**Clustering Algorithms :**

**Cast :**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | Nature | Yes | Usage | Applied in clustering of Gene expression data. | A non-hierarchical clustering algorithm14 was used to de®ne experimental clusters. This approach takes a graph theoretic approach, and makes no assumptions on the similarity function or the number of clusters sought. |
| 2 | Nature Genetics | Yes | Usage | Applied in clustering of DNA Gene expression data. | A sequence similarity score was computed between all pairs of statistically significant dimers that were then grouped using the ‘cast’ routine24. |
| 3 | Nature Genetics | No | review | Cited in review | A variation that allows samples to influence the location of neighboring clusters is known as the self-organizing map or Kohonen map41,42. Such maps are particularly valuable for describing the relationships between clusters. New centroid methods specifically for microarray data have also been proposed43. |
| 4 | human molecular genetics | No | review | Cited in review | For instance, the corrupted clique-graph model and the cluster affinity search technique (CAST) proposed by Ben-Dor and Yakhini (94) were succefully tested on two large multiple expriment data sets |
| 5 | genome research | No | others | cited as previous work | The third category groups genes into blocks, in which the correlation is maximized and between which the correlation is minimized (Ben-Dor et al. 1999). |
| 6 | biotechniques | No | review | Cited in review | A large number of distance measures and analytic techniquesnhave been implemented, including: • Hierarchical clustering (HCL) (26) • Bootstrapped/jackknifed HCL • k-means/k-medians clustering (KMC) (27) • k-means/k-medians support (iterative KMC) • Self-organizing maps (SOMs) (28) • Self-organizing trees (SOTA) (29) • Cluster affinity search technique (CAST) (30) |
| 7 | genome biology | No | others | Cited in background | Clustering approaches group genes and experiments with similar behavior [6-10], making the data simpler to analyze [11]. Clustering methods group genes that behave similarly under similar experimental conditions, assuming that they are are functionally related. |
| 8 | toxicological sciences | No | others | cited as previous work | Others have used similar types of visual representations; however in those cases the order of items on the x and y axes were determinated by the application of a clustering algorithms (Ben-Dor et al., 1999; Weinsten et al., 1997), |
| 9 | emerging infectious diseases | No | review | Cited in review | Most global gene expression analyses have used some form of unsupervised clustering algorithm (16,42-44) to find genes coregulated across the dataset (Figure 1). |
| 10 | genome research | No | others | Cited in discussion | Interesting algorithms based on threshold graphs are shown by Hartuv et al. (1999) and Ben-Dor and Yakhini (1998). The latter approach is enriched by heuristics that allow corrections of false joining of two data points. Another interesting approach to clustering are self-organizing maps. |
| 11 | biotechniques | No | review | Cited in review | A number of different methods: k-means, self-organizing maps, hierarchical clustering, vector machines and Bayesian statistics are employed for clustering analysis (2,7,12,25,34,67,68,75,77). |
| 12 | genome biology | No | comparison | Cited in comparison | We studied the performance of a wide variety of clustering algorithms, including several agglomerative hierarchical algorithms (average linkage, centroid linkage, complete linkage and single linkage), a divisive hierarchical algorithm called DIANA [20], k-means [7], a graph-theoretic algorithm called CAST [21], a finite Gaussian mixture model-based hierarchical clustering algorithm from MCLUST [8], and an IMM-based approach [18]. |
| **BIOINFO DOMAIN** | | | | | |
| 1 | journal of computational biology | No | others | cited in introduction and compared with the proposed method | Our approach is quite different from the clustering approach used by Alon et al. (1999), Ben-Dor et al. (1999), Eisen et al. (1999), Michaels et al. (1998), and Spellman et al. (1998) in that it attempts to learn a much richer structure from the data. |
| 2 | journal of computational biology | No | review | Cited in review | A variety of clustering algorithms have been used to group together similar temporal expression patterns and thus reveal clusters of genes that seem to be co-regulated in experiments [13, 25, 36, 141, 194, 235]. |
| 3 | bioinformatics | No | review | Cited in review | Ben-Dor et al. (1999) have developed a clustering algorithm based on random graph theory. |
| 4 | bioinformatics | No | comparison | Cited in comparison | We used three clustering algorithms ij our empirical study: the Cluster Affinity Search Technique (CAST) (Ben-Dor and Yakhini, 1999). |
| 5 | bioinformatics | No | others | cited in introduction and compared with the proposed method | We compare the performance of the model-based approach to CAST (Ben-Dor and Yakhini, 1999), a leading heuristic-based clustering algorithm. |
| 6 | journal of computational biology | Yes (self-citation) | Usage | Used for clustering of gene expression profiles for Tissue classification | The rest of the paper is organized as follows. In Section 2, we describe the principal classification methods we use in this study. These include two state of the art methods from machine learning, and a novel approach based on the clustering algorithm of (Ben-Dor et al. 1999). |
| 7 | biostatistics | No | others | cited in introduction and discussion sections | Several groups have studied different clustering algorithms ( Ben-Dor et al., 1999; Eisen et al., 1998; Heyer et al., 1999; Lazzeroni and Owen, 2000; Tamayo et al., 1999). |
| 8 | bioinformatics | No | comparison | Cited in comparison | In Figure 1, three clustering algorithms – hierarchical single-link, kmeans (with random initialization), and CAST (Ben-Dor and Yakhini, 1999) – are compared on a simulated data set. |
| 9 | bioinformatics | No | others | Cited in introduction and compared with the proposed method | A widely used technique for microarray data analysis is clustering analysis (Alon et al., 1999; Ben-Dor et al., 1999; Bittner et al., 1999; Getz et al., 2000 a,b; Hartuv et al., 2000). |
| 10 | journal of computational biology | No (self-citation) | others | Cited in introduction | These methods are successfully applied to yeast cell cycle data in [ 141. In [4] and [13] more direct approaches to clustering are taken, using graph theoretic methods. |
| 11 | bioinformatics | No | others | Cited in introduction | Other different clustering methods have recently been proposed (Heyer et al., 1999; Ben-Dor et al., 1999), but their performance remains to be evaluated by the user community. |
| 12 | bioinformatics | No | others | Cited in introduction | The analysis of gene expression microarray data using clustering techniques has an important role to play in the discovery, validation and undestanding of various classes and subclasses of cancer, see, for example, Eisen et al. (1998), Ben-Dor et al. (1999, 2000). |
| 13 | bioinformatics | No | review | Cited in review | It is thus not surprising that early work on gene expression analysis has focused on this level, and several clustering algorithms have been suggested for gene expression data (Ben-Dor et al., 1999; Tamayo et al., 1999; Sharan and Shamir, 2000). |
| 14 | bioinformatics | No | others | Cited in introduction | Ben-Dor et al. (1999) proposed an algorithm to recover clusters from noisy data. |
| 15 | bioinformatics | No | others | Cited in introduction | Popular non-model-based clustering methods include hierarchical clustering (Spellman et al., 1998), the graph-based CAST nalgorithm (Ben-Dor and Yakhini, 1999), self-organizing maps (Tamayo et al., 1999). |
| 16 | bioinformatics | No | others | Cited in introduction | Several algorithmic techniques were previously used in clustering gene expression data, including hierarchical clustering (Eisen et al., 1998), self organizing maps (Tamayo et al., 1999), K-means (Herwig et al., 1999), simulated annealing (Alon et al., 1999), and graph theoretic approaches: HCS (Hartuv & Shamir, 2000) and CAST (Ben-Dor et al., 1999). |
| 17 | bioinformatics | No | others | cited as previous work | To address this issue and to further accelerate the biclustering process, we generalize the model of bicluster to incorporate null values and propose a probabilistic algorithm (FLOC) that can discover a set of possibly overlapping biclusters simultaneously. |
| 18 | journal of computational biology | No | review | Cited in review | For instance, much of the large-scale analysis of genomic data to date focuses on gene expression patterns, and particularly on establishing gene clusters based on their expression (Ben-Dor et al., 1999; Sharan and Shamir, 2000; Spellman et al., 1998; Tamayo et al., 1999). |
| 19 | bioinformatics | No | others | cited in introduction | Hierarchical clustering (Eisen et al., 1998), self-organizing mapping (Tamayo et al., 1999) and other clustering methods (Ben-Dor et al., 1999) have been applied to the expression profile data. |
| 20 | bioinformatics | Yes | Usage | used for clustering of samples of gene expression data | In addition, we used the CAST algorithm (Ben-Dor et al., 1999) for clustering the samples. |
| 21 | bioinformatics | Yes | Usage | used for clustering of of ideal regulator profiles that are highly correlated | To add a new regulator, we apply the CLUST algorithm (Ben-Dor et al., 1999) to find clusters of ideal regulator profiles that are highly correlated (above 0.8), and may correspond to a new regulator of the genes for which these ideal profiles were generated. |
| 22 | bioinformatics | No | others | cited in introduction | Cluster Affinity Search Technique (Ben-Dor et al., 1999) has been implemented and succefully been used to analyze or cluster high-dimentional microarray data. |

**Eisen:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | Nature | Yes (self-citation) | Usage | Applied in clustering of gene expression data. | A hierarchical clustering method was used to group genes on the basis of similarity in the pattern with which their expression varied over all samples5. The same clustering method was used to group the experimental samples (cell lines and tissues separately) on the basis of similarity in their patterns of expression.We focus ®rst on a set of 1,753 genes (about 22% of the 8,102 genes analysed), whose transcripts varied in abundance by at least fourfold from their median abundance in this sample set in at least three of the samples (Fig. 1; see Supplementary Information Fig. 4 for the complete cluster diagram). |
| 2 | Nature Genetics | Yes (self-citation) | Usage | Applied in clustering of expression microarray experiments with Gene Ontology terms. | Fig. 2. Correspondence between hierarchical clustering of expression microarray experiments with GO terms. The coloured matrix represents the results of clustering many microarray expression experiments32. In the matrix, each row represents the yeast gene described to the right, and each column represents the expression of that gene in a particular microarray hybridization. |
| 3 | Nature | Yes (self-citation) | Usage | Applied in clustering of gene expression data. | In all, ,1.8-million measurements of gene expression were made in 96 normal and malignant lymphocyte samples using 128 Lymphochip microarrays. Figure 1 provides an overview of the variation in gene expression across these samples. A hierarchical clustering algorithm was used to group genes on the basis of similarity in the pattern with which their expression varied over all samples20. |
| 4 | Nature Protocols | Yes | Usage | Appled in clustering of real-time PCR expression data. | 16| Present the data as DCT and plot the heatmap as described18. An example is shown in Figure 3. |
| 5 | Genome Research | No | others | Cited in introduction | Similarly, for gene expression profiles and other molecular states, numerous programs such as GeneCluster (Tamayo et al. 1999), Tree-View (Eisen et al. 1998), and GeneSpring (www.silicongenetics.com) are available for clustering, classification, and visualization. |
| 6 | Nature | No | review | Cited in review | Although it is essentially impossible to draw meaningful conclusions from a single quantitative gene expression profile, the availability of multiple profiles from related samples allows the application of statistical tools80 to extract signature patterns containing diagnostic or functional information. |
| 7 | Cell | Yes | Usage | Applied in clustering of Stem cells. | The heat map in Figure 4C was produced using the SAGE tag counts in Table 1, which were visualized using MapleTree (developed by L. Simirenko). Before visualization, tag counts were log2-transformed (treating a count of 0 as if it were 1), median-centered by tag, and subjected to hierarchical complete linkage clustering by tags and SAGE libraries with uncentered correlation similarity metrics using Cluster (Eisen et al., 1998). |
| 8 | Molecular Biology of the Cell | Yes | Usage | Applied in clustering of Yeast genes. | In our initial experiments, 142 different mRNA samples were analyzed by whole-genome microarray hybridization. Each microarray used in this study contained ; 6,200 known or predicted yeast genes that had been identified at the time of our analysis (Ball et al., 2000). The resulting table of ; 9 3 105 quantitative measurements of transcript levels was organized by hierarchical clustering and displayed as previously described (Eisen et al., 1998) (Figure 1). |
| 9 | Nature Genetics | Yes | Usage | Applied in clustering of gene expression data. | These expression data were normalized to a mean of 0 and a variance of 1. Data were hierarchically clustered and visualized using the Cluster and TreeView software packages32. We parsed these 52 genes into 32 coregulated probe sets and 20 probe sets that are not coregulated, based on the dendrogram in Figure 3. |
| 10 | Science | No | review | Cited in review | Substantial research has been done on expression profiling, in which clustering analysis is used to identify genes that are coexpressed with genes of known function (5, 6). Although clustering analysis provides insight into the “correlation” among genes and biological phenomena, it does not reveal the “causality” of regulatory relationships. |
| 11 | Science | Yes | Usage | applied in clustering of topological overlap matrix corresponding to the small network | The rows and columns of the matrix were reordered by the application of an average linkage clustering method (22) to its elements, allowing us to identify and place close to each other those nodes that have high topological overlap. |
| 12 | Science | Yes | Usage | applied in clustering of gene expression data | Genes were hierarchically clustered across all experiments using the program Cluster22 and viewed in TreeView (http://rana.lbl.gov/ EisenSoftware.htm). |
| 13 | nature genetics | Yes (self-citation) | Usage | applied in clustering of microarray expression data | An early example of such a functional map (Fig. 2) was made using a hierarchical clustering strategy to group genes on the basis of similarity in their patterns of expression23 |
| 14 | genome research | No | others | Cited in introduction | Examples of this include the application of tree maps to show distribution of disk usage on a file system (Johnson and Shneiderman 1991) and hierarchical biological data (McConnell et al. 2002), directed graphs to depict networks, pathways, and phylogenetic information (Ciccarelli et al. 2006; Darwin 1859; Letunic and Bork 2007), and clustered heat maps to visualize array and expression data (Eisen et al. 1998; Sneath 1957). |
| 15 | Cell | Yes | Usage | Applied in clustering of Yeast genes. | Two-dimensional hierarchical clustering of 59 control experiments and 278 genes, selected to include only experiments with two or more genes up- or downregulated at a significance of P # 0.15; and only genes that are significant at P # 0.15 in two or more experiments, using a significance model which accounts for measurement error. |
| 16 | nature genetics | No | others | cited as previous work | Hierarchical clustering has also been used to organize genes into hierarchical dendograms on the basis of their expression across multiple growth conditions7. |
| 17 | nature | Yes | Usage | Applied in clustering of lincRNAs microarray data | Total RNA from cells was extracted using TRIzol and the RNeasy mini kit (Qiagen) and hybridized to Stanford human oligonucleotide (HEEBO) arrays as described29. Data analysis was done using CLUSTER28. |
| 18 | nature medicine | Yes | Usage | Applied in clustering of gene expression data | The specificity was 100% for all four diagnostic categories. Also, hierarchical clustering14 using the 96 genes, identified from the ANN models, correctly clustered all 20 of the test samples (Fig. 3c). |
| 19 | nature | No | review | Cited in review | These include western blots, two-dimensional gels, methods based on protein or peptide chromatographic separation and mass spectrometric detection37–40, methods that use specific protein-fusion reporter constructs and colorimetric readouts41–44, and methods based on characterization of actively translated, polysomal mRNA45–47. |
| 20 | nature medicine | Yes | Usage | Applied in clustering of B-cell lymphoma genes | Predictive value of GC-B-cell and activated B-cell signatures. a and b, The GC and activated B-cell markers common to our dataset (b) and that of Alizadeh et al.12 (a). (90 common UniGene clusters) were hierarchically clustered45 with respect to patient samples. |
| 21 | nature genetics | Yes | Usage | Applied in clustering of gene expression data | Hierarchical clustering. We used the Cluster and TreeView software to carry out average linkage clustering, which organizes all of the data elements into a single tree with the highest levels of the tree representing the discovered classes29. |
| 22 | nature genetics | Yes (self-citation) | Usage | Applied in clustering of DNA microarrays data | We studied gene expression in the 60 cell lines using DNA microarrays prepared by robotically spotting 9,703 human cDNAs on glass microscope slides17,18. |
| 23 | Science | Yes (self-citation) | Usage | Applied in clustering of gene expression data | This subset of genes was clustered hierarchically into groups on the basis of the similarity of their expression proÞles by the procedure of Eisen et al. |
| 24 | Science | Yes | Usage | Applied in clustering of mRNA data | Tree comparing gene-expression changes resulting from different perturbations to the GAL pathway. We used the Neighbor and Drawtree programs (38) to construct a hierarchical-clustering tree (39) based on Euclidean distance between perturbation proÞles, where each proÞle consists of log10 mRNA expression ratios over the set of 997 signiÞcantly affected genes. |
| 25 | cancer cell | Yes (self-citation) | Usage | Applied in clustering of probe sets and cell lines data | Agglomerative clustering (Eisen et al., 1998) was applied to probe sets and cell lines using the uncentered Pearson’s correlations. |
| 26 | nature | Yes | Usage | Applied in clustering of tumor data | Normal brain was easily separable from the brain tumours, and the different tumour types were similarly separable. Separation of tumour types was also seen using hierarchical clustering (Fig. 1D)5. |
| 27 | nature | Yes | Usage | Applied in clustering of gene expression data | Using a matrix of Pearson correlation coef®cients from the complete pair-wise comparison of all experiments11, the 31 melanoma experiments are displayed as a hierarchical clustering dendrogram7,12 and as a three-dimensional multidimensional scaling (MDS) plot7,13. |
| 28 | nature | Yes | Usage | Applied in clustering of prostate cancer gene expression | We carried out SAM analysis by comparing gene expression profiles of metastatic prostate cancer samples with those of clinically localized prostate cancer samples from our previous work1. Genes were analysed using Cluster25, implementing average linkage hierarchical clustering of genes, and the output (Supplementary Information) was visualized by Treeview25. |
| **Bioinfo Domain** | | | | | |
| 1 | journal of computational biology | No | others | cited in introduction and compared with the proposed method | Our approach is quite different from the clustering approach used by Alon et al. (1999), Ben-Dor et al. (1999), Eisen et al. (1999), Michaels et al. (1998), and Spellman et al. (1998) in that it attempts to learn a much richer structure from the data. |
| 2 | journal of computational biology | No | review | Cited in review | A variety of clustering algorithms have been used to group together similar temporal expression patterns and thus reveal clusters of genes that seem to be co-regulated in experiments [13, 25, 36, 141, 194, 235]. |
| 3 | bioinformatics | No (self-citation) | others | Cited in introduction | Commonly used analysis methods include clustering techniques (Eisen et al., 1998; Tamayo et al., 1999), techniques based on partitioning of data (Heyer et al., 1999; Tamayo et al., 1999), as well as various supervised learning algorithms (Alter et al., 2000; Brown et al., 2000; Golub et al., 1999; Raychaudhuri et al., 2000; Hastie et al., 2000). |
| 4 | bioinformatics | No | others | Cited in introduction | Those of the first type have allowed for the identification of functionally related genes due to common expression patterns (Brown et al., 2000; Eisen et al., 1998; Wen et al., 1998; Roberts et al., 2000). |

**K-means:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | |  |  |  |  |
| 1 | Nature | No | review | Cited in review | In yeast studies, more than 50% of the genes that are transcribed in a cell cyclespecific manner and whose transcript abundance peaks in the G1 phase of the cell cycle have an MCB (Mlu cell-cycle box) within 500 base pairs (bp) of their translational start site24,68,69. |
| 2 | Nature | No | others | Cited in introduction | Direct identification of the repertoire of regulatory elements in a genome is even more challenging. The best approach so far has relied on clustering genes into functionally related subsets (for example, genes with a common biochemical function or coordinated transcription) and then searching for common sequence motifs in the general vicinity of the genes (for example, using computer programs such as MEME8 or AlignACE9; reviewed in ref. 10). |
| 3 | Nature Reviews Genetics | No | review | Cited in review | k-means clustering. If there is advanced knowledge about the number of clusters that should be represented in the data, k-means clustering is a good alternative to hierarchical methods22. |
| 4 | Annual Review of Genomics and Human Genetics | No | review | Cited in review | Some analyses achieve more specific and/or accurate functional predictions by integrating gene-expression clusters with complementary types of global data: for example, searching for shared regulatory sequences in the promoters of coexpressed genes (29, 89, 97, 111, 114, 119). |
| 5 | Nature Genetics | No | others | Cited in discussion | Although other approaches identify modules of coregulated genes and their shared cis-regulatory motifs 6,10, they do not directly suggest the regulators themselves. |
| 6 | Science | Yes | Usage | Applied in clustering of Hematopoietic Stem cells (HSCs). | We found 822 human homologs for murine HSC-related genes that are expressed in fetal liver (Database S3). Of these, 322 (39%) were enriched in human fetal HSCs. The probability of observing such an overlap by chance as estimated using hypergeometrical distribution (6, 16) is extremely low (P 10 11 ). These genes likely represent the conserved molecular components expressed in HSCs. |
| 7 | Journal of Molecular Biology | Yes (self-citation) | Usage | Applied in clustering of Saccharomyces cerevisiae expression data. | AlignACE is based on a Gibbs sampling algorithm and returns a series of motifs that are overrepresented in the input set. It previously has been used to ®nd transcriptional regulatory DNA motifs in Saccharomyces cerevisiae using groups of genes derived from genome-wide mRNA expression data (Roth et al., 1998; Tavazoie et al., 1999). While many known cis-acting elements were identi®ed, AlignACE returned many more motifs about which no literature information was found. |
| 8 | Science | No | others | Cited in introduction | For example, when relating transcription factor binding sites in the promoter regions of genes to their expression profiles, we can start by finding clusters of coexpressed genes and then search for overrepresented elements in the promoters of the genes in each cluster (2). Alternatively, we can group genes with similar binding sites in their promoter regions, and then test whether they are coexpressed (3). |
| 9 | Genome Research | No | others | Cited in background | An application of k-means clustering to expression data is provided in Tavazoie et al. (1999). One of the main problems with this method is that the number of clusters, k, must be specified prior to running the algorithm. For our data set, the number of clusters is not known in advance, and the final clustering depends heavily on the choice of k. Furthermore, clusters formed by kmeans do not satisfy a quality guarantee. |
| 10 | nucleic acids research | No | others | cited as previous work | The capabilities of the FunCat as a controlled (structured) vocabulary and classification scheme are not only limited to the functional annotation of genomes but also provide a powerful tool to analyse genome- and proteome-wide data which have been generated by large-scale transcriptome/ proteome experiments (15,20–22) as well as computational analysis of the functional networks (23,24). |
| 11 | nature | Yes | Usage | Applied in clustering of Hubs TF data | Hubs were clustered according to their propensity values using a k-means clustering algorithm (15) with k = 6, where k is the pre-defined number of clusters. |
| 12 | Science | Yes | Usage | Used for the calculation of the functional enrichment and the associated P values of *Saccharomyces* Genomes | Functional enrichment was based on the Munich Information Center for Protein Sequences (MIPS) classification of S. cerevisiae genes. Functional enrichment and the associated P values were calculated as in (53). |
| 13 | nature genetics | Yes | Usage | Use the results obtained by the clustering algorithm to detect motifs | Tavazoie et al.14 analysed the cell-cycle data of Cho et al.12 and obtained 17 motifs (many nonperiodic), of which we identified nine (MCB, SCB, STRE, Met31/32p, M3a, M3b, M27, M4 and M14a). We failed to identify Cbf1, Rap1p, M5 and ECB, the latter two because their weight matrices have more than one gap, so go beyond our dimer class. The four remaining motifs (M1a, M14b, M13 and M26) show no significant correlation with expression for any time point, and we found no similar motifs. |
| 14 | nature genetics | Yes (self-citation) | Usage | Used as reference for the analysis of cell-cycle | Both mRRPE and PAC have been identified from the same expression cluster in analyses of cell-cycle10 and stress response28,29 microarray data sets, but these studies did not capture the impressive synergy between the two motifs. |
| 15 | febs letters | No | review | Cited in review | Tavazoie et al. [8] clustered expression profiles of the 3 000 most variable yeast genes during the cell cycle (15 time points, data from Cho et al. [18]) into 30 clusters by the K-means algorithm. |
| 16 | genome biology | No | others | Cited in discussion | One could then investigate the clusters for the presence of shared regulatory motifs among the genes [31]. |
| 17 | current opinion in structural biology | No | review | Cited in review | In one of the original studies by Tavazoie et al. [10], it was reported that some of the major co-expression clusters coincided with functional groupings of genes. |
| 18 | cell | Yes | Usage | applied for clustering of gene expression data | In order to explore the statistical significance of the overlap between the set of targets of a factor and the genes expressed in a particular cell cycle stage, we used the hypergeometric distribution as described (Tavazoie et al., 1999). |
| 19 | genome research | Yes | Usage | Use the results obtained by the clustering algorithm to develop an analysis model for the differentially expressed genes | The results presented here were obtained using the hypergeometric model (Tavazoie et al. 1999; Draghici et al. 2003) in which pi is the probability of obtaining at least the observed number of differentially expressed gene, Nde, just by chance. |
| 20 | nature genetics | No | review | Cited in review | Several groups have considered ways of directly combining expression data with analysis of gene regulatory motifs91–93. |
| 21 | nature genetics | Yes (self-citation) | Usage | Use the results obtained by the clustering algorithm to generate a transcriptome-interactome correlation map for yest cell-cycle expression data | To generate a transcriptome-interactome correlation map, we first used the results of a clustering analysis 8 carried out by a k-means clustering algorithm 9 with yest cell-cycle expression data. |
| 22 | nature cell biology | No | review | Cited in review | Great advances have been made in yeast, in which it has been possible to identify common DNA motifs in the promoter regions of co-ordinately regulated genes10,12,20,21. |
| 23 | cell | Yes (self-citation) | Usage | use the results obtained by the application of the clustering algorithm on gene expression data | Two computationally discovered sequence elements, PAC and RRPE (Hughes et al., 2000; Tavazoie et al., 1999), exemplify the type of combinatorial regulation our network must describe in order to achieve predictive accuracy. |
| 24 | genomics | Yes | Usage | Use the same hypergeometric distribution for regulated genes | The probability that a certain category occurs x times just by chance in the list of differentially regulated genes is appropriately modelled by a hypergeometric distribution with parameters (N;M;K) [26]: |
| 25 | cell | Yes (self-citation) | Usage | Use the same results obtained in their previous work | We searched for overrepresented DNA sequence motifs upstream of genes within each acetylation cluster as previously described (Hughes et al., 2000; Tavazoie et al., 1999). |
| 26 | nucleic acids research | Yes | Usage | Use the same P-values as parameters | As an isolated quasi-clique may involve different functional categories, P-values (27,28) are used as criteria to assign each quasi-clique a main function. |
| **Bioinfo Domain** | | | | | |
| 1 | journal of computational biology | No | review | Cited in review | In addition, it might involve the analysis of DNA sequences to check whether a regulated gene has upstream targets for the hypothesized regulatory proteins [201]. |
| 2 | bioinformatics | Yes | Usage | used in a C Clustering Library tool | We have implemented k-means clustering, hierarchical clustering and self-organizing maps in a single multipurpose open-source library of C routines, callable from other C and C++programs. |
| 3 | bioinformatics | Yes | Usage | implemented with other clustering algorithms | To enable thorough cluster analysis and data mining the common clustering algorithms were implemented: hierarchical clustering (Eisen et al., 1998), k-means (Tavazoie et al., 1999),… |
| 4 | bioinformatics | No | review | Cited in review | The K-means algorithm (MacQueen, 1967) can be used to partition N genes into k clusters, where k is pre-determined by the user (see Tavazoie et al., 1999) for an application to yeast gene expression. |
| 5 | bioinformatics | No | others | Cited in introduction | A wide range of clustering algorithms have been proposed to analyze gene expression data, including hierarchical clustering (Eisen et al., 1998), self-organizing maps (Tamayo et al., 1999), k-means (Tavazoie et al., 1999), … |
| 6 | bioinformatics | No | review | Cited in review | Despite such advances, traditional clustering techniques such as hierarchical clustering algorithms (Eisen, 1998), k-means (Tavazoie et al., 1999), fuzzy c-means (Gasch and Eisen, 2002), finite mixture models (Yeung et al., 2001b) and SOMs (Tamayo et al., 1999) remain the predominant methods in post-genomics—a fact that is arguably more owing to their conceptual simplicity and their wide availability in standard software packages than to their intrinsic merits. |
| 7 | bmc bioinformatics | No | others | Cited in background | Since inception of the microarray technology, significant technological and analytical improvements have been introduced to meet these challenges, from experimental design [1], probe-level analysis of oligonucleotide chips [3,4], data normalization [5], statistical analysis [6], clustering techniques [7-9], to various data mining tools [10-12]. |

**Model-Based:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **BIOMED DOMAIN** | | | | | |
| 1 | Cancer Cell | Yes | usage | Applied in clustering of Gene expression data. | During this process, computational demands were reduced by randomly sampling and clustering 2000 probe intensities using mclust (Yeung et al., 2001, 2004) with two clusters and unequal variance. |
| 2 | Nature Reviews Cancer | No | review | Cited in review | Other approaches, such as CLUSTER ANALYSIS or SUPPORT VECTOR MACHINES — a technique adapted from computer science — identify groups of genes that seem to discriminate between cases and controls when used together14–16. |
| 3 | Nature Genetics | No | review | Cited in review | Model-based methods assume the data can be generated by a specified statistical model (such as a mixture of Gaussian distributions), and search for model parameters that best fit the data47,48. |
| 4 | Journal of Proteome Research | No | review | Cited in review | Several such clustering algorithms, used in microarray data analysis, could be applied, including fuzzy k-means clustering,32 model-based clustering,33 non-negative matrix factorization,34 and cliquebased clustering.35 |
| 5 | genome biology | No | others | Cited as a possible method for clustering but not used | The optimal number of clusters, C, can be inferred using one of various approaches. One possibility is to use the EM algorithm to learn the parameters for the two different models C = 1 and C = 2, and perform model selection using the Bayesian Information Criterion (BIC) score [51,52]. |
| 6 | nature reviews molecular cell biology | No | review | Cited in review | Therefore, we focus here on agglomerative clustering and refer the reader elsewhere for more detailed methods17,18. |
| 7 | trends in genetics | No | review | Cited in review | Nevertheless, we have fitted mixture models [46] to data from 58 lines of Drosophila and using a Bayesian information criterion found evidence for a slight excess of multimodal transcript distributions suggesting that at least 5% of the transcriptome could show distinguishable frequency classes of transcription (W. Hsieh et al., unpublished data). |
| 8 | genome biology | No (self-citation) | review | Cited in review | There are numerous algorithms and associated programs to perform cluster analysis (for example, hierarchical methods [5], self-organizing maps [6], k-means [7] and model-based approaches [8-10]) and many of these techniques have been applied to expression data (for example [1,11-14]). |
| 9 | journal of clinical oncology | Yes | usage | Applied for clustering of both ER and PR mRNA measurements | Using a Gaussian model-based clustering technique,24,29 both ER and PR mRNA measurements were found to be bimodal, which was most pronounced for ER (Fig 1). |
| 10 | cytometry part A | No | others | cited as previous work | Furthermore, confirmed by results of our simulation studies, our proposed approach is robust against model misspecification and can avoid the problem of Gaussian mixture models that excessive clusters are often needed to provide a reasonable fit in case of model misspecification (34). |
| 11 | nature genetics | Yes | Usage | Applied in clustering of *lymphoblastic leukemia* expressiondata | A threshold of 10% was the antimode that discriminated two major modes in the US ALL population (Supplementary Fig. 1)29 and approximately the same 10% value maximally differentiated relapse risk (Supplementary Note). |
| 12 | genome biology | No | others | Cited in background | It has a clear definition that a cluster is a subpopulation with a certain distribution, and several statistical methods can be applied to estimate the number of clusters. Some authors have considered its application to cluster gene-expression patterns [10-12]. |
| **Bioinfo Domain** | | | | | |
| 1 | bioinformatics | No | review | Cited in review | Despite such advances, traditional clustering techniques such as hierarchical clustering algorithms (Eisen, 1998), k-means (Tavazoie et al., 1999), fuzzy c-means (Gasch and Eisen, 2002), finite mixture models (Yeung et al., 2001b) and SOMs (Tamayo et al., 1999) remain the predominant methods in post-genomics—a fact that is arguably more owing to their conceptual simplicity and their wide availability in standard software packages than to their intrinsic merits. |
| 2 | bioinformatics | No | others | Cited in introduction | Recently, clustering method based on multivariate normal mixture model, in which each gene is assumed to have come from a mixture of multivariate normal densities with different means and certain parameterization of the covariance matrix (Fraley and Raftery, 1998), was also used to cluster genes based on their expression profiles over time or over different samples (Yeung et al., 2001). |
| 3 | bioinformatics | No | others | Cited in introduction and compared with the proposed method | Based on these estimates, clusters are then formed based on the pesterior probabilities of individual profiles of being generated by a particular mixture component ( McLachlan and Basford, 1987; Frely and Reftery, 1999; Yeung et al., 2001a). Results of this analysis were compared to our infinite mixture model. |
| 4 | bmc bioinformatics | No | others | Cited in background | Another family of fuzzy clustering approaches is based on Gaussian Mixture Model (GMM) [18-20], where the dataset is assumed to be generated by a mixture of Gaussian distributions with certain probability, and an objective function is calculated based on the mixture Gaussians as the likelihood of the dataset being generated by such model. |
| 5 | bioinformatics | No | comparison | Cited in comparison | Forcing all these genes into cluster formation can introduce more false positives and distort the structure of identified clusters. Model-based clustering (Yeung et al., 2001a; Fraley and Raftery, 2002b; Medvedovic and Sivaganesan, 2002; McLachlan et al., 2002; Medvedovic et al., 2004) and tight clustering (Tseng and Wong, 2005) are two examples among this category. |
| 6 | biometrics | No | others | Cited in introduction | In contrast to heuristic methods, model-based clustering methods make inferences based on a probabilistic assumption of the data distribution (Fraley and Raftery, 1998; Yeung et al., 2001; Medvedovic and Sivaganesan, 2002). |
| 7 | bioinformatics | No (self-citation) | others | cited in the introduction | Virtually all classical clustering algorithms (Eisen et al., 1998; Tavazoie et al., 1999; Tamayo et al., 1999; Yeung et al., 2001), as well as a multitude of brand new procedures (Herrero et al., 2001), have been applied in the context of clustering microarray data. |
| 8 | bioinformatics | No | comparison | Cited in comparison | Yeung et al. (2001a) and Ghosh and Chinnaiyan (2002) have considered Gaussian mixture clustering of microarray data, but they did not apply the method to missing value estimations. In a Gaussian mixture, each component is modeled by a multivariate normal distribution. |

PCA:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **Biomed Domain** | | | | | |
| 1 | nature reviews genetics | No | review | Cited in review | PCA (also called singular value decomposition) is a mathematical technique that exploits these factors to pick out patterns in the data, while reducing the effective dimensionality of gene-expression space without significant loss of information23. |
| 2 | journal of immunology | Yes | Usage | Applied in clustering of expression microarray data to assess Transcriptional profile analysis | Principal component analysis (PCA)3 was conducted on all genes analyzed to assign the general variability in the data to a reduced set of variables called principal components (22). |
| 3 | nature genetics | No | review | Cited in review | Principal components analysis31,32, singular value decomposition33,34 and multidimensional scaling35,36 are related dimensionreduction techniques that can be used for visualizing large data sets. |
| 4 | genome research | Yes | Usage | Appled in clustering of microarray experiment. | The SVD has been applied to microarray experiment analysis in order to find underlying temporal and tumor patterns (Alter et al. 2000; Holter et al. 2000; Raychaudhuri et al. 2000; Lian et al. 2001). |
| 5 | nature reviews drug discovery | No | review | Cited in review | Principal-components analysis. Principal-components analysis is more useful as a visualization technique than as an analytical method33,36. |
| 6 | laboratory investigation | Yes | Usage | Appled in clustering of of gene expression within the set of 13 housekeeping genes. | Analysis for patterns of gene expression within the set of 13 housekeeping genes was performed with PCA using the SAS package (SAS Institute Inc.). PCA is a linear dimensionality reduction technique, which identifies (orthogonal) directions of maximum variance in the original data. The data are projected in a lower dimension, formed of (a subset of) components with the highest variance.21,22 |
| 7 | nature immunology | Yes | Usage | Applied in clustering of D1 cells with Gram-negative Escherichia coli and transcriptionally analyzed immature cells. | To get an approximate visualization of our entire dataset, without losing experimental information (variance), we first applied the principal component analysis (PCA) method, which allows the dimensionality of complex data to be reduced12. Thus, we were able to globally describe features of the kinetic points that best explained the corresponding transcriptome (Table 1). |
| 8 | genome biology | No | others | cited as previous work | Many of the eigen vectors identified by PCA seemed to contain little information about the dataset, as previously noted for this type of analysis [18,19]. |
| 9 | current opinion in immunology | No | review | Cited in review | SVD has also been applied (in this case referred to as principal components) by Raychaudhuri et al. 24 to the sporulation data of Chu et al., 25 and was successful in readucing the feautures of the data to their principal components. |
| 10 | trends in genetics | No | review | Cited in review | Some commonly used examples include principal component analysis (PCA) [26] or singular value decomposition (SVD) [27] for dimensionality reduction, as well as hierarchical clustering [28], K-means clustering [29] and self organizing maps (SOMs) [30] for clustering. There are already several excellent reviews on various unsupervised analyses and their applications in microarray data mining [31–33], therefore we do not discuss their details here. |
| 11 | biotechniques | No | review | Cited in review | A large number of distance measures and analytic techniques have been implemented, including: • Hierarchical clustering (HCL) (26) • Bootstrapped/jackknifed HCL • k-means/k-medians clustering (KMC) (27) • k-means/k-medians support (iterative KMC) • Self-organizing maps (SOMs) (28) • Self-organizing trees (SOTA) (29) • Cluster affinity search technique (CAST) (30) • Figure of merit for CAST and KMC (soon also for SOMs) (31) • QT-clust (32) • Principal component analysis (PCA) (33) |
| 12 | nucleic acids research | No | others | cited in introduction | Analysis results obtained using clustering algorithms, such as hierarchical clustering (8), K-means clustering and selforganizing maps (9), or data dimension reduction and projection methods such as singular value decomposition (10) or principal component analysis (11), will be in¯uenced by the estimates replacing the missing values. |
| 13 | neurogenetics | Yes | Usage | applied in clustering of microarray expression data from 20 regions of the human CNS. | To gain a global perspective of the transcriptional profile similarities and differences among the 65 tissues processed, PCA and hierarchical clustering were performed using all probe sets present on the array. Figure 1a shows the results of PCA. |
| 14 | genes & development | No | review | Cited in review | This is often performed using principal component analysis (PCA) (Raychaudhuri et al. 2000) or multidimensional scaling (MDS) (Khan et al. 1998). |
| 15 | nature | Yes | Usage | Applied in clustering of WT, ASO-treated, and saline-treated microarray samples | Comparisons between groups were summarized and rank-ordered by fold-changes of mean expression level and t tests. The software package R 38 was used to perform principal components analysis (PCA)39, 40 on WT, ASO-treated, and saline-treated microarray samples. The principal components allowed the capture of the majority of the expression variation in each sample within 3 dimensions. |
| 16 | Science | Yes | Usage | Applied in clustering of IRA B cells data | Unsupervised hierarchical clustering (Fig. 2D) and principal component analysis (PCA) (Fig. 2E) grouped IRA B cells in a separate population from T1, FO, MZ, B1a, and PC. IRA B cells also gave rise to a unique transcriptome signature (fig. S5, B to D, and table S1) and expressed genes relevant to B cell biology (fig. S5D). |
| 17 | journal of immunology | Yes | Usage | Applied to identify sources of variability among genes | From this restricted gene comparison, we moved to a global gene profiling comparison of TC-Mf, TAM, M1, and M2 cells. To identify sources of variability among these entire databases, a PCA was performed (50, 54, 55). |
| 18 | molecular and cellular biology | Yes | Usage | Applied in clustering of mRNA gene expression data | Data analysis. Normalization of the image files was accomplished with quantile normalization, and signal was estimated with the positional dependent nearest neighbor (PDNN) algorithm (39). To assess quality, principal component analysis (28) was used; plots were generated and points were colored according to experimental condition (WT or KO), time (0 to 120 min), and replicate (samples 1 to 5). |
| 19 | current opinion in genetics & development | No | review | Cited in review | As mRNA levels are the output of gene-expression regulatory networks, it is theoretically possible to use expression data to reverse engineer the architecture of the controlling regulatory networks. A number of groups have tackled this problem using singular value decomposition (SVD) analysis [32–35]. |
| 20 | nature cell biology | Yes | Usage | Applied in clustering of gene expression data | Principal component analysis (PCA) was used to explore correlations among gene expression microarray datasets for three biological replicates for each of the four different conditions (Tcf3+/+(C), Tcf3−/−(C), Tcf3+/+(W), Tcf3−/−(W)) (see Methods)25. |
| 21 | journal of leukocyte biology | Yes | Usage | Applied in clustering of microarray gene expression data | Data were extracted and analyzed (Genespring GX software, Agilent). Normalization was performed according to the Feature Extraction One-color protocol. Principal component analysis (PCA) was conducted on all genes analyzed [38], and data are presented in Supplemental Fig. 1. |
| 22 | placenta | Yes | Usage | Applied in clustering of gene expression profile of the three patients | Principal component analysis [21] and hierarchical clustering showed that placental gene expression profile of the three patients with HELLP syndrome was similar to that of other preeclamptic women. |
| 23 | current opinion in structural biology | No (self-citation) | review | Cited in review | The major initial efforts at clustering and linear decomposition (such as principal components analysis) not only assist humans in understanding the data, but also demonstrate that the amount of independent new information may be much smaller than the number of raw data points suggests [2,11]. (Some microarray analysis tools are available at http://classify.stanford.edu/) |
| 24 | blood | Yes | Usage | Applied in clustering of samples according to the similarity of their gene expression patterns | To cluster the samples according to the similarity of their gene expression patterns, we performed unsupervised principal component analysis (PCA) using SPSS software (SPSS, Chicago, IL). Data were first log transformed and median centered. To select genes that are most informative, the analysis was limited to a subset of 1174 genes with the highest standard deviation across the 49 samples (SD 1200). The goal of PCA is to transform the original variables into new independent and uncorrelated variables called principal components that explain the variability 25. |
| **BIOINFO DOMAIN** | | | | | |
| 1 | bioinformatics | No (self-citation) | comparison | Cited in comparison | Many analysis methods, such as principle components analysis or singular value decomposition, require complete matrices (Alter et al., 2000; Raychaudhuri et al., 2000). |
| 2 | bioinformatics | Yes | Usage | Used as a clustering algorithm in Genesis tool | To enable thorough cluster analysis and data mining the common clustering algorithms were implemented: hierarchical clustering (Eisen et al., 1998), k-means (Tavazoie et al., 1999), self-organizing maps (Tamayo et al., 1999), principal component analysis (Raychaudhuri et al., 2000), and support vector machines (Brown et al., 2000). |
| 3 | bioinformatics | No | others | Cited in introduction | However, Raychaudhuri et al. (2000) reported that these seven clusters are very poorly separated when the data is visualized in the space of the first two PCs , even though they account for over 85% of the variation on the data. |
| 4 | bioinformatics | No | others | Cited in introduction | Moreover, many multivariate statistical analyses, like principal component analysis (PCA) (Raychaudhuri et al., 2000) and singular value decomposition (SVD) (Alter et al., 2000), cannot be applied to data with missing values. Thus, in order to avoid improper analyses, missing value estimation is an important preprocess. |
| 5 | bioinformatics | No | others | cited as previous work | Though it has points of contact with Fourier analysis (e.g. as used in Spellman et al., 1998) and SVD (Alter et al., 2000; Holter et al., 2000; Raychaudhuri et al., 2000), it is not a method of decomposing signals into a weighted sum of basis functions or vectors and simplifying them by focusing on principal terms |
| 6 | bioinformatics | No | others | Cited in introduction and compared with the proposed method | PCA assumes its first components to capture a maximal amount of data variance. This constrains the modes, as well as their influences, to be orthogonal. Although the biological interpretation of individual principal components is not obvious, PCA can be expected to seperate a subspace of biological effects from the subspace of weaker noise components. |
| 7 | bioinformatics | Yes | Usage | Used as a clustering algorithm in J-Express tool | The methods implemented have been successfully applied to the analysis of gene expression data (Eisen et al., 1998; Tavazoie et al., 1999; Raychaudhuri et al., 2000; Tamayo et al., 1999). |
| 8 | bioinformatics | Yes | Usage | Used as a clustering algorithm in MADE4 R package | Although PCA is available in several R packages, including stats and amap, the R package ade4 contains many additional multivariate statistical methods including methods for analysis of one-data matrix, coupling of two-data matrices or multi-table analysis, http://cran.univ-lyon1.fr/doc/Rnews/4Rnews\_2004-1.pdf (Thioulouse et al., 1997; Chessel et al., 2004). |

**SOM:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **BIOMED DOMAIN** | | | | | |
| 1 | Science | Yes (self-citation) | Usage | Applied in clustering of gene expression monitoring data. | To cluster tumors, we used a technique called self-organizing maps (SOMs), which is particularly well suited to the task of identifying a small number of prominent classes in a data set (32). In this approach, the user specifies the number of clusters to be identified. The SOM finds an optimal set of “centroids” around which the data points appear to aggregate. It then partitions the data set, with each centroid defining a cluster consisting of the data points nearest to it. |
| 2 | Genome Research | No | others | Cited in introduction | Similarly, for gene expression profiles and other molecular states, numerous programs such as GeneCluster (Tamayo et al. 1999), Tree-View (Eisen et al. 1998), and GeneSpring (www.silicongenetics.com) are available for clustering, classification, and visualization. |
| 3 | Nature Genetics | Yes (self-citation) | Usage | Applied in clustering of gene expression data. | To identify sets of coregulatedgenes, we used self-organizing maps to group the GNF mouse expression atlas into 36 clusters12,31. |
| 4 | Nature Genetics | No | others | Cited in the abstract | Technologies to measure whole-genome mRNA abundances1−3 and methods to organize and display such data4−10 are emerging as valuable tools for systems-level exploration of transcriptional regulatory networks. |
| 5 | Nature | No | review | Cited in review | An increasingly common approach involves using the gene expression behaviour observed over multiple experiments to first cluster genes together into groups (see Fig. 3), either by manual examination of the data24, or by using statistical methods such as selforganizing maps64. |
| 6 | Nature Medicine | Yes (self-citation) | Usage | Applied in clustering of gene expression data. | Arrays were subsequently developed with phycoerythrin-conjugated streptavidin (SAPE) and biotinylated antibody against streptavidin, and scanned to obtain quantitative gene-expression levels14,43. |
| 7 | Nature Genetics | No | others | Cited in introduction | The assessment of gene expression patterns in a multitude of cell and tissue types, such as the diverse set of cell lines we studied here, under diverse conditions in vitro and in vivo, should lead to increasingly detailed maps of the human gene expression program and provide clues as to the physiological roles of uncharacterized genes11–16. |
| 8 | Science | Yes | Usage | Applied in clustering of gene expression data (modified). | The 997 affected genes were clustered based on Euclidean distance between their log10 expression ratios over all perturbation conditions, using a 4 row by 4 column self-organizing map (SOM) implemented by the GeneCluster application (28), with 100 epochs. |
| 9 | Nature | Yes (self-citation) | Usage | Applied in clustering of tumors data. | The tumours were clustered into two groups using self-organizing maps (SOM); an unsupervised algorithm that groups samples into a predetermined number of clusters on the basis of their gene expression patterns6,25. |
| 10 | Nature Genetics | No | review | Cited in review | Even if data-mining analysis is going to be done using, for example, one or more of the widely used clustering methods10–12, it is still extremely useful to reduce the data set to those genes that are most variable between samples. |
| 11 | nature reviews genetics | No | review | Cited in review | Several techniques have been used for the analysis of gene-expression data, including hierarchical clustering5-9,mutual information5,10 and self-organizing maps (SOMs)11,12. |
| 12 | annual review of genomics and human genetics | No | review | Cited in review | Genes with similar responses over multiple conditions are often clustered together to form functional groups or to reveal coordinated patterns of expression. Several clustering methods have been proposed: Most are excellent and have associated software packages that are publicly available (4, 23, 25, 36, 39, 57, 62, 118). |
| 13 | immunity | Yes | Usage | Applied to clustering of gene expression data | These genes were clustered using the self-organizing map (SOM) algorithm, a statistical method for grouping genes based on expression patterns independent of expressionmagnitude (Tamayo et al., 1999) (Figure 2A). |
| 14 | biotechniques | No | others | cited in conclusion | Recently, hierarchical clustering (4) and self-organizing maps (11) have been applied to the analysis of microarray expression data across multiple experiments. |
| 15 | genome biology | No | others | Cited as previous work | Cluster analysis is a popular method for analyzing the data of a series of microarrays [6, 7]. |
| 16 | blood | No | review | Cited in review | Current efforts towards the identification of ATRA target genes are using microarrays of immobilized cDNAs152,153 or oligonucleotides, 154 which can monitor expression of thousands of gene simultaneously.155,156 |
| 17 | genome research | No | others | Cited as previous work | The method of self-organizing maps (SOM) has been applied recently to expression data (Tamayo et al.1999). This method is closely related to the k-means procedure (Kohonen 1997). |
| 18 | trends in genetics | No | others | Cited as previous work | Statistical and bioinformatical analyses of these profile data reveal genes and gene regulation events by either (non)hierarchical cluster analysis7, referenced or supervised classification approaches8,9 or correlationbased analyses10. |
| 19 | nature genetics | Yes | Usage | Applied for clustering of gene expression profiles among the 413 *SAR-associated ESTs* | We applied two commonly used algorithms to analyse gene expression profiles among the 413 SAR-associated ESTs. A hierarchical ‘clustergram’ of genes was grouped by both related regulation patterns and expression amplitudes20 (Fig. 2). ‘Self-organizing maps’ (SOMs) generate expression profiles organized by shape, essentially independent of amplitude21 (Fig. 3). |
| 20 | genome biology | No | others | Cited in background | This can be done using data-driven methods, such as hierarchical clustering and selforganizing maps [1,2], which identify groups of genes with similar expression patterns. |
| 21 | plant cell | Yes | Usage | Used for clustering of transcription factor genes according to their expression profiles | We clustered the transcription factor genes according to their expression profiles at different stages of leaf development. The two-dimensional self-organizing map algorithm (Tamayo et al., 1999) was used to gain an overview of the behavior of each gene relative to the others during the course of leaf development. |
| 22 | cancer cell | Yes (self-citation) | Usage | Applied for clustering of RNA microarray data | A total of 10 g of total RNA from each sample was used to prepare biotinilated target RNA as previously described with minor modifications (Golub et al., 1999; Tamayo et al., 1999; Wodicka et al., 1997). |
| 23 | Science | Yes | Usage | Applied for clustering of expression profiles of strictly maternal genes | A self-organizing map (SOM) algorithm (10), applied to the data from all 1212 maternally deposited genes, identified a cluster of 27 “strictly” maternal genes. Transcripts from almost all 27 of these genes were degraded after fertilization and were not subsequently expressed at high levels until they appeared in the female germ line during oogenesis (Fig. 2C). |
| **Bioinfo Domain** | | | | | |
| 1 | journal of computational biology | No | review | Cited in review | A variety of clustering algorithms have been used to group together similar temporal expression patterns and thus reveal clusters of genes that seem to be co-regulated in experiments [13, 25, 36, 141, 194, 235]. |
| 2 | bioinformatics | No | comparison | Cited in the introduction in a comparative study | Commonly used analysis methods include clustering techniques (Eisen et al., 1998; Tamayo et al., 1999), techniques based on partitioning of data (Heyer et al., 1999; Tamayo et al., 1999), as well as various supervised learning algorithms (Alter et al., 2000; Brown et al., 2000; Golub et al., 1999; Raychaudhuri et al., 2000; Hastie et al., 2000). |
| 3 | bioinformatics | No | others | Cited in introduction | In order to extract information from gene expression measurements, different methods have been employed this data including … self-organizing maps (Tamayo et al., 1999; Golub et al., 1999) … |
| 4 | bioinformatics | Yes | Usage | Implemented in a open-souce library | We have implemented k-means clustering, hierarchical clustering and self-organizing maps in a single multipurpose open-source library of C routines, callable from other C and C++programs. Using this library, we have created an improved version of Michael Eisen’s well-known Cluster program for Windows, Mac OS X and Linux/Unix. |
| 5 | bioinformatics | No | others | Cited as previous work | A number of ad hoc clustering procedures have been applied to DNA microarray data (Eisen et al., 1998; Alon et al., 1999; Furey et al., 2000; Heyer et al., 1999; Tamayo et al., 1999) without any clear emerging consensus. |
| 6 | journal of computational biology | No | others | Cited in introduction | Looking ahead, we believe they will also be useful in assessing the quality of the results from higher-order analyses such as clustering (Eisen et al., 1999; Tamayo et al., 1999). |
| 7 | journal of computational biology | No | others | Cited in introduction | Self-Organizing Maps were used by Tamayo et. al [33]. |
| 8 | bioinformatics | Yes | Usage | integrated in the Genesis platform | To enable thorough cluster analysis and data mining the common clustering algorithms were implemented: hierarchical clustering (Eisen et al., 1998), k-means (Tavazoie et al., 1999), self-organizing maps (Tamayo et al., 1999), principal component analysis (Raychaudhuri et al., 2000), and support vector machines (Brown et al., 2000). Detailed information on the supported data formats and used methods is available on the web site and in the accompanying documentation. |
| 9 | bioinformatics | No | review | Cited in review | The Self-Organized Map (SOM) method is closely related to K-means and has been applied to mRNA expression data of yeast-cell cycles as well as hematopoietic differentiation of four well-studied model cell lines (Tamayo et al., 1999). |
| 10 | bioinformatics | No | others | Cited in introduction and compared with the proposed methods | It is instructive to compare existing methods for clustering genes by expression, e.g., (Eisen et al., 1998; Tamayo et al., 1999), to those introduced here. |
| 11 | bioinformatics | No | others | Cited as previous work | The data was normalized to have mean 0 and variance 1 across each cell cycle as suggested in Tamayo et al. (1999). |
| 12 | bioinformatics | No | others | Cited in introduction | A wide range of clustering algorithms have been proposed to analyze gene expression data, including hierarchical clustering (Eisen et al., 1998), self-organizing map (Tamayo et al., 1999), … |
| 13 | biostatistics | No | others | Cited in introduction | Following the precedent of the leading experiments, 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. (1999), Lazzeroni and Owen (2000), and Tamayo et al. (1999) to name a few). |

**SVG:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ID** | **Journal** | **Application** | **Type** | **Comments** | **Excerpt** |
| **BIOMED DOMAIN** | | | | | |
| 1 | nature medicine | No | others | Cited in discussion | Other approaches that share the parametric nature of ANNs and have been utilized to classify geneexpression profiles include Support Vector Machines23. Thus far, these other methods have not been fully explored to extract the genes or features that are most important for the classification performance and which also will be of interest to cancer biologists24. |
| 2 | nature medicine | Yes | Usage | Applied in clustering of gene expression data from large B-cell lymphoma | The detailed results from using the SVM to predict outcome are in the DLBCL Outcome Prediction section. |
| 3 | Nature | Yes | Usage | Applied in clustering of tumors samples data | We subsequently tested several other classi®cation algorithms including weighted voting6,27, support vector machines28,29 and IBM SPLASH30, all of which performed with similarly high accuracy (Supplementary Information I and III). |
| 4 | nature reviews genetics | No | review | Cited in review | The techniques discussed so far are unsupervised methods for identifying patterns of gene expression. Supervised methods represent a powerful alternative that can be applied if one has some previous information about which genes are expected to cluster together. One widely used example is the support vector machine (SVM)24. |
| 5 | annual review of genomics and human genetics | No | review | Cited in review | Several clustering methods have been proposed: Most are excellent and have associated software packages that are publicly available (4, 23, 25, 36, 39, 57, 62, 118). |
| 6 | oncogene | Yes | Usage | Applied in clustering of miRNAs samples data | SVMs have been shown to perform well in multiple areas of biological analysis including evaluating microarray date (Brown et al., 2000). We attempted to use this algorithm to predict the classification of samples into cancer and non-cancer groups. |
| 7 | trends in genetics | No | others | cited as previous work | Statistical and bioinformatical analyses of these profile data reveal genes and gene regulation events by either (non)hierarchical cluster analysis7, referenced or supervised classification approaches8,9 or correlationbased analyses10. |
| 8 | neuroimage | No | others | Cited in introduction | Pattern recognition techniques have been applied to a wide range of practical problems, from face recognition (Papgeorgiou and Poggio, 1999) to the analysis of DNA microarray data (Brown et al., 2000). |
| 9 | genetics research | No | others | Cited in introduction | After one has estimates of relative expression in hand, there are problems in data visualization, dimension reduction (Hilsenbeck et al., 1999), and pattern recognition (Brown et al., 2000). |
| 10 | febs letters | No | review | Cited in review | Brown et al. [22] have applied various supervised learning algorithms to six functional classes of yeast genes using gene expression matrices from 79 samples [6]. Genes from some of the classes such as ribosomal proteins and histones are expected to be coexpressed. |
| 11 | nature genetics | No | review | Cited in review | Support vector machines (SVMs) are a family of statistical machine-learning methods that have been proposed as particularly suitable to the dimensions of microarray learning problems68,69. |
| 12 | genome research | No | others | cited as previous work | Analysis of gene expression data is currently one of the most exciting areas in genomics. Computationally, it involves clustering and grouping individual expression measurements and interrelating them to other sources of information, such as phenotypes, functional classifications, or cellular responses (Golub et al. 1999; Brown et al. 2000; Califano et al. 2000; Gaasterland and Bekiranov 2000; Raychaudhuri et al. 2001; Subrahmanyam et al. 2001). |
| 13 | nature cell biology | No | review | Cited in review | In general, one can discriminate between unsupervised methods, such as k-means clustering, principal-component analysis and self-organizing maps17,18, and supervised methods, in which a subset of the data is used to train the system to discriminate within the remaining set19. |
| 14 | genome research | No | others | Cited as previous work | Examples include attempts to identify (yet unknown) subclasses of tumors, or work on identifying clusters of genes that are coregulated or share the same function (Brown et al. 2000; Mateos et al. 2002). |
| 15 | journal of molecular biology | No | others | Cited in discussion | While this article was being ®nalized, other papers appeared which had used the SVM method to functionally classify genes by using gene expression data from DNA microarray hybridization experiments,2 |
| 16 | nature reviews drug discovery | No | review | Cited in review | Support vector machines. Support vector machines address the problem of finding combinations of genes that better split sets of biological samples30. |
| 17 | trends in neurosciences | No | others | cited as previous work | Reporting the ‘most changed genes’ reduces the microarray approach to a high throughput northern hybridization strategy, failing to take advantage of simultaneously obtained transcriptome differences between experimental and control subjects. Many complex methods for data analysis are being developed26,58–65. |
| 18 | nucleic acids research | No | others | cited as previous work | Known for its high accuracy and good performance, SVM is a widely used classification tool in bioinformatics analysis such as microarray-based cancer classification (19,20), prediction of protein function (21,22) and prediction of subcellular localization (23,24). |
| **Bioinfo Domain** | | | | | |
| 1 | journal of computational biology | No | review | Cited in review | A variety of clustering algorithms have been used to group together genes with similar temporal expression patterns (Alon et al., 1999; Ben-Dor et al., 1999; Brown et al., 2000; Cho et al., 1998; Eisen et al., 1998; Holter et al., 2001; Michaels et al., 1998; Spellman et al., 1998; Tamayo et al., 1999; Wen et al., 1998). |
| 2 | bioinformatics | No | comparison | Cited in comparison | Commonly used analysis methods include clustering techniques (Eisen et al., 1998; Tamayo et al., 1999), techniques based on partitioning of data (Heyer et al., 1999; Tamayo et al., 1999), as well as various supervised learning algorithms (Alter et al., 2000; Brown et al., 2000; Golub et al., 1999; Raychaudhuri et al., 2000; Hastie et al., 2000). |
| 3 | bioinformatics | Yes (self-citation) | Usage | Applied to clustering of DNA microarray experiments | This in turn corresponds to an asymmetric margin; i.e. the class with smaller *d* will be kept further away from the decision boundary (Brown et al., 2000). |
| 4 | bioinformatics | Yes | Usage | Used as a clustering algorithm in Genesis tool | To enable thorough cluster analysis and data mining the common clustering algorithms were implemented: hierarchical clustering (Eisen et al., 1998), k-means (Tavazoie et al., 1999), self-organizing maps (Tamayo et al., 1999), principal component analysis (Raychaudhuri et al., 2000), and support vector machines (Brown et al., 2000). |
| 5 | bioinformatics | No | others | Cited in introduction | The SVM method has been successfully applied to isolated handwritten digit recognition (Cortes and Vapnik, 1995; Scholkopf et al., 1995), object recognition (Roobaert and Hulle, 1999), text categorization (Drucker et al., 1999), microarray data analysis (Brown et al., 2000), protein secondary structure prediction (Hua and Sun, 2001), etc. |
| 6 | bioinformatics | No | others | Cited in introduction | SVM is a new descriminative method (Vapnik, 1995), which has demenstrated high classification accuracy in protein family (evolutionary relationship) prediction (Jaakkola et al. 1999), gene expression classification (Brown et al., 2000). |
| 7 | bioinformatics | No | others | Cited in introduction | A wide range of clustering algorithms have been proposed to analyze gene expression data, inclusding hierarchical clustering (Eisen et al., 1998), self-organizing maps (Tamayo et al., 1999), K-means (Tavazoie et al., 1999), graph-theoretics approaches ( for example, Ben-Dor and Yakhini, 1999; Hartuv et al., 1999), and support vector machines (Brown et al., 2000). |
| 8 | journal of computational biology | No | others | cited as previous work in the conclusions section | Brown et al. (1999) use support vector machines in the context of gene expression data. In contrast to our approach, they attempt to classify the genes rather than samples. Thus, they deal with the dual classi cation problem. The characteristics of their classi cation problem are quite different: many examples (i.e., thousands of genes) and few attributes (i.e., expression in different samples). |
| 9 | bioinformatics | No | comparison | Cited in comparison | Fig. 2. MC-SVM algorithms applied to a three-class diagnostic problem. (a) MC-SVM OVR constructs three classifiers: (1) class 1 versus classes 2 and 3; (2) class 2 versus classes 1 and 3; and (3) classes 3 versus classes 1 and 2. (b) MC-SVM OVO constructs three classifiers: (1) class 1 versus class 2; (2) class 2 versus class 3; and (3) class 1 versus class 3. (c) MC-SVM DAGSVM constructs a decision tree on the basis of OVO SVM classifiers. (d) MC-SVM methods by Weston and Watkins and by Crammer and Singer construct a single classifier by maximizing margin between all classes simultaneously. |
| 10 | bioinformatics | Yes (self-citation) | Usage | Applied to clustering of microarray data | Among the 30 train/test splits, seven proteins are consistently mislabeled by SDP/SVM (see online supplement). These include one, YLR406C (RPL31B), that was previously misclassified as non-ribosomal in an SVM-based study using a smaller microarray expression dataset (Brown et al., 2000). |
| 11 | bioinformatics | No | others | Cited in introduction | However, while binary classification and clustering are heavily studied, (see, e.g. Alizadeh et al., 2000; Ben-Dor et al., 2000; Brown et al., 2000; Der et al., 1998; Eisen et al., 1998; Friedman et al., 2000; Raychaudhuri et al., 2002; Tamayo et al., 1999;Welsh et al., 2001), only a small amount of work has been made on multiclass classification, i.e. classification involving more than two classes (Alizadeh et al., 2000; Golub et al., 1999; Khan et al., 2001; Ross et al., 2000; Tamayo et al., 1999). |
| 12 | bioinformatics | No | others | Cited as previous work | Little exprerience exists in the application of SVMs to biomolecular problems (we only know of work on remote homology detection (Jaakkola et al., 1999) and on gene expression analysis (Brown et al., 2000)). |