | Based on a Bayesian statistical model framework proposed by Gu and Liu,(28) but with modifications to increase the robustness and efficiency of the |
| 3 | BMC Genomics | No | others | Cited in introduction | Eight papers discuss novel mathematical or statistical approaches to analyze microarray datasets. Gu and Liu [1] proposed a Bayesian biclustering model, and imple- mented a Gibbs sampling procedure and illustrated that such Bayesian biclustering approach |
| 4 | Neural Networks | No | others | Cited in introduction | Gu and Liu (2007) proposed a Bayesian biclustering model with a Gibbs sampling procedure for statistical inference. |
| 5 | BMC Genomics | No | review | Cited as selected paper in conference summary | The 2007 International Conference on Bioinformatics and Computational Biology (Biocomp) was designed in response to the fast development of above inter/multidisciplinary fields [28], 25 high-quality papers [29-53] were selected for a BMC Genomics supple- mentary issue. |
| 6 | ieee-acm transactions on computational biology and bioinformatics | No | others | Cited in background | Gibbs sampling is shown to be an effective option to derive deterministic solutions. A similar Bayesian biclustering model was also proposed by Gu and Liu [18]. |
| 7 | ieee-acm transactions on computational biology and bioinformatics | No | comparison | Cited in comparison | We applied the same biclustering algorithms used in [13]. The bbc algorithm [16] found only \_B2 and \_B3, but Table 11 shows that several measures evaluated the bbc’s solution as a perfect one, which again can be explained by their lack of the coverage property given by Definition 2. |
| 8 | journal of biopharmaceutical statistics | No | others | Cited in background | The biclustering technique identifies submatrices (biclusters) in a large AERS data matrix, each representing an association between a subset of AEs (rows) with a corresponding subset of drugs (columns) (Hartigan, 1972; Getz et al., 2000; Tanay et al., 2002; Ben-Dor et al., 2003; Yang et al., 2007; Gan et al., 2008; Kaiser and Leisch, 2008; Prelic et al., 2006; Barkow et al., 2006; Cheng and Church, 2000; Kluger et al., 2003; Carmona-Saez et al., 2006; Liu et al., 2003; Murali and Kasif, 2003; Lazzeroni and Owen, 2002; Hochreiter et al., 2010; Gu and Liu, 2008). |
| 9 | Methods | No | others | Cited in introduction and compared with the proposed method in evaluation section | We have performed the comparative experiments on the synthesized data sets by choosing other eight representative methods from literature, which are (1) MSR-based biclustering method by Cheng and Church (CC) [14]; (2) the order preserving submatrix algorithm (OPSM) [7]; (3) the iterative signature algorithm (ISA) [6]; (4) the Bayesian biclustering (BBC) [4]; (5) the extensive maximum similarity biclustering method (MSBE) [19]; |
| **BIOINFO DOMAIN** | | | | | |
| 1 | bioinformatics | No | others | Cited in introduction | Gu and Liu (2008) generalized the plaid models to fully generative models called Bayesian BiClustering model (BBC). To avoid the high percentage of overlap in the plaid models, BBC constrains the overlapping of biclusters to only one dimension. Further it allows different error variances per bicluster |
| 2 | breifings in bioinformatics | No | comparison | Cited in comparison | Bayesian biclustering (BBC) uses Gibbs sampling to fit a hierarchical Bayesian version of the plaid model [27]. It restricts overlaps to occur only in rows or columns, not both, so that two biclusters may not share the same data elements. The sampled posteriors for cluster membership of each row and column represent fuzzy membership; thresholding yields crisp clusters. |
| 3 | breifings in bioinformatics | No | others | Cited in future work | Such a Bayesian approach with an additive model was recently proposed for bi-clustering of gene expression data with missing data [73]. |
| 4 | current bioinformatics | No | review | Cited in review | A Bayesian BiCustering (BBC) model is proposed in [69] that uses Gibbs sampling. For a single bicluster, the same model as in the plaid model is assumed. Whereas for multiple biclusters, the overlapping of biclusters is allowed either in genes or conditions. Moreover, the authors used a flexible error model, which permits the error term of each bicluster to have a different variance. |
| 5 | international journal of data mining and bioinformatics | No | others | Cited as previous work | The proposed WF-MSB method was evaluated in comparison with MSBE on a real yeast microarray data and synthetic data sets. The experimental results show that WF-MSB can effectively find the biclusters with significant GO-based functional meanings. |
| 6 | current bioinformatics | No | comparison | Cited in comparison | In [46], Gu and Liu proposed a fully generative models called Bayesian biclustering algorithm (BBC) for gene expression data. |
| 7 | bmc bioinformatics | No | others | Cited in background | Methods for the identification of tight clusters, such as gene-shaving [15], bi-clustering [16] and contextdependent clustering [17], cannot identify hidden factors that act in linear combinations. |
| 8 | statistical methods in medical research | No | review | Cited in review | Spellman et al.18 Since a set of genes may be co-regulated only under a subset of the experimental conditions, a number of methods have been developed to cluster genes under specific conditions,55–57 and the gene clusters thus identified are sometimes called conditional specific expression modules. |
| 9 | bioinformatics | No | others | Cited in introduction | Because biclustering approach FABIA (factor analysis for bicluster acquisition) has shown good performance in recognizing biclusters and has been successfully used in identifying gene expression modulesand associated TFs (Gu and Liu, 2008; Hochreiter et al.,2010) |
| 10 | bmc systems biology | No | others | Cited in discussion | First of all, we analyzed CPM on a synthetic data in a simple but typical strategy adopted in the previous studies [23]. We produced a random data matrix with 10 rows and 15 columns. |
| 11 | plos computational biology | No | others | Cited in results | Biclustering approaches fall into four general categories. The first category assumes that each observed gene expression level for one sample is a linear combination of a mean effect, a row (gene) effect, and a column (sample) effect, some of which may be zero [23]. One approach in this category, Plaid, captures gene expression levels as a sum of many sparse submatrix components, where each submatrix includes non-zero values only for a subset of genes and subset of samples [30, 31]. |

**BiBit:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | DNA Research | No | others | Cited as Introduction | Sequence annotation is also important in transcriptomic experiments that aim to identify gene clusters with similar expression patterns that are linked to a particular biological process or experimental condition. Biological function can then be inferred from annotations shared within these clusters.7 |
| 2 | IEEE/ACM Transactions on Computational Biology and Bioinformatics | No | comparison | Cited as Comparison | Binary Biclustering methods that compare their solution to known solutions are in the following references: [24], [34], [35], [36]. |
| 3 | IEEE/ACM Transactions on Computational Biology and Bioinformatics | No | comparison | Cited as Comparison | The bibit algorithm [55] found 31 biclusters. All of them encompass a large portion of the data matrix, which led Sprec, Su, and Serec to attain 1. This is not dramatic per se because these three measures should be taken together with their pairs Sprel, Se, and Serel in an analysis. However, we believe that a good measure should evaluate the found solution as very poor because of the big difference in the number of biclusters |
| 4 | bmc genomics | No | others | Cited with other methods in the method section | In the field of biclustering, this is a well-known task that can be tackled with approximate pattern matching [17, 19, 20], non-negative matrix decomposition [21, 22], bipartite graph partitioning [23] or heuristic algorithms [24–27]. |
| **Bioinfo Domain** | | | | | |
| 1 | bioinformatics | No | others | Cited in discussion | Many other bi-clustering algorithms, including Bimax and the recently described BiBiT (Rodriguez-Baena et al., 2011), discover only homogenous bi-clusters and have low tolerance to noise and missing data. |
| 2 | bmc bioinformatics | Yes | usage | Applied in biclustering gene expression data from a tool named Mass-Up | Biclustering is performed with Bimax [42], a powerful algorithm capable of generating all optimal biclusters, and BiBit [43], a novel approach for the extraction of biclusters from binary datasets that can obtain similar results to Bimax by using significantly less computation time and reducing the total number of generated biclusters. |
| 3 | bmc bioinformatics | No | comparison | Cited in comparison | Bit-Pattern Biclustering Algorithm (BiBit) [31], which searches for maximal biclusters in binary datasets by applying the logical AND operator over all possible gene pairs; |

Bimax:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | Nature | Yes | usage | Applied for breakpoint detection on copy number analysis | We also applied biclustering16 to construct a heat map of breakpoints, and ordered it on the basis of the copy number tree to show which breakpoints were common or divergent between the major subpopulations (Supplemen- tary Fig. 7a) |
| 2 | Molecular Systems Biology | No | review | Cited in review | Such systematic evaluation efforts were recently performed in other fields, for example, discovery of TF-binding sites (Tompa et al, 2005), biclustering of expression data (Prelic et al, 2006) and protein structure prediction (Kryshtafovych et al, 2005), which has a long and successful history of community evaluations |
| 3 | Nature Reviews Genetics | No | review | Cited in review | Given the rapid emergence of integrative methodologies, some effort has been made to compare their accuracy (pre- cision), sensitivity (recall) or computational efficiency within individual method classes44–46. |
| 4 | Plant Physiology | No | others | Cited in introduction to describe problems with traditional clustering. However, the use clustering over biclustering stating some reasons | The main disadvantages of the approach are the complexity of its data output, the lack of predefined boundaries between clusters, and its weaker perfor- mance in identifying local expression similarities in a small subset of the samples (Prelic et al., 2006). How- ever, most of these challenges can be overcome by applying efficient postprocessing methods of the ob- tained dendrograms, such as tree cutting methods (Gutierrez et al., 2007). |
| 5 | Nucleic Acids Research | No | others | They used benchmark datasets introduced also in Bimax paper | We now show the application results of QUBIC first on a number of benchmark data sets developed by Prelic et al. (13) |
| 6 | Nature Chemical Biology | No | review | Cited in review | Newer methods allow for genes to participate in multiple clusters and for those clusters to be condition-specific7,40–47 |
| 7 | Plant Cell | Yes | usage | Used as only option for clustering analysis of gene expression | To identify groups of genes with similar expression profiles in a subset of conditions, the Biclustering tool (BiMax algorithm; Prelic et al., 2006) of Genevestigator V3 (https://www.genevestigator.ethz.ch; Zimmermann et al., 2004) was used. |
| 8 | Plant Cell | No | others | Not applied, but they applied another biclustering algorithm (LPSA) | We applied this method to the entire data set to discover groups of genes sharing compatible expression patterns across subsets of samples (Madeira and Oliveira, 2004; Prelic et al., 2006). |
| 9 | Plant Physiology | No | others | Cited in introduction as an example | Breaking down these effects into local patterns (e.g. by using a biclustering algorithm; Prelic et al., 2006) helps in finding out conditions that are more directly linked to the gene of interest. |
| 10 | ieee-acm transactions on computational biology and bioinformatics | No | others | Cited in introduction as an example | Moreover, the inherent difficulty of this problem when dealing with the original expression matrix and the great interest in finding coherent behaviors regardless of the exact numeric values in the matrix, has led many authors to a formulation based on a discretized version of the expression matrix [3], [12], [14], [16]–[20], [23], [29], [32], [34], [36], [40]. The discretized versions remain, in general, NP-hard. |
| 11 | toxicological sciences | No | review | Cited in review | However, methods such as biclustering (Cheng and Church, 2000; Prelic et al., 2006) were developed to partition the two-dimensional matrix of gene expression data into subsets of genes sharing compatible expression patterns across subsets of samples (so-called cliques). |
| 12 | ieee-acm transactions on computational biology and bioinformatics | No | comparison | Cited in comparison | Recently, Preli c et al. [45] evaluate the performance of fiveprominent coclustering algorithms and provide synthetic gene expression data sets. |
| 13 | plant cell | No | others | Cited in current methods | A variety of network clustering algorithms have been developed to identify modules in networks, including molecular complex detection (Bader and Hogue, 2003), Markov clustering algorithm (Enright et al., 2002), MCLUST (Fraley and Raftery, 2003), biclustering (Preli c et al., 2006), and the heuristic cluster chiseling algorithm (Mutwil et al., 2010). |
| 14 | molecular systems biology | Yes | usage | Applied for biclustering correlated genes with drugs | Along those lines, we used a biclustering approach that is well-suited for revealing the modular organization of transcriptional responses to drug perturbation (Ihmels et al, 2002; Prelic´ et al, 2006), as it can group coregulated genes with the drugs they respond to (technically, each bicluster consists of both a gene and a drug subset). |
| 15 | clinical pharmacology & therapeutics | Yes | usage | Applied for biclustering drug events in FDA’s Spontaneous Reporting System | Binary inclusion-maximal biclustering (Bimax)24 used in work, is one of the several biclustering algorithms included in Biclustering Analysis Toolbox25, which is a software platform for biclustering data analysis and visualization. |
| **Bioinfo Domain** | | | | | |
| 1 | bioinformatics | Yes (self-citation) | usage | Used as a biclustering method in BicAT tool | Bimax, an exact biclustering algorithm based on a divide-andconquer strategy that is capable of finding all maximal bicliques in a corresponding graph-based matrix representation (Prelic et al., 2005). |
| 2 | bmc bioinformatics | No | others | Cited in background | So-called biclustering (clustering both genes and experimental conditions), is a widely studied problem and many different approaches to it have been published [6,25,52,76,80,86,98]. |
| 3 | bioinformatics | No | others | Cited in introduction but they use the same evaluating measure | Tanay et al. (2002) and Prelic´ et al. (2006) focused on finding bi-clusters of up-regulated expression values or down-regulated expression values. |
| 4 | bioinformatics | No | others | Cited in introduction but they use the same evaluating data | To simplify this task, some methods discretize the data in a first step, like xMOTIF (Murali and Kasif, 2003) or Bimax (Prelic et al., 2006) which even binarizes the data and searches for blocks with an enrichment of ones. |
| 5 | breifings in bioinformatics | No | comparison | Cited in comparison | Prelic´ et al. compared several algorithms on both synthetic data with constant and constant-column biclusters and on real data [14] |
| 6 | bmc bioinformatics | No | others | Cited in background and they used the same method of evaluating | Prelic et al. [18] have compared many biclustering algorithms using the additive model. |
| 7 | bmc bioinformatics | No | comparison | Cited in introduction but they use the same evaluating measure | This requires simultaneous clustering along both the row and column directions, and is often called biclustering [11-16]. |
| 8 | bioinformatics | No | others | Cited in methods as reference for biclustering comparison | Biclustering has been popular in expression profiles studies as it attempts to find a subset of genes having similar expression patterns under a group of conditions. Such an entity is often called a module. For a comparison of various biclustering algorithms, see Prelic et al. |
| 9 | bmc bioinformatics | No | others | Cited in discussion | In its current form, the AutoSOME method does not identify genes whose co-regulation is restricted to a subset of arrays (see e.g., [37-39]). |
| 10 | bmc bioinformatics | No | comparison | Cited in backgroud, comparison and discussion. They used the same measures of comparison | The biclustering methods Bimax [32] and Samba [33] discretize the expression level which allows them to enumerate a large number of biclusters in less time than more complicated models. We also tested our proposed methodology on a data set corresponding to synthetic gene expression data created by an artificial model [32]. |
| 11 | bioinformatics | No | others | Cited in introduction | Prelic et al. (2006) have compared performance of different biclustering algorithms, and proposed a fast divide-and-conquer biclustering algorithm (Bimax). |
| 12 | bmc bioinformatics | No | others | Cited in introduction | For an integrative analysis of 3D gene-condition-time datasets with standard clustering or biclustering approaches, these datasets have often been projected onto a single gene-condition matrix, with each time-point labeled as a separate condition [14,19,20]. |
| 13 | BioData Mining | No | others | Cited in introduction | Representative examples of systematic search methods include, among others, greedy algorithms [1,10-14], divide and conquer algorithms [7,15] and enumeration algorithms [16-18]. |
| 14 | plos computational biology | Yes | usage | Applied for biclustering of human proteins that share the same set of HIV-1 interactions | To find sets of human proteins thatshare the same set of HIV-1 interactions in this matrix, biclustering was performed using the Bimax algorithm [101]. |
| 15 | bioinformatics | Yes | usage | Used as a biclustering method in BicOverlapper tool | For demonstration purposes, the tool has been applied to the biclustering results of the Order Preserving SubMatrix search algorithm (OPSM) (Ben-Dor et al., 2003) and Bimax (Prelic et al., 2006) in the analysis of a microarray data matrix containing two types of Diffuse Large B-Cell Lymphomas |
| 16 | algorithms for molecular biology | No | others | Cited in backgroud | Furthermore, the inherent difficulty of this problem when dealing with the original real-valued expression matrix and the great interest in finding coherent behaviors regardless of the exact numeric values in the matrix, has led many authors to a formulation based on a discretized version of the expression matrix [7-9,12-23]. |
| 17 | journal of theoretical biology | No | others | Cited in introduction | A systematic comparison of some biclustering methods is made in [30]. |

**CCA:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **BIOMED DOMAIN** | | | | | |
| 1 | Nature | No | others | Cited in future work | Previous studies of sequence alignment of regulatory domains of orthologous genes in multiple species has shown a remarkable correlation between sequence conservation, dubbed `phylogenetic footprints'441, and the presence of binding motifs for transcription factors. This approach could be particularly powerful if combined with expression array technologies that identify cohorts of genes that are coordinately regulated, implicating a common set of cisacting regulatory sequences442. |
| 2 | IEEE/ACM Transactions on Computational Biology and Bioinformatics | No | review | Cited in review | Cheng and Church [6] were the first to apply biclustering to gene expression data. Given a data matrix and a maximum acceptable mean squared residue score (see (17)), Æ , the goal is to find Æ-biclusters, that is, subsets of rows and subsets of columns, with a score no larger than Æ (see Section III-D). |
| 3 | Nature Genetics | No | others | Cited in introduction | Gene classification is an essential task in studying the global structure of the transcriptional network. But although standard clustering methods classify genes successfully when applied to relatively small data sets, their use in the analysis of large-scale expression data is limited by two well-recognized drawbacks1–3. |
| 4 | Genome Research | No | others | Cited in background | The idea of simultaneous clustering of rows and columns of a matrix goes back to (Hartigan 1972). Methods for simultaneous clustering of genes and conditions were more recently proposed (Cheng and Church 20002000; Getz et al. 2000; Lazzeroni and Owen 2002). |
| 5 | Cell | No | review | Cited in review | A particular advantage of PSMF over another multiway factorizating technique, bi-clustering (Cheng and Church, 2000), is that PSMF allows each cluster to be defined by an arbitrary set of genes and compounds, whereas bi-clustering is restricted so that any two clusters containing the same gene (or compound) must be defined by exactly the same set of genes (or compounds). |
| 6 | Genome Biology | No | others | Cited in background | Biclustering was developed to address better the full com- plexity of finding co-regulated genes under multifactor con- trol by grouping genes on the basis of coherence under subsets of observed conditions [10,16-22]. |
| 7 | Drug Discovery Today | No | review | Cited in review | Global analysis of combination screens can reveal patterns of synergy between particular drug classes or mechanisms, as recently shown for antibacterial combinations [31]. The synergy scores for each combination can be displayed on a color scale in a grid format (Figure 2c) where the scores in each row or column represent the synergy profile of an individual drug with the rest of the library. The drugs on each axis of the grid can be sorted by mechanistic class or primary target, and clusters of high synergy scores show pairs of molecular targets that interact synergistically (Figure 2c). Alternatively, the synergy profiles for the components of the combination screen can be clustered or bi-clustered [32] to identify regions of synergy that can reveal multi-target interactions between alternate targets of the components. |
| 8 | nature reviews microbiology | No | review | Cited in review | Global versus query-driven inference. Global module inference methods22,52–59 search for the modules that explain most of the data. |
| 9 | molecular systems biology | No | others | cited as previous work | we employed a biclustering algorithm (Materials and methods). Unlike the single-dimension clustering scheme used to group the lowpleiotropy mutants, biclustering methods (Cheng and Church, 2000; Getz et al, 2000; Segal et al, 2001; Tanay et al, 2002) use statistical parameters to select sets of genes that share common phenotypes across a subset of conditions in a profile. |
| 10 | ieee-acm transactions on computational biology and bioinformatics | No | comparison | Cited in comparison | For the purpose of comparison, we applied the k-means algorithm [41], Kohonen’s SOM [33], and the biclustering algorithm [14] to the synthetic dataset. |
| **Bioinfo Domain** | | | | | |
| 1 | biometrics | No | others | Cited in a book |  |
| 2 | bioinformatics | No | others | Cited in introduction and compared with the proposed algorithm | Biclustering was introduced in the seventies (Hartigan, 1975), Cheng and Church (2000) were the first to apply it to gene expression data, |
| 3 | bioinformatics | No | comparison | Cited in comparison | This modified clustering concept can be useful to uncover processes that are active only over some but not all samples as has been demonstrated in several studies (Cheng and Church, 2000; Ihmels et al., 2002; Ben-Dor et al., 2002; Tanay et al., 2002; Murali and Kasif, 2003), |
| 4 | journal of computational biology | No | others | Cited in introduction and compared with the proposed algorithm | This modified clustering concept can be useful to uncover processes that are active only over some but not all samples as has been demonstrated in several studies (Cheng and Church, 2000; Ihmels et al., 2002; Ben-Dor et al., 2002; Tanay et al., 2002; Murali and Kasif, 2003), |
| 5 | bioinformatics | yes | usage | Applied for biclustering of gene expression data | Bi-clustering Bi-clusters were generated using the program ‘Bi-cluster’ published in (Cheng and Church, 2000) (available at http://cheng.ececs.uc.edu/biclustering). |
| 6 | bioinformatics | No | others | Cited in future work | CODENSEcan be further modified and extended to identify more subtle patterns. For example, instead of clustering edges with overall similar edge support vectors, one may use biclustering algorithms (Cheng and Church, 2000) to identify edges showing similar supports in a subset of datasets. |
| 7 | bioinformatics | yes | usage | Applied as a biclustering method on BicAT tool | BicAT implements the following biclustering methods: (i) Cheng and Church’s algorithm (CC) which is based on a mean squared residue score (Cheng and Church, 2000); |
| 8 | bioinformatics | No | others | Cited in introduction | Other recent biclustering approaches (Lazzeroni and Owen, 2000; Cheng and Church, 2000; Segal et al., 2001; Ben-Dor et al., 2001; Tanay et al., 2002) rely on a variety of optimization procedures. |
| 9 | bmc bioinformatics | No | others | Cited in background and compared with the proposed algorithm | For our modeling purposes, only methods which derive biclusters with coherent, or correlated, gene profiles, such as those of Cheng and Church [25], Yang et al. [98], and Lazzeroni and Owen [53] are suitable. |
| 10 | bmc bioinformatics | No | others | Cited in introduction in a paper about comparison of clustering methods | A further aggressive approach known as bi-clustering or two-way clustering searches for groups of genes that have similar expression pattern only in a subset of samples or time periods (Cheng and Church, 2000). |
| 11 | computers in biology and medicine | No | comparison | Cited in comparison | Cheng and Church [18] adapted the method of Hartigan [46] to GE data. Their algorithm proceeds by scoring sub-matrices of the GE matrix using the H-score (Eq. 1, Fig. 1) [11]. |
| 12 | bioinformatics | No | others | Cited in introduction | Bi-clustering methods seek gene similarity in only a subset of available conditions, which is more appropriate for functionally heterogeneous data (Cheng and Church, 2000; Madeira and Oliveira, 2004). |

COALESCE:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | Nucleic Acids Research | No | others | Cited in introduction | These federated data sets allow the construction of interaction networks from various data sources, which can be used to explain biological processes, interpret different orthogonal data types and predict new biological functions (6,7). |
| 2 | Genetics | No | others | Cited as an interesting approach for the future (2011) | While exciting approaches for combining heterogeneous data are being developed (Lage et al. 2008; Huttenhower et al. 2009; Lee et al. 2009; Battle et al. 2010), these issues must be taken into consideration in the design and interpretation of truly integrative systems genetics analyses. |
| 3 | Nature Reviews Microbiology | No\* | review | Cited in review, several times, by name | COALESCE (combinatorial algorithm for expression- and sequence-based cluster extraction)48 is an integrative, non-supervised module inference procedure that uses |
| 4 | Nature Reviews Cancer | No | review | Cited in review | Bayesian variable selection has been suc- cessfully applied to situations that comprise one data type90, but it could be extended to multiple data types using similar fundamental biological assumptions as in Huttenhower et al.69 |
| 5 | Molecular Systems Biology | No | others | Cited in introduction | Computa- tional methods have also been developed for integrating various types of functional genomics data and using them to create models of regulatory networks and other interactions in the cell (Alon, 2006; Huttenhower et al, 2009; Myers et al, 2009). |
| 6 | Genome Biology | No\* | comparison | Used in biclustering comparison with good results | Additional genome- wide data (such as association networks and transcription factor binding sites) greatly improves the performance of these approaches [19-22]. |
| 7 | Molecular and Cellular Endocrinology | No | review | Cited in review | As a concrete example Huttenhower et al. (2009) have shown the effectiveness of Bayesian data integration in predicting regulatory modules from metazoan data collections. |
| 8 | Nucleic Acids Research | No\* | comparison | Used in biclustering comparison with good results | e evaluated its performance relative to both cMonkey1, to other popu- lar clustering methods––k-means (41) and WGCNA (42), and to published data integration/module detection algo- rithms COALESCE (11), DISTILLER (10) and LeMoNe (4). |
| 9 | bmc genomics | No | others | Cited in background | On the other hand, gene-module based methods take advantages of biclustering algorithms [7,8] or item-set mining algorithms [9] to detect a cluster of genes which share similar patterns on a subset of gene conditions. |
| **Bioinfo Domain** | | | | | |
| 1 | breifings in bioinformatics | No | comparison | Used in biclustering comparison | Combinatorial algorithm for expression and sequence-based cluster extraction (COALESCE) is a nondeterministic greedy algorithm that seeks biclusters representing regulatory modules in genetics [28]. |
| 2 | bmc bioinformatics | No | others | Cited in backgroud | gene expression modules, which in addition to bi-clustering structure also incorporate information about gene expression regulation [9-12]. |
| 3 | bioinformatics | No | comparison | Cited in the comparison section with good results and use of similar process as COALESCE | Iterative module extraction The bi-clustering approach described so far is able to find one module at a time. In order to find all possible clusters, an iterative approach similar to that described by Huttenhower et al. (2009) was applied. |
| 4 | bmc bioinformatics | No | review | Cited in review | Coalesce [30] is a combinatorial algorithm specifically developed for the identification of regulatory modules from the analysis of gene expression and DNA sequence data. |
| 5 | bioinformatics | No | others | Cited in introduction and as a method of inferring regulatory networks | Current methods (Ernst et al., 2008; Faith et al., 2007; Friedman et al., 1998; Hecker et al., 2009; Huttenhower et al., 2009; Ma et al., 2006; Marbach et al., 2010, 2012; Margolin et al., 2013; Michoel et al., 2009; Mordelet and Vert, 2008; Park et al., 2010; Pique-Regi et al., 2011; Poultney et al., 2012; Prill et al., 2010, 2011; Yu et al., 2004; Zhu and Guan 2014; Zhu et al., 2014; Zou and Conzen, 2005) for modeling regulatory networks based on expression data can be divided into two categories. |
| 6 | bmc bioinformatics | No | others | Cited as previous work | While several approaches were considered [7,8,16], we have chosen to compare TIDAL’s results with those of the Dynamic Regulatory Events Miner (DREM) [13], |
| 7 | plos computational biology | No | others | Cited in introduction | This statistical model encodes a general framework known as biclustering [23–35]. A biclustering model decomposes a matrix into clusters that each correspond to a subset of samples and a subset of features that exhibit latent structure unique to those subsets. |
| 8 | bioinformatics | No | others | Cited in introduction | These integrative methods have been the most successful to date in predicting protein function (Fraser and Marcotte, 2004; Guan et al., 2008; Llewellyn and Eisenberg, 2008) and networks of functional relationships between proteins (Hess et al., 2009; Huttenhower et al., 2009; Ideker et al., 2002; Jansen et al., 2003; Jensen et al., 2009; Lee et al., 2004; Myers et al., 2005; Troyanskaya et al., 2003; von Mering et al., 2005) from large collections of data. |
| 9 | bmc bioinformatics | No | others | Cited in discussion | During last years, numerous integrative approaches that search for regulatory elements have been developed by incorporating co-expression datasets and/or ontology annotation within an unique algorithm to improve the discovery of regulatory modules in various organisms such as archeae [45], bacteria [46], yeast [47] and human [48,49]. |
| 10 | bmc bioinformatics | No | comparison | Cited in comparison | Combinatorial Algorithm for Expression and Sequence-Based Cluster Extraction (COALESCE) [28], which initializes a bicluster with two maximally correlated genes across all experimental conditions, adding columns to it according to a standard z-test and inserting rows based on posterior probability. |

**CPB:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | BMC Genomics | No | others | Cited in background | Bozda˘ g et al. [15] proposed the Correlated Pattern Biclusters (CPB) algorithm, which discovers biclusters by setting reference genes with randomly selected columns, and then adding rows with high correlation and determining columns that have a smaller Root Mean Squared Error. In this case, the search space can be restricted again by the randomly selected seeds of columns. |
| 2 | Cancer biology & Therapy | Yes (Self-citation) | usage | Applied in biclustering of microarray data | We performed a computational analysis of publicly available breast cancer microarray data sets to identify genes that are co-regulated with the reference genes BRCA1, BRCA2, and BARD1,34-36 and also a novel two-step query based biclustering approach, called correlated patterns biclusters (CPB).37,38 |
| 3 | IEEE Transactions on Nanobioscience | No | others | Cited in background | Several correlation functions, based on explicit thermodynamic modeling of gene regulatory networks, have been combined with biclustering to identify functionally enriched groups followed by classification [23]. |
| 4 | nucleic acids research | No | others | Cited in the comparison section | MCbiclust was compared with the FABIA (36), FABIAS (36), biMax (37), CC (28), Plaid (38), ISA (39), FLOC (40), QUBIC (41), CPB (42) and CTWC (43) biclustering methods (see Supplementary Table S1) all run withdefault parameters. |
| **Bioinfo Domain** | | | | | |
| 1 | breifings in bioinformatics | No\* | comprison | Cited in the comparison | Correlated pattern biclusters (CPB) is a nondeterministic greedy algorithm that seeks biclusters with high row-wise correlation according to the Pearson Correlation Coefficient (PCC) [29]. |
| 2 | plos computational biology | No | others | Cited in the results section | The second category of biclustering methods uses hierarchical clustering to group together similar samples and features [3]. For example, samples may be clustered by considering some measure of feature similarity [24, 25, 28, 32, 35]. |
| 3 | bioinformatics | No | review | Cited in the related work as a review | These methods are CC (Cheng and Church, 2000), Plaid (Lazzeroni and Owen, 2002; Turner et al., 2005), FLOC (Yang et al., 2003), ISA (Bergmann et al., 2003), xMOTIFs (Murali and Kasif, 2003), Spectral (Kluger et al., 2003), SAMBA (Tanay et al., 2004), Bimax (Preli c et al., 2006), BBC (Gu and Liu, 2008), QUBIC (Li et al., 2009), CPB (Bozda g et al., 2009), iBBiG (Gusenleitner et al., 2012) and BicPAM (Henriques and Madeira, 2014). |
| 4 | bmc bioinformatics | No | comprison | Cited in comparison | Correlated Pattern Biclusters (CPB) [29], which starts with randomly selected subsets of rows and columns, and searches for biclusters with highly correlated rows regarding some reference genes; |

**DeBi:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | ieee-acm transactions on computational biology and bioinformatics | No | others | Cited in background | This work integrates principles to discover biclusters under a plaid assumption with principles from exhaustive yet scalable searches to discover flexible structures of biclusters [3], [22], [39], [47]. |
| 2 | Gene | No | others | Cited in background | Another recently proposed differentially expressed biclustering algorithm called DeBi (Akdes and Martin, 2011) uses frequent pattern mining method to discover maximum size homogeneous biclusters in which all genes are co-expressed under a subset of samples. In fact, DeBi aims to find biclusters in onemicroarray dataset, which is not the same as our proposed differential co-expression bicluster that is produced from two microarray datasets. |
| 3 | ieee-acm transactions on computational biology and bioinformatics | No | others | Cited as previous work | We propose an algorithm that can analyze datasets with a large attribute set at different densities, and can operate on a laptop, which makes it accessible to practitioners. EMFP produces biclusters that have a very low Root Mean Squared Error and false positive rate, with very few type II errors. |
| **Bioinfo Domain** | | | | | |
| 1 | bioinformatics | Yes | Usage | used in the step of finding multiple modules for time-series data | The second heuristic, called filter, takes a set of biclusters U as input and produces a reduced set. It first uses the overlap reduction method of Serin and Vingron (2011): |
| 2 | breifings in bioinformatics | No | review | Cited in review | While biclustering is not exclusively a frequent itemset mining problem, frequent itemset mining-based algorithms have been shown to perform equally or superior to various other methods [29]. |
| 3 | algorithms for molecular biology | No | others | Cited background | Other works are based on a previous binarization of the data, such as DeBi [44]. After binarizing the data, DeBi consist of three stages for finding, extending and filtering seed bicluters. Although no evaluation measure for biclusters is defined, Fisher exact text is used in the extending phase. |
| 4 | algorithms for molecular biology | No | others | Cited in introduction | Interestingly, the existing patternbased approaches for biclustering – such as BiModule [13], DeBi [10], RAP [14] and GenMiner [15] – provide complementary principles of interest for this field. |
| 5 | bmc bioinformatics | No | others | Cited in background | Complementary, statistical evaluation has been proposed based on biclusters’ expected probability of occurrence [18,19] or based on their enrichment p-values against real datasets [20-22]. |
| 6 | algorithms for molecular biology | No | others | Cited in background | This mapping1 led to the development of several pattern-based approaches for biclustering [15, 22–24]. |
| 7 | bioinformatics | No | others | Cited in related work | Since then, extensive efforts have been made in both computer science and statistics to develop different types of biclustering methods (e.g. Bergmann et al. 2003; Ben-Dor et al., 2003; Bozda g et al., 2009; Cho et al., 2004; Gu and Liu, 2008; Gusenleitner et al., 2012; Henriques and Madeira, 2014, 2015; Hochreiter et al., 2010; Huttenhower et al., 2009; Kluger et al. 2003; Lazzeroni and Owen, 2002; Liu and Wang, 2003; Li et al., 2009; Murali and Kasif, 2003; Pandey et al., 2009; Preli c et al., 2006; Serin and Vingron, 2011; Sheng et al., 2003; Tanay et al., 2002, 2004; Turner et al., 2005; Yang et al., 2002, 2003; Wang et al., 2002). |
| 8 | algorithms for molecular biology | No | others | Cited in introduction | Recent attempts to perform biclustering based on enhanced pattern mining searches [8, 12, 13], termed as pattern-based biclustering, showed the unprecedented possibility to efficiently discover arbitrarily positioned biclusters with parameterizable size, coherency and quality [2, 14]. |
| 9 | bmc bioinformatics | No | comparison | Cited in comparison | Differentially Expressed Biclusters (DeBi) [32], an algorithm based on a frequent itemset approach that applies a depthfirst traversal on an enumeration tree to discover hidden patterns in data. |
| 10 | bmc bioinformatics | No | others | Cited in background | BicPAMS makes available earlier pattern-based biclustering algorithms (including BicPAM [11], BiModule [16] and DeBi [15]), well suited for expression data analysis. |
| 11 | bioinformatics | No | others | Cited in introduction | Bimax (Prelic et al., 2006) is one of most popular biclustering algorithms (Bhattacharya et al., 2009; DiMaggio et al., 2008; Harpaz et al., 2011; Serin et al., 2011) and finds all the inclusive maximal biclusters such that all their elements are 1. |

**FABIA:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | Science | Yes | Usage | Used for biclustering microbiome data | Next, we identified sample subsets with specific taxonomic signatures using a biclustering approach (33). Two stable biclusters were detected, spanning 410 and 374 samples respectively, with an intersection of 92 (Table S14). |
| 2 | Nucleic Acids Research | Yes (self-citation) | Usage | Used for haplotype identification | We use the ‘FABIA’ biclustering model (36). In contrast to other biclustering methods such as BIMAX (37) and QUBIC (38), FABIA can represent homozygous regions where the same IBD segment may be present in one diploid individual two times. |
| 3 | Drug Discovery Today | Yes (self-citation) | Usage | Used for transcriptional module identification | Differentially expressed genes [53] were called and transcriptional modules [54] (i.e. genes where expression is simultaneously up- or down- regulated in a subset of samples) were identified |
| 4 | Nucleic Acids Research | No | others | Implemented on a web server | Additionally, other basic biclustering methods have been implemented including Cheng and Church (21), FLexible Overlapped biClustering (FLOC) (22) and Factor Analysis for BIcluster Acquisition (FABIA) (23) (Supplementary Table S1), which are also extended by the generalized fac- torization method. |
| 5 | ieee-acm transactions on computational biology and bioinformatics | No | others | Cited in background | Stochastic approaches assume that biclusters follow multivariate distributions [26] and learn their parameters by maximizing a likelihood (merit) function. |
| 6 | ieee-acm transactions on computational biology and bioinformatics | No | others | Cited in related work | FABIA [15] uses factor analysis in a multiplicative model to extract biclusters in a gene expression data set. The authors represent linear dependency between a subset of rows and columns as an outer product of a prototype column vector and a factor vector with which the prototype vector is scaled over samples. |
| 7 | international journal of molecular sciences | Yes | Usage | Used for biclustering gene expression profiles | Then, a biclustering analysis was performed using FABIA method to identify gene modules. The biological functions of gene modules provide preliminary insights into pharmacological mechanisms of both natural medicines. |
| 8 | nucleic acids research | No | others | Cited in the evaluation section to perform the defined method | To evaluate the performance of Bi-Force on biclustering problems, we referenced to the work of Eren et al. (17). Eight (out of twelve) prevalent online available biclustering algorithms were downloaded, including Cheng and Church (7), BiMax (16), FABIA (32), ISA (33), Plaid (26), QUBIC (34), Spectral (35) and xMOTIFs (36). |
| 9 | ieee-acm transactions on computational biology and bioinformatics | No | comparison | Cited in comparison | Hochreiter et al. [37] stated that previous measures designed specifically for biclusterings neither account for overlapping biclusters nor consider the number of biclusters in the found and reference solutions. |
| 10 | bmc genomics | No | others | Cited in background | Given a microarray gene-expression matrix, comprised of the rows of genes and the columns of samples (or conditions), biclustering has been the most common method extracting RMs defined as a bi-set of co-regulated genes and coregulating conditions [5-11]. |
| **Bioinfo Domain** | | | | | |
| 1 | breifings in bioinformatics | No | comparison | Cited in comparison | Factor analysis for bicluster acquisition (FABIA) models the data matrix X as the sum of p biclusters plus additive noise W, where each bicluster is the outer product of two sparse vectors [31]: |
| 2 | algorithms for molecular biology | No | others | use a measure of evaluation to evaluate the proposed method | In order to assess the performance of different biclustering algorithms, we used two measures from Prelic et al. [10] and Hochreiter et al. [21], respectively. |
| 3 | bioinformatics | No | others | Cited in introduction | the Bi-correlation clustering algorithm (BCCA) by Bhattacharya and De (2009) and factor analysis for bicluster acquisition (FABIA; Hochreiter et al., 2010). |
| 4 | bmc bioinformatics | No | others | Cited in discussion of the results | For applications of component models with sparse loadings to microarray gene expression data, see [28] and [13]. |
| 5 | algorithms for molecular biology | No | others | Cited in background | Hochreiter et al. [41] have developed a generative multiplicative model for the biclustering problem, assuming realistic non-Gaussian signal distributions with heavy tails. |
| 6 | bioinformatics | No | others | Cited in introduction and compared with the proposed method | It scales well when applied to hundreds of datasets is tolerant to noise characteristic of genomics data and when applied on simulated data, outperforms clustering and bi-clustering methods including hierarchical and k-means clustering, FABIA (Hochreiter et al., 2010), COALESCE (Huttenhower et al., 2009) and Bimax (Prelic, 2005). |
| 7 | bmc bioinformatics | Yes (self-citation) | Usage | Used as principle method in the visualization tool | We use FABIA to demonstrate the proposed technique; however, note that any other biclustering algorithm that produces overlapping clusters can be used in the same way. |
| 8 | algorithms for molecular biology | No | others | Cited in introduction | Biclustering has been also applied to group mutations and copy number variations [2], to analyze biological networks [3], and to study translational [4], chemical [5] or nutritional data [6]. |
| 9 | biodata mining | No | comparison | Cited in comparison | The selected algorithms are BIMAX [5], FABIA(Factor Analysis for Bicluster Acquisition) [13], ISA (Iterate Signature Algorithm) [3], QUBIC (Qualitative Biclustering algorithm) [14] and SAMBA (Statistical-Algorithmic Method for Bicluster Analysis) [4]. |
| 10 | plos computational biology | Yes | Usage | Used for biclustering of T-box (TBX) genes | Using the Factor Analysis for Bicluster Acquisition (FABIA) algorithm [44], five, ten, then twenty biclusters were identified. Several probe sets grouped together in this way were also in transcriptional regulatory relationships per the CLR and ARACNE learning. |
| 11 | bmc bioinformatics | No | others | Cited in background | The evaluation of order-preserving solutions does not significantly differ from the evaluation of traditional biclustering solutions. When considering the knowledge of hidden biclusters, relative non-intersecting area (RNIA) [15], match scores [3,16] and clustering metrics (e.g. entropy, recall and precision) have been adopted. RNIA [15] measures the overlap area between the hidden and found biclusters. |
| 12 | bioinformatics | Yes | Usage | Used for biclustering of gene expression data | In this study, by using biclustering approach FABIA (factor analysis for bicluster acquisition), we generate 49 modules for gene expression profiles on 1309 agent treatments. |
| 13 | journal of theoretical biology | No | comparison | cited in comparison | As a result, the accuracy o fFabia for identifying the column clusters is 67%. For Clusters 1, 3, and 4, GSGBC obtains the highest accuracy while GBC provide highest accuracy in Cluster 2. However, the bicluster size obtained from GBC is slightly larger than that by GSGBC. |
| 14 | bioinformatics | Yes | Usage | Used and extended to define sparse group factor analysis as a way of inferring biclusters from heterogeneous multi-source data. | The parameters are inferred with variational expectation maximization, see Hochreiter et al. (2010) for details. |

**ISA:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | Nature | No | comparison | Cited in comparison | To assess the extent of this correlation, we made use of a study that identified 33 transcriptional ‘modules’ of genes with marked co-regulation based on analysis of over 1,000 microarray data sets reflecting the results of different mutant strain backgrounds or environmental perturbations38,40. |
| 2 | Nature | No | comparison | Cited in comparison | we compared our results with global transcriptional array data. A recent analysis of mRNA expression profiles from ,1,000 published microarray experiments allowed for the identification of 33 ‘modules’ of transcriptionally co-regulated genes18,19. |
| 3 | PLOS Biology | Yes (self-citation) | Usage | Comparison of regulatory structures | To compare the higher-order regulatory struc- tures more systematically, we decomposed the expression data of each organism into a set of transcription modules using the iterative signature algorithm (ISA) we proposed recently (Bergmann et al. 2003; J. Ihmels, unpublished data). |
| 4 | Toxicology Pathology | No | others | Citation in background | To some degree, the wide use of cluster analysis in classifying toxicity and disease (Abe et al., 2003; Bergmann et al., 2003) seems to favor toxicity or disease characterization by transcript signature profile over a single biomarker. |
| 5 | PLOS Genetics | Yes (self-citation) | Usage | Identify regulatory relationships | In order to identify novel regulatory relationships that are not confined to specific function-related genes, we conducted a complementary, unsupervised analysis of the C. albicans expression data. To this end, we used the iterative signature algorithm (ISA) [31,44] to determine the modular organization of the C. albicans transcription program |
| 6 | BMC Genomics | No | comparison | Used in intercomparison | Bergmann et al [4] introduced the iterative signature algorithm (ISA), which searches bicluster modules iteratively based on two pre-determined thresholds. |
| 7 | Science | No | others | Cited in conclusion | We expect that these signatures therefore represent fundamental small-molecule response systems that are present across eukaryotic cells. Accordingly, we expect that many of our 317 chemical-genetic probes will be directly applicable to mammalian cell biology and may support novel targets as opportunities to pursue for therapeutic intervention (5, 22, 23). |
| 8 | genome biology | No | others | Cited in background | This observation has spurred the development of condition-specific classification of multiple or large-scale gene expression data. [7-11]. |
| 9 | nature reviews genetics | No | review | Cited in review | However, a more powerful approach may be the joint analysis of gene modules (that is, groups of genes with coherent expression patterns across a subset of samples), which could potentially include different combinations of species, organs and conditions102. |
| 10 | molecular systems biology | Yes | Usage | Used for biclustering of genes and drugs that coherently regulate these genes | Drug-induced transcriptional modules were detected and tested for statistical significance separately in each of the four matrices of expression data using an unsupervised biclustering approach that has previously been shown to maintain high accuracy even with noisy input data (Iterative Signature Algorithm (ISA) |
| 11 | human molecular genetics | Yes | Usage | Used for biclustering of gene expression data from different RI strains. | We used the ISA (1,6) to identify transcription modules. We first tested the performance of ISA when applied to gene expression traits using simulated data and selected parameters for optimal performance. |
| 12 | nucleic acids research | No | others | Cited in discussion | Subspace-clustering algorithms have been developed to analyse 2D microarray data (also known as biclustering in microarray data analysis literature), associating subsets of genes whose expression are coherent under a subset of conditions (46–50). |
| 13 | nucleic acids research | Yes | Usage | Used for biclustering of gene expression data | Biclusters were built with the Iterative Signature Algorithm (ISA) (35) with default parameters and using the known target genes for each TF as the bicluster seed. |
| 14 | current genomics | No | review | Cited in review | The iterative signature algorithm (ISA) [64] was applied to decompose the compendium of each organism in co-expressed modules. Next, they compared to what extent each of these organisms shared homologous modules, i.e. a module of coexpressed genes in the reference species (yeast in their study) of which the orthologs or homologs are also coexpressed in the other species. |
| **Bioinfo Domain** | | | | | |
| 1 | bioinformatics | No | comparison | Cited in comparison | In the context of biclustering, mainly external validation has been used. Biological analyses and interpretations by human experts are most common for the evaluation of a single, newly proposed biclustering algorithm (Cheng and Church, 2000; Getz et al., 2000; Ben-Dor et al., 2002; Murali and Kasif, 2003; Bergmann et al., 2003; Getz et al., 2003; Ihmels et al., 2004); they are usually descriptive and qualitative only, and therefore are not suitable for comparing multiple methods. |
| 2 | bioinformatics | Yes (self-citation) | Usage | used to identify modules by iteratively refining random input gene sets | We proposed to identify modules by iteratively refining random input gene sets, using the signature algorithm introduced previously (Ihmels et al., 2002). |
| 3 | bmc bioinformatics | No | others | Cited in results | In our assessment of cMonkey's performance, we compared cMonkey-generated biclusters against those generated using the following algorithms: Cheng-Church (CC [25]), Order Preserving Sub-matrix (OPSM [18]), Iterative Signature (ISA [19]), xMOTIF [55], BIMAX [6], and SAMBA [86]. |
| 4 | bioinformatics | No | others | Cited in introduction | Many methods have been proposed for extracting biological meaning from microarray data, including normalization and meta-analysis (Choi et al., 2003; Griffith et al., 2005; Hu et al., 2005; Moreau et al., 2003), clustering (Allison et al., 2006; Butte et al., 2000; Cheng and Church 2000; Eisen et al., 1998; Heyer et al., 1999), signature algorithms (Bergmann et al., 2003; Ihmels et al., 2005 |
| 5 | bmc bioinformatics | No | others | Cited as a reference for the data used as benchmark | Yeast Module 9 YLR183C cell cycle yes [19, 63] ND transcription factor regulating several promoters of genes involved in pheromone response and cell cycle; |
| 6 | plos computational biology | No | others | Cited in introduction | Alternatively, biclustering methods look for gene sets that are co-expressed in a subset of the conditions [10–12]. Other methods compute the differential expression (DE) of a gene between two profiles or between two classes of profiles (e.g., cases and controls) [13,14]. |
| 7 | breifings in bioinformatics | No | comparison | Cited in comparison | Iterative signature algorithm (ISA) is a nondeterministic greedy algorithm that seeks biclusters with two symmetric requirements [24]: each column in the bicluster must have an average value above some threshold TC; likewise each row must have an average value above some threshold TR. |
| 8 | bmc bioinformatics | No | others | Cited in background | Other bilinear modeling approaches have also been used (e.g. [12,13]), but these will not be described in detail, except where they relate to the current work below. |
| 9 | bioinformatics | No | others | Cited in introduction | Numerous computational and experimental studies have used the analysis of compendia to discover consensus patterns in experiments and to uncover functional information (Bergmann et al., 2003; Eisen et al., 1998; Hughes et al., 2000). |
| 10 | bmc bioinformatics | Yes | New version | Used with extensions to bicluster time-series data | In order to mine 3D gene-condition-time datasets with different module definitions we established the EDISA (Extended Dimension Iterative Signature Algorithm), which is based on the Iterative Signature Algorithm (ISA) proposed by Bergman et al. [30] in 2003. We chose to extend the ISA algorithm because it was successfully applied to Saccharomyces cerevisiae microarray data [31], ranked among the best biclustering algorithms in a comparative study [20], |
| 11 | bioinformatics | Yes | New version | Used with extensions to bicluster yeast gene-expression data | Based on the iterative signature algorithm [Bergmann,S., Ihmels,J. and Barkai,N. (2002) Phys. Rev. E 67, 031902], we present an algorithm—the progressive iterative signature algorithm (PISA)— that, by sequentially eliminating modules, allows unsupervised identification of both large and small regulatory modules. We applied PISA to a large set of yeast gene-expression data |
| 12 | plos computational biology | No | others | Cited in introduction | Other methods performed iterative analysis as SEREND does here [18,19]. |
| 13 | bioinformatics | No | others | Cited in introduction | Bergmann et al. (2003) (who iteratively refine sets of genes and conditions until they match the definition of a special kind of bicluster that they call transcription modules). |
| 14 | bioinformatics | No | others | Cited in discussion | First, it can discover subtle associations between biological elements that are too weak to detect by considering all of their features as a whole(Bergmann et al., 2003), which also was noted in microarray data analysis (Hu and Qin, 2009). |
| 15 | algorithms for molecular biology | No | others | Cited in background | These algorithms use different strategies to compute biclusters such as exhaustive enumeration [16,19,20], iterated improvement [5,6], repeated random sampling [11], and expectation maximization [12]. |
| 16 | algorithms for molecular biology | No | others | Cited in background | Bergmann et al. [7] identified biclusters which consist of the set of co-regulated genes and the conditions that induce their co-regulation. |
| 17 | bioinformatics | No | comparison | Cited in comparison | For both scenarios the performance of the S4VD algorithm was examined in comparison to the original SSVD algorithm, the improved Plaid Model (PM; Turner et al., 2005) and the ISA (Bergmann et al., 2003). |
| 18 | bmc bioinformatics | No | others | Cited in background | The clustering method of Barkai et al. [15,16] addresses this issue of multiple membership in a different way, by using randomly-selected gene sets to iteratively search for and refine self-consistent groups. |
| 19 | bioinformatics | No | others | Cited in discussion | Although nominally ISA runtime scales linearly with number of genes and samples (Bergmann et al.,2003), it scales linearly with the number of seeds, which, to get a good modularization, is required to be larger for a bigger set of data. |
| 20 | biodata mining | No | others | Cited in background | In the ISA algorithm [12] a simple linear model for gene expression is used assuming a normally distributed expression level for each gene or condition in a specific way. |
| 21 | bmc bioinformatics | Yes | New version | Used with extensions to bicluster large expression data from Saccharomyces cerevisiae | The USA is a variant of the signature algorithm (SA) [8-,10]. We used notation similar to that described by Ihmelset al [9]. |

**LAS:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | Neuro-Oncology | Yes | Usage | Applied in biclustering a gene set matrix | We used large average submatrices (LAS), an unsupervised significance-based biclustering method, to identify groups of coordinately expressed genes.29 |
| 2 | ieee-acm transactions on computational biology and bioinformatics | No | comparison | cited in a comparison | We created another data set with 20 rows and 10 columns and only one constant column wise bicluster, depicted in Fig. 7c. The las algorithm [58] found the biclusters |
| 3 | acs infectious diseases | Yes | Usage | Applied in biclustering of uropathogenic E. coli VF genotypes from a previously described hospitalized UTI patient | To determine which antivirulence target combinations predominate in patients, we applied mathematical network community detection and statistical biclustering to uropathogenic E. coli VF genotypes from a previously described hospitalized UTI patient cohort with a high incidence of antibiotic resistance, pyelonephritis, and bacteremia.20 The mathematical tools used here21−23 |
| 4 | Cell | Yes | Usage | Applied in biclustering of Interferon Signature Genes | To visually clarify regulator/target relationships we applied the LAS biclustering algorithm (Shabalin et al., 2015) and identified a solution with five regulatory modules (C1 through C5). |
| **Bioinfo Domain** | | | | | |
| 1 | biometrics | No | others | Cited in introduction | See also Shabalin et al. (2009) for a more recent development of biclustering methods. |
| 2 | bioinformatics | Yes | New version | Used with some extensions to propose a new version | Large Average Submatrices (LAS) proposed by Shabalin et al. (2009) is yet another recent algorithm for which we utilize the localization preprocessing for further improvement. Assuming a Gaussian null model for the data, the significance score of a submatrix is defined using a Bonferroni-corrected p-value based on the size and the average of the data values in the submatrix. The main goal then is defined as that of extracting the submatrix with maximum score. As with the previous algorithms this optimization goal is computationally hard to achieve. Therefore rather than solving it to the optimum a greedy iterative heuristic is proposed for the search procedure. The heuristic starts out with a random initial submatrix. Fixing alternatively the column (row) set, the set of rows (columns) locally optimizing the significance score is searched iteratively until convergence. |
| 3 | bmc bioinformatics | No | others | Cited as previous work then compared with the proposed method | Eight different biclustering methods were applied to the combined datasets, namely Cheng and Church (CC) [15], Plaid [16], Bimax [17], Spectral [33], FLOC [34], XMOTIFS [35], large average sub-matrices (LAS) [36], bipartite spectral graph partitioning (BSGP) [37]. At all genome sizes, Spectral and XMOTIFS produced no clusters, while CC produced a single trivial cluster that encompasses the entire genome and all of the data samples. Comparison between the UNCLES method and the five biclustering methods that neither produced no clusters nor included the entire dataset in a single cluster is shown in Table 3. |
| 4 | bmc bioinformatics | No | others | Cited in related work | For example, in [28] the quality of a submatrix is defined based on the average expression value in the submatrix. Biclustering methods have been shown to be useful for finding informative patterns which are not necessarily present across the entire data set. |

**MSSRCC:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | ieee-acm transactions on computational biology and bioinformatics | No | comparison | cited in a comparison | Cho et al. [11] also used the variance to find constant biclusters together with an alternative measure to enable the discovery of more complex biclusters (see Section 3.4). |
| 2 | ieee-acm transactions on computational biology and bioinformatics | Yes (self-citation) | Usage | Applied for biclustering ofhuman cancer microarrays | In this paper, we propose specific strategies that enable MSSRCC to escape poor local minima and resolve the degeneracy problem in partitional clustering algorithms. |
| **Bioinfo Domain** | | | | | |
| 1 | bioinformatics | No | others | Cited in related work | Cho et al. (2004) proposed an algorithm, named COCLUS, also based on the mean squared residue, to find non-overlapping biclusters. |
| 2 | bmc bioinformatics | No | others | Cited in background | Cho et al. [16] have improved this mean-squaredresidue based method by using the variance as the second measure. |
| 3 | computational intelligence and neuroscience | No | others | Cited in introduction | This method has found a variety of real-world applications in the areas such as blind separation of images and nonnegative signals [1–6], spectra recovering [7–10], pattern recognition and feature extraction [11–16], dimensionality reduction, segmentation and clustering [17–32], language modeling, textmining [25, 33],music transcription [34], and neurobiology (gene separation) [35, 36]. |

**OPSM:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | | **Application** | | **Type** | | **Comments** | | **Excerpt** | |
| **Biomed Domain** | | | | | | | | | | |
| 1 | IEEE\_Transactions\_on\_computational\_Biology\_and\_Bioinformatics | | | No | | review | | Cited in review | | Ben-Dor et al. [6] defined a bicluster as an order-preserving submatrix (OPSM). According to their definition, a bicluster is a group of rows whose values induce a linear order across a subset of the columns. |
| 2 | nature reviews microbiology | | | No | | comparison | | Cited in comparison the available tools according to the strategies that they use to deal with the problem of underdetermination. | | Global versus query-driven inference. Global module inference methods22,52–59 search for the modules that explain most of the data. |
| 3 | genome research | | | No | | others | | Citation in results section | | Other variations, such as finding subsets of genes that preserve order of expression (Ben-Dor et al. 2003) or that cover many patients (Ulitsky et al. 2008; Kim et al. 2010), have been proposed |
| 4 | nucleic acids research | | | No | | comparison | | Citation in introduction and in comparison with four biclustring algorithms | | We have compared our algorithm with four existing algorithms, BIMAX (13), Iterative Signature Algorithm (ISA) (20), SAMBA (14) and HCL but did not include three earlier biclustering algorithms, Cheng–Church method (CC) (8), xMotif (24) and OPSM (12), since they were shown to have rather low performance accuracy (below 50%) in recovering implanted biclusters by previous studies (13,22). |
| 5 | bmc genomics | | | No | | comparison | | Citation in background and in comparison with five biclustring algorithms | | We compared the performance of the BBC algorithm for several different types of simulated datasets with that of the plaid model [5], the ISA [4], the method of Cheng and Church [1], the SAMBA method [6] and the OPSMs [7]. Finally, we applied the BBC algorithm to the yeast expression dataset and identified many biologically significant biclusters. |
| 6 | ieee-acm transactions on computational biology and bioinformatics | | | No | | others | | Cited in introduction | | Moreover, the inherent difficulty of this problem when dealing with the original expression matrix and the great interest in finding coherent behaviors regardless of the exact numeric values in the matrix, has led many authors to a formulation based on a discretized version of the expression matrix [3], [12], [14], [16]–[20],[23], [29], [32], [34], [36], [40]. The discretized versions remain, in general, NP-hard. |
| 7 | ieee-acm transactions on computational biology and bioinformatics | | | No | | others | | Cited in introduction | | Several clustering methods have been developed in recent years that cater to one or more of these scenarios. These include gene-shaving [B], context-specific Bayesian clustering [2], EMMIX-GENE 1741, interrelated two-way clustering [25], simultaneous clustering [17], coupled twoway clustering [7], rich probabilistic models 1221, double conjugated clusterirg [4], SAMBA 1241, order preserving submatrix clustering [3], biclustering 16l,1231a,n d the piaid model [12]. |
| 8 | ieee-acm transactions on computational biology and bioinformatics | | | No | | others | | Cited in introduction | | the discovery of such local expression patterns may be the key to uncovering many genetic pathways that are not apparent otherwise [11]. |
| 9 | ieee-acm transactions on computational biology and bioinformatics | | | No | | others | | Cited in background | | Some biclustering algorithms seek to find biclusters with coherent evolutions across the rows regardless of their exact numerical values. Ben-Dor et al. [4] looked for orderpreserving submatrices (OPSMs), in which the expression levels of all genes induce the same linear ordering of the experiments. |
| **Bioinfo Domain** | | | | | | | | | | |
| 1 | | bioinformatics | No | | others | | Cited in introduction | | Ben-Dor et al. (2002) defined a bicluster as an order preserving submatrix, or equivalenty, a group of genes whose expression levels induce some linear order across a subset of the conditions, | |
| 2 | | bioinformatics | No | | comparison | | cited in comparison | | The selected algorithms, which all are based on greedy search strategies, are Cheng and Church’s algorithm CC (Cheng and Church, 2000); Samba (Tanay et al., 2002); Order Preserving Submatrix Algorithm, OPSM (Ben-Dor et al., 2002); Iterative Signature Algorithm, ISA (Ihmels et al., 2002, 2004); xMotif (Murali and Kasif, 2003). A brief description of the corresponding approaches can be found in the Supplementary Material. | |
| 3 | | bioinformatics | Yes | | usage | | Used as a biclustering method in BicAT tool | | the Orderpreserving Submatrix Algorithm (OPSM) which tries to identify large submatrices for which the induced linear order of the columns is identical for all rows (Ben-Dor et al., 2003) | |
| 4 | | bmc bioinformatics | No | | comparison | | Cited in results in a comparison section | | In our assessment of cMonkey's performance, we compared cMonkey-generated biclusters against those generated using the following algorithms: Cheng-Church (CC [25]), Order Preserving Sub-matrix (OPSM [18]), Iterative Signature (ISA [19]), xMOTIF [55], BIMAX [6], and SAMBA [86]. We also compared our method to hierarchical clustering and k-means clustering [30] with k varying between 10 and 300 (see Methods for details). | |
| 5 | | computers in biology and medicine | No | | review | | Cited in review | | This is a major strength of bi-clustering as cellular processes are understood to rely on subsets of genes, which are co-regulated and co-expressed under certain conditions and behave independently under others, [10]. | |
| 6 | | bioinformatics | No | | others | | Cited in introduction and compared with other methods to the proposed method | | Order-Preserving Sub-Matrices (OPSM; Ben-Dor et al., 2003) searches for blocks having the same order of values in their columns. Using partial models, only the column order on subsets must be preserved. | |
| 7 | | bioinformatics | No | | others | | Cited in introduction | | Ben-Dor et al. (2002) proposed to find the order-preserving sub-matrix (OPSM) in which all genes have same linear ordering and gave a heuristic algorithm for the OPSM problem. | |
| 8 | | breifings in bioinformatics | No | | others | | Cited in comparison | | OPSM is a deterministic greedy algorithm that seeks biclusters with ordered rows [23]. The OPSM model defines a bicluster as an order-preserving submatrix, in which there exists a linear ordering of the columns in which the expression values of all rows of that submatrix linearly increase. | |
| 9 | | bioinformatics | No | | others | | Cited in introduction | | There exist other methods that search efficiently for informative submatrices in expression data (Ben-Dor et al., 2003; Getz et al., 2000; Lepre et al., 2004). | |
| 10 | | bmc bioinformatics | No | | others | | Cited in results and compared with other methods to the proposed method | | We compare our algorithm with six existing algorithms, i.e., OPSM [24], Bimax [18], Iterative Signature Algorithm, ISA [25], SAMBA [10], Cheng and Church's algorithm, CC [15] and xMotif [26], using the procedure proposed by Prelic et al. [18]. | |
| 11 | | bmc bioinformatics | No | | others | | Cited in introduction and compared with other methods to the proposed method | | Comparative studies were performed in the three aspects with several existing biclustering algorithms such as C&C, iterative signature algorithm (ISA) [32,33], order-preserving submatrix (OPSM) approach [1] and xMotifs [34], which are available in [27]. | |
| 12 | | bmc bioinformatics | No | | others | | Cited in background | | In contrast to the plaid model, which focuses on the uniformity of expression levels, biclusters defined by order-preserving submatrices focus on the relative order of the columns [31] in an attempt to identify biclusters with coherent evolutions. | |
| 13 | | bioinformatics | No | | others | | Cited in introduction | | There also exist several biclustering algorithms in this regard. These include greedy biclustering algorithms of Cheng and Church (2000) and Ben-Dor et al. (2002), iterative algorithms of Getz et al. (2000) and Ihmels et al. (2004), SAMBA of Tanay et al. (2002), Flexible Overlapped biClustering (FLOC) Kluger et al. (2003), a graph theoretic algorithm of Alexe et al. (2002). | |
| 14 | | biodata mining | No | | others | | Cited in methods section and compared with other methods to the proposed method | | Representative examples of systematic search methods include, among others, greedy algorithms [1,10-14], divide and conquer algorithms [7,15] and enumeration algorithms [16, 18]. | |
| 15 | | bioinformatics | Yes | | usage | | Used as a biclustering method in BicOverlapper tool | | For demonstration purposes, the tool has been applied to the biclustering results of the Order Preserving SubMatrix search algorithm (OPSM) (Ben-Dor et al., 2003) and Bimax (Prelic et al., 2006) in the analysis of a microarray data matrix containing two types of Diffuse Large B-Cell Lymphomas | |
| 16 | | algorithms for molecular biology | No | | others | | Cited in background | | The advantages of biclustering in the discovery of local expression patterns, described by a coherent behavior of a subset of genes in a subset of the conditions under study, have been extensively studied and documented [4-8]. | |
| 17 | | journal of theoretical biology | No | | others | | Cited in introduction section and compared with other methods to the proposed method | | We first investigate the performance of our algorithm with noisy data in comparison with other biclustering algorithms. Here we consider the following algorithms, CC [8], ISA [19, 20], OPSM [3], Bimax [30] which can be downloaded from the software toolbox BicAT. | |

**Plaid:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | | **Application** | | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | | | |
| 1 | IEEE/ACM Trans on Comp Biol and Bioinfo | | No | | review | Cited in review | The plaid model [17], defined in (25), can be used to describe most of these different biclusters structures. The restriction that every row and every column are in exactly one bicluster correspond to |
| 2 | Genome Research | | No | | others | Cited in previous work | The idea of simultaneous clustering of rows and columns of a matrix goes back to (Hartigan 1972). Recently, methods for simultaneous clustering of genes and conditions have been proposed (Cheng and Church 2000; Getz et al. 2000; Lazzeroni and Owen 2002). |
| 3 | Genome Research | | No | | others | Cited in introduction | Inadditiontotheselong-establishedalgorithms,new ones have been developed specifically for analysis of gene- expressiondata(Getzetal.2000;LazzeroniandOwen2000; Shamir and Sharan 2000; Ben-Hur et al. 2002; Sinkkonen and Kaski 2002). |
| 4 | nature reviews microbiology | | No | | review | Cited in review | Global versus query-driven inference. Global module inference methods22,52–59 search for the modules that explain most of the data. |
| 5 | genome biology | | No | | comparison | Cited in comparison of the proposed method and other biclustering and clustering methods | Comparison of ICA with Plaid model Lazzeroni and Owen [14] proposed the Plaid model for microarray analysis. The Plaid model takes the input expression data in the form of a matrix Xij (where i ranges over N samplesand j ranges over K genes). |
| 6 | bmc genomics | | No | | others | Cited in background and compared with other methods to the proposed method | We compared the performance of the BBC algorithm for several different types of simulated datasets with that of the plaid model [5], the ISA [4], the method of Cheng and Church [1], the SAMBA method [6] |
| **Bioinfo Domain** | | | | | | | |
| 1 | | bioistatistics | | No | others | Cited in introduction and discussion | There have been many new ideas proposed about the best way to cluster genes (Ben-Dor et al. (1999), Eisen et al. (1998), Heyer et al. (1998), Lazzeroni and Owen (2000), and Tamayo et al. (1999) to name a few). |
| 2 | | bioinformatics | | No | others | cited in introduction | Lazzeroni and Owen (2000) introduced the notion of a plaid model, which describes the input matrix as a linear function of variables corresponding to its biclusters. |
| 3 | | bioistatistics | | No | others | cited as an other existing matrix decompositions method | We have developed a method for finding a PMD in an efficient manner. This decomposition builds upon a variety of existing matrix decompositions, such as the SVD, the NNMF (Lee and Seung, 1999, 2001), and the plaid model (Lazzeroni and Owen, 2002). |
| 4 | | journal of computational biology | | No | others | Cited in introduction | The plaid model [12] is a statistical model for gene expression and other data. The plaid mode1 describes two-sided clusters where overlap is allowed. Like our model, their two-sided clusters are not necessarily supported on the entire set of either genes or tissues. |
| 5 | | bioinformatics | | No | others | Cited in introduction | The ‘plaid model’ (Lazzeroni and Owen, 2000) decomposes the expression matrix into a sum of submatrices, each related to specific subsets of genes and samples. |
| 6 | | bioinformatics | | No | others | Cited as previous work | A two-sided clustering (Lazzeroni and Owen, 1999; Hofmann et al., 1999) partitions the set Gene to gene clusters |
| 7 | | bioinformatics | | No | others | Cited in introduction | Other recent biclustering approaches (Lazzeroni and Owen, 2000; Cheng and Church, 2000; Segal et al., 2001; Ben-Dor et al., 2001; Tanay et al., 2002) rely on a variety of optimization procedures. |
| 8 | | biometrics | | No | review | Cited in review | Microarray data have stimulated research in such diverse areas as transposable data (Lazzeroni and Owen, 2002), multiple testing and false discovery rates (Storey, 2002), and classification tools (Dudoit, Fridlyand, and Speed, 2002). |
| 9 | | bmc bioinformatics | | No | others | Cited in background | For our modeling purposes, only methods which derive biclusters with coherent, or correlated, gene profiles, such as those of Cheng and Church [25], Yang et al. [98], and Lazzeroni and Owen [53] are suitable. |
| 10 | | biometrics | | No | others | Cited in conclusion | Some modifications of the plaid models (Lazzeroni and Owen, 2002) to include a baseline reference group of tissues (non-cancer in our case) might be useful for this purpose. |
| 11 | | computers in biology and medicine | | No | review | Cited in review | The Plaid Model [20] assumes that bi-clusters can be generated using a statistical model and tries to identify the distribution of the parameters that best the available data by minimising the sum of squared error |
| 12 | | bioinformatics | | No | others | Cited in introduction and compared with other methods to the proposed method | In the plaid model family (Lazzeroni and Owen, 2002), the i-th bicluster is extracted by row and column indicator variables ki and ij . |
| 13 | | biometrics | | No | others | Cited in introduction and compared with other methods to the proposed method | We then conclude in Section 5. Results from additional simulation studies, one real application of analyzing the food nutritional data of Lazzeroni and Owen (2002), and more analysis of the cancer data can be found in the Web Appendix. |
| 14 | | current bioinformatics | | No | review | Cited in review | The plaid model introduced by Lazzeroni and Owen [126] regards the gene expression data matrix as a linear function of multiple “layers” corresponding to its biclusters. |
| 15 | | bioinformatics | | No | others | cited in related work | In Lazzeroni and Owen (2002) they used ‘plaid models’ in the same context, where the concept of ‘layers’ (bicluster) is used to compute the values in the data matrix, which is described as a linear function of layers. |
| 16 | | bioinformatics | | No | others | Cited in introduction | Many clustering algorithms have been used in and developed for this problem, from hierarchical clustering (Eisen et al., 1998) to more intricate probabilistic models (Lazzeroni and Owen, 2002; Rogers et al., 2004; Segal et al., 2003), each of which is advantageous in certain situations. |

QUBIC:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | Nucleic Acids Research | Yes (self-citation) | Usage | Applied in clustering of gene expression data. | Our in-house bi-clustering QUBIC (Qualitative BI-Clustering) program (12) was used to identify statistically significant bi-clusters in the expression data |
| 2 | Genome Biology | No | comparison | Cited in comparison | To validate MSCM, we compared it to several multispecies and single-species methods (Table 1; Table S3 in Additional file 1). Among the single-species methods, we included the single-species version of cMonkey (SSCM; which was previously shown to be competitive with other biclustering methods [20]) as well as two recent single-species biclustering methods, QUBIC (QUalitative BIClustering algorithm) [45] and Coalesce [22] (COAL). |
| 3 | BMC Plant Biology | Yes (self-citation) | Usage | Applied in clustering of gene expression data relevant to plant cell-wall (PCW) synthesis. | We specifically employed QUBIC, a bi-clustering algorithm that we recently developed for solving this type of generalized clustering problem [21]. |
| 4 | Neural Networks | No | others | Cited in introduction | Biclustering algorithms that are used for this purpose include the interrelated twoway clustering (ITWC) algorithm (Tang & Zhang, 2005), the coupled two-way clustering (CTWC) algorithm (Getz et al., 2003), xMOTIF (Murali & Kasif, 2003), Binary State Pattern Clustering (BSPC) (Beattie & Robinson, 2006), the qualitative biclustering algorithm (QUBIC) (Li et al., 2009), the spectral biclustering algorithm (Kluger et al., 2003), the double conjugated clustering (DCC) algorithm (Busygin, Jacobsen, & Kramer, 2002), and fuzzy adaptive subspace iteration-based two-way clustering (FASIC) (Shaik & Yeasin, 2009), among others. |
| 5 | Nucleic Acids Research | No | comparison | Cited in comparison | We use the ‘FABIA’ biclustering model (36). In contrast to other biclustering methods such as BIMAX (37) and QUBIC (38), FABIA can represent homozygous regions where the same IBD segment may be present in one diploid individual two times. |
| 6 | Nucleic Acids Research | Yes (self-citation) | Usage | Applied in clustering of gene expression data. | To examine this issue, we have carried out a de novo biclustering (45,46) of the expression data of E. coli in M3D and obtained 17 condition clusters with high statistical significance (shown in Supplementary Table S6). |
| 7 | frontiers in plant science | No | review | Cited in review | For instance, Liet al.(2009) implemented a pipeline based on QUBIC, a Qualitative BIClustering algorithm, to select the conditions under which seed genes of the plant cell-wall biosynthesis pathway in Arabidopsis were found to be co-expressed among a total set of 351conditions. |
| 8 | bmc genomics | No | others | Cited in background and compared with other methods to the proposed method | The Qualitative Biclustering algorithm (QUBIC) is a recently proposed gene-wise discretization-based biclustering algorithm to solve the general form of the biclustering problem efficiently, including constant, shifting, and scaling patterns [20]. QUBIC converts a microarray data matrix into a simplified integer matrix called a representing matrix, from which it finds biclusters. |
| 9 | nucleic acids research | No | comparison | Cited in comparison | To evaluate the performance of Bi-Force on biclustering problems, we referenced to the work of Eren et al. (17). Eight (out of twelve) prevalent online available biclustering algorithms were downloaded, including Cheng and Church (7), BiMax (16), FABIA (32), ISA (33), Plaid (26), QUBIC (34), Spectral (35) and xMOTIFs (36). Five of the eight methods are integrated in the R package ‘biclust’. The three remaining software packages (FABIA, ISA and QUBIC) were downloaded from the project web sites. |
| **Bioinfo Domain** | | | | | |
| 1 | bioinformatics | No | others | Cited as previous work published benchmark data sets | Benchmark data sets published in (Prelic et al., 2006) and (Li et al., 2009) are small (50 to 100 genes), have low noise, have equally sized biclusters, and have only simultaneous row and column overlaps. |
| 2 | breifings in bioinformatics | No | comparison | Cited in comparison | QUBIC is a deterministic algorithm that reduces the biclustering problem to finding heavy subgraphs in a bipartite graph representation of the data [30]. |
| 3 | algorithms for molecular biology | No | others | Cited in backgroud and compared with other methods to the proposed method | We show that DeBi compares well with existing biclustering methods such as BIMAX, SAMBA, Cheng and Church’s algorithm (CC), Order Preserving Submatrix Algorithm (OPSM), Iterative Signature Algorithm (ISA) and Qualitative Biclustering (QUBIC) [5-7,9,10]. |
| 4 | algorithms for molecular biology | No | others | cited in background | More recently, QUBIC has been presented as a qualitative biclustering algorithm, in which the input data matrix is first represented as a matrix of integer values. |
| 5 | international journal of data mining and bioinformatics | No | others | Cited as previous work | The proposed WF-MSB method was evaluated in comparison with MSBE on a real yeast microarray data and synthetic data sets. The experimental results show that WF-MSB can effectively find the biclusters with significant GO-based functional meanings. |
| 6 | biodata mining | No | comparison | Cited in comparison | In this study, five biclustering algorithms (i.e. BIMAX, FABIA, ISA, QUBIC and SAMBA) were compared with each other in the cases where they were used to handle two expression datasets (GDS1620 and pathway) with different dimensions in Arabidopsis thaliana (A. thaliana) |
| 7 | bmc bioinformatics | No | others | Cited in backgroud | fourth approach, QUBIC [11] is a graph theoretic algorithm that identifies sets of genes with similar classconditional coexpression patterns (biclusters) by employing a network representation of the gene expression data and agglomeratively finding heavy subgraphs of co-expressed genes. |
| 8 | journal of theoretical biology | No | comparison | Cited in comparison | GSGBC and qubic shows slightly worse result. When the significant level is low, all the biclusters are functionally enriched. However, at a higher significance level, the proportion of the highly enriched biclusters decreases. |
| 9 | plos computational biology | No | others | Cited in intrioduction | Many biclustering methods have been previously described, for example, SAMBA [24], QUBIC [25], ISA [26], BIMAX [27], and NNN [28], and other algorithms [4,29,30,31,32]. |
| 10 | journal of computational biology | No | others | Cited in introduction and compared with other methods to the proposed method | Our method is also comparable with, if not better than, QUBIC (Li et al., 2009), FABIA (Hochreiter et al., 2010), ISA (Mclachian and Basford, 1998; Ihmels et al., 2004), and BCCA in identifying additive and scaling patterns. An indirect comparison suggests that CLiP is also superior to Scatter Search and the geometric algorithm (both methods do not provide publicly available code). Source code for CLiP is available upon request. |

**SAMBA:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | Nature | Yes | Usage | Appled in biclustering a gene set matrix | We next performed biclustering on the gene set matrix to identify sets of lincRNAs that are associated with distinct sets of functional categories24. This analysis revealed numerous sets of lincRNAs associated with distinct and diverse biological processes (Fig. 3a). |
| 2 | IEEE/ACM Transactions on Computational Biology and Bioinformatics | No | review | Cited in review | Tanay et al. [24] defined a bicluster as a subset of genes (rows) that jointly respond across a subset of conditions (columns). |
| 3 | Nature Genetics | No | others | Cited in introduction | Current approaches for analyzing gene expression data4-8 allow identification of groups of co-expressed genes |
| 4 | Molecular Systems Biology | No | review | Cited in review | Maximum likelihood-based scoring has been used for detecting molecular complexes in NetworkBlast (Sharan et al, 2005) and for integration of heterogeneous data through biclustering in SAMBA (Tanay et al, 2002, 2004, 2005). |
| 5 | Nature Reviews Genetics | No | review | Cited in review | Tanay and colleagues34 have shown that coexpression modules can be maintained in evolution despite changes in the regulatory genes that activate them. |
| 6 | Genome Biology | No | others | Cited in background | Biclustering was developed to address better the full com- plexity of finding co-regulated genes under multifactor con- trol by grouping genes on the basis of coherence under subsets of observed conditions [10,16-22]. |
| 7 | Nature Reviews Microbiology | No | others | Cited in introduction | By complementing gene expression with additional transcriptional information (such as motif data or DNA– protein interaction data), integrative network inference methods8,30,41–45 can extend the scope of their predic- tions beyond interactions that can be inferred from co- expression behaviour and usually result in more reliable predictions (FIG. 3) |
| 8 | Molecular Systems Biology | Yes (self-citation) | Usage | Usage of the algorithm (modified) to gene expression data biclustering. | The challenge in using such an approach lies in the fact that there are many possible submatrices, and thus many possible biclusters that may be highly redundant or not statistically significant. In this study, we adapted the SAMBA (statistical-algorith- mic method for bicluster analysis) biclustering algorithm (Tanay et al, 2004) to exhaustively search the 216 gene 21 condition matrix for all significant biclusters. |
| 9 | genome research | No | others | Cited in results | For example, biclustering techniques are commonly used to identify subsets of genes with similar expression in subsets of patients (Cheng and Church 2000; Getz et al. 2000; Tanay et al. 2002; Murali and Kasif 2003; Segal et al. 2003; Madeira and Oliveira 2004). |
| 10 | nucleic acids research | No | others | Cited in introduction and compared with other methods to the proposed method | The application results on these data sets indicate that our program outperforms the existing and popular biclustering tools, such as SAMBA (14), ISA (20), BIMAX (13), RMSBE (22) and a hierarchical clustering method (HCL) in both the identification accuracy and the computational efficiency. |
| 11 | nature protocols | Yes (self-citation) | usage | Used as a biclustering method in Expander tool | A biclustering algorithm can detect a collection of biclusters in a large gene expression data set. In this collection, genes or conditions can take part in more than one bicluster. Expander implements the Statistical-Algorithmic Method for Bicluster Analysis (SAMBA 2.0) biclustering algorithm, which can handle data sets with hundreds to thousands of conditions5,6,35. |
| 12 | nature chemical biology | No | review | Cited in review | Early methods for clustering genes assumed that genes cluster across all observed cell states (or genetic backgrounds) and that genes participate in only one cluster. Newer methods allow for genes to participate in multiple clusters and for those clustersto be condition-specific7,40–47. |
| 13 | bmc genomics | No | others | Cited in background and compared with other methods to the proposed method | Tanay et al. [6] proposed a SAMBA biclustering scheme using bipartite graphs containing both conditions and genes. |
| 14 | IEEE\_Transactions\_on\_computational\_Biology\_and\_Bioinformatics | No | review | Cited in review | SAMBA can be executed using two statistical models of the resulting bipartite graph. When the simpler model is used, Tanay et al. show how to compute an upper-bound on the probability of an observed bicluster. |
| **Bioinfo Domain** | | | | | |
| 1 | bioinformatics | No | comparison | Cited in comparison | Tanay et al. (2002) and Yang et al. (2002, 2003), provide a comparison with the algorithm by Cheng and Church, (2000) regarding synthetic data respectively the problem formulation introduced in Cheng and Church (2000). In Ihmels et al. (2004), |
| 2 | bioinformatics | No | others | Cited in introduction | Several methods have been put forward to address these issues (Cheng and Church, 2000; Gasch and Eisen, 2002; Getz et al., 2000; Hastie et al., 2000; Tanay et al., 2002). |
| 3 | bmc bioinformatics | Yes (self-citation) | usage | Used as a biclustering method in Expander tool | EXPANDER serves as the major platform in which we integrate various gene expression analysis algorithms that were developed in our lab, including CLICK for clustering [9], SAMBA for biclustering [10], PRIMA for promoter elements analysis [7] |
| 4 | bioinformatics | No | others | Cited in introduction | Some of the existing biclustering algorithms are based on the idea to perform standard clustering algorithms iteratively on both genes and conditions (Getz et al., 2000; Busygin et al., 2002). Other recent biclustering approaches (Lazzeroni and Owen, 2000; Cheng and Church, 2000; Segal et al., 2001; Ben-Dor et al., 2001; Tanay et al., 2002) rely on a variety of optimization procedures. |
| 5 | bmc bioinformatics | No | others | Cited in background and compared with other methods to the proposed method | In our assessment of cMonkey's performance, we compared cMonkey-generated biclusters against those generated using the following algorithms: Cheng-Church (CC [25]), Order Preserving Sub-matrix (OPSM [18]), Iterative Signature (ISA [19]), xMOTIF [55], BIMAX [6], and SAMBA [86]. |
| 6 | journal of computational biology | Yes (self-citation) | Usage | Used for identification of models of yeast and bacterial protein interaction data | Similarly to the probabilistic approach taken in [20], we define the two models as follows: The protein-complex model, Mc, assumes that every two proteins in a complex interact with some high probability b. |
| 7 | computers in biology and medicine | No | review | Cited in review | The Statistical Algorithmic Method for Bi-cluster Analysis (SAMBA) method nds bi-clusters based on the coherent evolution model (Section 3.1) [21]. |
| 8 | bioinformatics | No | others | cited in introduction | For example, some approaches obtain faster running times by limiting the types of bi-clusters they can identify (Tanay et al., 2002), or by focusing on specific types of data, such as time courses (Madeira and Oliveira, 2005). |
| 9 | bmc bioinformatics | No | others | Cited in background | In the last few years several methods have been proposed to avoid these drawbacks [15-18]. |
| 10 | breifings in bioinformatics | No | review | Cited in review | The clusters contain similar objects with respect to a metric, representing genes of similar functionality, or tissues of similar profiles, or patients of similar clinical outcomes [32–37]. |
| 11 | bioinformatics | No | others | Cited in introduction and compared with other methods to the proposed method | Probabilistic and generative methods use model-based techniques to define biclusters. Statistical-Algorithmic Method for Bicluster Analysis (SAMBA; Tanay et al., 2002) uses a bipartitioned graph, where both conditions and genes are nodes. |
| 12 | bioinformatics | No | others | Cited in introduction and compared with other methods to the proposed method | We compare different bi-clustering methods on constant biclusters. Here we consider CC (Cheng and Church, 2000), Samba (Tanay et al., 2002), ISA (Ihmels et al., 2002, 2004) and Bimax (Prelic´ et al., 2006). |
| 13 | current bioinformatics | No | review | Cited in review | Tanay et al. [127] introduced SAMBA (Statistical Algorithmic Method for Bicluster Analysis) method which combines graph theory with statistical data modeling. |
| 14 | bioinformatics | No | others | Cited in related work | Also in Tanay et al., (2002) the authors identified biclusters by means of a bipartite graph-based model and using a greedy approach to add/remove vertices in order to find maximum weight subgraphs, which is related to its statistical significance. |
| 15 | bmc bioinformatics | No | others | Cited in background and compared with other methods to the proposed method | It should be pointed out that some symbolic, coherent evolution or numerical biclusters, such as those producedn by cMonkey [9], SAMBA [10] and some statistical criteria, cannot be classified as additive or multiplicative patterns directly. |

**Spectral:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | IEEE/ACM Trans on Comp Biol and Bioinfo | No | review | Cited in Review | Kluger et al. [16] used a spectral approach to biclustering by assuming that the data matrix contains a checkerboard structure after normalization. |
| 2 | Genome Biology | No | others | Cited as Background | Biclustering was developed to address better the full com- plexity of finding co-regulated genes under multifactor con- trol by grouping genes on the basis of coherence under subsets of observed conditions [10,16-22]. |
| 3 | Nature Reviews Microbiology | No | review | Cited in Review | Global module inference methods22,52–59 search for the modules that explain most of the data. |
| 4 | ieee transactions on nanobioscience | Yes | Usage | Used as a biclustering method for the determination of gene selection | The proposed gene selection method is based on the spectral biclustering algorithm proposed by Kluger [14]. |
| 5 | nucleic acids research | No | others | Cited as previous work | Actually, there have been several studies that used biclustering as part of a larger analysis pipeline to do cancer subtyping (34). |
| 6 | nature chemical biology | No | review | Cited in Review | Newer methods allow for genes to participate in multiple clusters and for those clusters to be condition-specific7,40–47. |
| 7 | journal of leukocyte biology | Yes (self-citation) | Usage | Used in biclustering of gene expression data | The latter is obtained by a preprocessing, normalization step, as described recently [37]. We selected a normalization procedure based on the concept that two genes and likewise, two samples, whose expression profiles differ only by a multiplicative constant of proportionality, are really behaving in the same way. |
| 8 | cancer research | Yes | Usage | used in biclustering of gene expression data for 16 SCC, 19 dysplasia, and 4 normal mucosa | The gene expression data for 16 SCC, 19 dysplasia, and 4 normal mucosa cultures were analyzed by ‘‘spectral clustering,’’ which is a useful and sensitive tool for analyzing complex microarray data. |
| 9 | laboratory investigation | Yes (self-citation) | Usage | Used in biclustering of gene expression microarray data | Separation of these samples was also seen using our spectral bidirectional clustering method.31 |
| 10 | ieee-acm transactions on computational biology and bioinformatics | No | others | Cited in related work | Kluger et al. [28] apply a spectral coclustering algorithm, motivated by Dhillon’s bipartite formulation [29], on gene expression data to produce a “checkerboard” structure. |
| 11 | journal of clinical oncology | No | others | Cited in introduction | 14-19 Other popular clustering methods, hierarchical clustering, and principle components analysis are more reductionist. |
| **Bioinfo Domain** | | | | | |
| 1 | bioinformatics | No | others | Cited in related work | Most biclustering papers are concerned with the introduction and validation of a new approach, while only a few contain quantitative comparisons with existing methods. Cheng and Church (2000) and Kluger et al. (2003), |
| 2 | bmc bioinformatics | No | others | Cited in background | So-called biclustering (clustering both genes and experimental conditions), is a widely studied problem and many different approaches to it have been published [6,25,52,76,80,86,98]. |
| 3 | computers in biology and medicine | No | review | Cited in Review | These authors used spectral clustering principals [36] to create a partition matrix, Z, of the original GE matrix. Z is then clustered using a possibilistic approach. |
| 4 | bmc bioinformatics | No | others | cited in discussion | Kluger et al. [33] suggested the use of singular value decomposition to find checkerboard patterns in gene expression matrices and more recently Dueck et al. [34] proposed a probabilistic sparse factorization method to discover gene expression biclusters. |
| 5 | breifings in bioinformatics | No | review | Cited in Review | Biclustering of gene expression data allows simultaneous clustering of the rows and columns of a matrix, where rows correspond to genes and columns to conditions [33, 124–127]. |
| 6 | bioinformatics | No | others | Cited as previous work | This is a standard transformation for microarray analysis. The motivation of the logarithm transformation is that the multiplicative model becomes the additive model. Another reason is that after the transformation, the distribution properties will be better, i.e. the distributions are closer to normal distributions. (Causton et al., 2003; Kluger et al., 2003). |
| 7 | bioinformatics | No | others | Cited in introduction and compared with other methods to the proposed method | Using partial models, only the column order on subsets must be preserved. Spectral clustering (SPEC; Kluger et al., 2003) performs a singular value decomposition of the data matrix after normalization. |
| 8 | breifings in bioinformatics | No | comparison | cited in comparison | Spectral uses singular value decomposition to find a checkerboard pattern in the data in which each bicluster is up- or downregulated [25]. Only biclusters with variance lower than a given threshold are returned. |
| 9 | bmc bioinformatics | No | others | Cited as previous work | Kluger et al. [19] have studied the checkerboard structure of this type of biclusters using a normalization scheme based on the above equation. |
| 10 | bmc bioinformatics | No | others | cited in discussion | In its current form, the AutoSOME method does not identify genes whose co-regulation is restricted to a subset of arrays (see e.g., [37-39]). |
| 11 | bmc bioinformatics | No | comparison | cited in comparison | Other methods for biclustering, such as plaid [26] and spectra models [29], are related to projection methods which regenerate the data matrix by biclusters. The plaid model expresses the value of each element in the gene expression data as a series of additive layers [26] and the spectra model uses singular value decomposition to identify eigenvectors that reveal the existence of checkerboard structures within the rearranged genes and conditions [29]. |
| 12 | bioinformatics | No | others | cited in introduction | There also exist several biclustering algorithms in this regard. These include greedy biclustering algorithms of Cheng and Church (2000) and Ben-Dor et al. (2002), iterative algorithms of Getz et al. (2000) and Ihmels et al. (2004), SAMBA of Tanay et al. (2002), Flexible Overlapped biClustering (FLOC) Kluger et al. (2003), a graph theoretic algorithm of Alexe et al. (2002). |
| 13 | bmc bioinformatics | No | others | Cited in background | Further publications followed [8,14-18]. These biclustering methods aim at finding subsets of genes and conditions by clustering them simultaneously. |
| 14 | biometrics | No | others | cited as previous work | Several other biclustering methods are also based on SVD, such as the coclustering algorithm of Dhillon, Mallela, and Modha (2003); the spectral method of Kluger et al. (2003) |

**xMOTIFs:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | IEEE/ACM Trans on Comp Biol and Bioinfo | No\* | review | Cited broadly in review | Murali and Kasif [19] aimed at finding conserved gene expression motifs (xMOTIFs). They defined an xMOTIF as a subset of genes (rows) that is simultaneously conserved across a subset of the conditions (columns). |
| 2 | Nature reviews microbiology | No | review | Cited in review | Global module inference methods22,52–59 search for the modules that explain most of the data. |
| 3 | Genome research | No | others | Cited in context | The problem of extracting subsets of genes with particular prop- erties has also been studied in the context of gene expression data. For example, biclustering techniques are commonly used to identify subsets of genes with similar expression in subsets of pa- tients (Cheng and Church 2000; Getz et al. 2000; Tanay et al. 2002; Murali and Kasif 2003; Segal et al. 2003; Madeira and Oliveira 2004). |
| 4 | Nucleic Acids Research | No | comparison | Cited in comparison | We have compared our algorithm with four existing algorithms, BIMAX (13), Iterative Signature Algorithm (ISA) (20), SAMBA (14) and HCL but did not include three earlier biclustering algorithms, Cheng–Church method (CC) (8), xMotif (24) and OPSM (12), since they were shown to have rather low performance accuracy (below 50%) in recovering implanted biclusters by previ- ous studies (13,22). |
| 5 | BMC Genomics | No | others | Cited as background | Murali and Kasif [8] discretized gene expression data into several symbols and searched for conservative symbol motifs (xMOTIFs). |
| 6 | Annual Review of Genomics and Human Genetics | No | review | Cited in review | Recently, several “bi-clustering” algorithms were developed that attempt to group genes together within the context of a subset of experimental conditions (20, 36, 73). |
| 7 | IEEE/ACM Trans on Comp Biol and Bioinfo | No | others | Cited as background | Most discretization techniques commonly applied to gene expression data use absolute expression values based on the following concepts: average and standard deviation [14], [33], [37], percentage of values [2], [32], equal-width intervals [2], [32], equal frequency [35], linear order between the conditions [3], [4], [16], [17], [18], [19], and statistically significant states [30], [41]. |
| 8 | ieee-acm transactions on computational biology and bioinformatics | No | others | cited as previous work | Murali and Kasif [21] proposed a representation for gene expression data called conserved gene expression motifs (xMOTIFs). An xMOTIF is a subset of genes that is simultaneously conserved across a subset of experimental conditions. |
| 9 | genome biology | No | others | Cited as background | Bicluster strategies are well suited to map both the condition dependency and the modularity of the transcriptional network from microarray compendia [8-11], but do not give any information on the transcriptional program of the modules. |
| 10 | nucleic acids research | No (self-citation) | others | Cited as introduction | Heuristic biclustering algorithms have been reported (7), such as Cheng and Church (8), coupled two-way clustering (9), plaid model (10), SPLASH (11), SAMBA (12), xMotif (13) and Gibbs sampling (14). |
| **Bioinfo Domain** | | | | | |
| 1 | bioinformatics | No | comparison | Cited in comparison | In the absence of noise, ISA, Samba and Bimax are able to identify a high percentage (>90%) of implanted transcription modules; as expected, the same holds for the hierarchical clusteringn approach, if the number k of clusters to be generated corresponds to the actual number of implanted modules. In contrast, the scores obtained by CC and xMotif are substantially lower. |
| 2 | bioinformatics | Yes | usage | Used as a biclustering method in BicAT tool | the xMotif algorithm, an iterative search method which seeks biclusters with quasi-constant expression values (Murali and Kasif, 2003); |
| 3 | bmc bioinformatics | No | others | Cited in backgroud and compared with other methods to the proposed method | The different bi/clustering algorithms used for the comparative analysis included: Cheng Church [25], Order Preserving Submatrix (OPSM [18]), Iterative Signature (ISA [19]), xMOTIF [55], and BIMAX [6] (all of these algorithms were run using the BICAT implementation [17]) |
| 4 | bioinformatics | No | others | Cited in introduction and compared with other methods to the proposed method | In Murali and Kasif (2003), the best two bi-clusters generated by the software XMOTIF are B1 and B2 (Table 2). B1 contains 11 genes and 15 samples. Among the 15 samples, 14 of them are tumor samples and 1 of them is a normal sample. B2 contains 13 genes and 18 samples. |
| 5 | bioinformatics | No | others | Cited in introduction and compared with other methods to the proposed method | To simplify this task, some methods discretize the data in a first step, like xMOTIF (Murali and Kasif, 2003) or Bimax (Prelic et al., 2006) which even binarizes the data and searches for blocks with an enrichment of ones. |
| 6 | bioinformatics | No | others | Cited in introduction | Although slightly different, the xMOTIF (conserved gene expression motif) is also a special case of bicluster. A probabilistic algorithm to find xMOTIFs is presented in Murali and Kasif (2003). Other heuristics, such as simulated annealing, have been also tested together with the mean squared residue (Bryan et al. 2005). |
| 7 | breifings in bioinformatics | No | comparison | Cited in comparison | xMOTIFs is a nondeterministic greedy algorithm that seeks biclusters with conserved rows in discretized dataset [26]. |
| 8 | bmc bioinformatics | No | others | Cited in backgroud and compared with other methods to the proposed method | We compare our algorithm with six existing algorithms, i.e., OPSM [24], Bimax [18], Iterative Signature Algorithm, ISA [25], SAMBA [10], Cheng and Church's algorithm, CC [15] and xMotif [26], using the procedure proposed by Prelic et al. [18]. |
| 9 | bmc bioinformatics | No | others | Cited in backgroud and compared with other methods to the proposed method | Experiments have been conducted on the yeast cell cycle dataset using the proposed algorithms and Cheng and Church (C&C) algorithm [12], iterative signature algorithm (ISA) [32,33], order-preserving submatrix (OPSM) approach [1] and xMotifs [34]. Post-filtering was applied to the biclustering results in order to eliminate insignificant biclusters as well as impose common constraints for comparison. |
| 10 | bioinformatics | No | others | Cited in introduction | There also exist several biclustering algorithms in this regard. These include greedy biclustering algorithms of Cheng and Church (2000) and Ben-Dor et al. (2002), iterative algorithms of Getz et al. (2000) and Ihmels et al. (2004), SAMBA of Tanay et al. (2002), Flexible Overlapped biClustering (FLOC) Kluger et al. (2003), a graph theoretic algorithm of Alexe et al. (2002). Murali and Kasif (2003) have defined biclusters as conserved gene expression motif i.e. xMOTIF, and devised an algorithm to find largest xMOTIF. |
| 11 | bmc bioinformatics | No | others | Cited in background | Such datasets are often processed by full-space clustering approaches, such as k-means clustering [10], hierarchical clustering [11], and spectral clustering [12]. In 2000, the biclustering approach was introduced by Cheng et al. [13]. Further publications followed [8,14-18]. |
| 12 | algorithms for molecular biology | No | others | Cited in background | Furthermore, the inherent difficulty of this problem when dealing with the original real-valued expression matrix and the great interest in finding coherent behaviors regardless of the exact numeric values in the matrix, has led many authors to a formulation based on a discretized version of the expression matrix [7-9,12-23]. |
| 13 | journal of theoretical biology | No | others | Cited in introduction | interrelated two-way clustering (ITWC) [35], double conjugated clustering (DCC) [6], coupled two-way clustering (CTWC) [14], statisticalalgorithmic method for bicluster analysis algorithm (SAMBA) [33], iterative signature algorithm (ISA) [19, 20], xMOTIF [28, 39], plaid model [24, 37], a fast divide-andconquer algorithm (Bimax) [30], |
| 14 | bioinformatics | No | others | Cited in introduction | Many gene expression analysis methods (Brazma and Vilo, 2000; Friedman et al., 2000; Murali and Kasif, 2003; Sherlock, 2000) as well as combination data methods are model-based and require assumptions to be made on the data. Such methods are generally computationally intensive since parameters are estimated in an iterative fashion. |
| 15 | algorithms for molecular biology | Yes (self-citation) | usage | Used for biclustering of leukaemia microarray data | First, we compute layouts for biclusters extracted from leukaemia microarray data by the xMotif biclustering algorithm [11,21]. |
| 16 | algorithms for molecular biology | No | others | Cited in background | Murali and Kasif [8] found subsets of genes that are simultaneously similarly expressed across a subset of the samples. The algorithm uses prior knowledge about the sample phenotypes. |
| 17 | bioinformatics | No | others | Cited in introduction | They include, among others, Block Clustering by Hartigan (1972), δ-biclusters by Cheng and Church (2000), Coupled Two-Way Clustering (CTWC) by Getz et al. (2000), FLOC by Yang et al. (2002, 2003), δ-pClusters by Wang et al. (2002), Spectral biclustering by Kluger et al. (2008), Iterative Signature Algorithm (ISA) algorithm by Ihmels et al. (2002, 2004), Interrelated Two-Way Clustering (ITWC) algorithm by Tang et al. (2001), Plaid model by Lazzeroni and Owen (2002), Order Preserving Sub Matrix (OPSM) algorithm by Ben-Dor et al. (2002), SAMBA by Tanay et al. (2002) and xMOTIF by Murali and Kasif (2003). |
| 18 | bioinformatics | No | others | Cited as a comparative method with other ones for the proposed method | The performance of the Algorithm 1 has been compared with biclustering methods such as CC [9], OPSM [14], ISA [40], BiMax [17], xMotifs [41] and Samba [13] for the GaschYeast data set and CCC-Biclustering [30] for Yeast data set. Also, random biclusters have been generated as naive reference method. |