BMC Bioinformatics

BioMed Central

Research article

Open Access

Selection of informative clusters from hierarchical cluster tree with gene classes

Petri Toronen*

Address: A. I. Virtanen Institute for Molecular Sciences, Neulaniementie 2, P.O. Box 1627, FIN-70211 Kuopio, Finland

Email: Petri Toronen* - toronen@hytti.uku.fi

* Corresponding author

Published: 25 March 2004

BMC Bioinformatics 2004, **5**:32

Received: 29 August 2003

Accepted: 25 March 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/32

© 2004 Toronen; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

Abstract

Background: A common clustering method in the analysis of gene expression data has been hierarchical clustering. Usually the analysis involves selection of clusters by cutting the tree at a suitable level and/or analysis of a sorted gene list that is obtained with the tree. Cutting of the hierarchical tree requires the selection of a suitable level and it results in the loss of information on the other level. Sorted gene lists depend on the sorting method of the joined clusters. Author proposes that the clusters should be selected using the gene classifications.

Results: This article presents a simple method for searching for clusters with the strongest enrichment of gene classes from a cluster tree. The clusters found are presented in the estimated order of importance. The method is demonstrated with a yeast gene expression data set and with two database classifications. The obtained clusters demonstrated a very strong enrichment of functional classes. The obtained clusters are also able to present similar gene groups to those that were observed from the data set in the original analysis and also many gene groups that were not reported in the original analysis. Visualization of the results on top of a cluster tree shows that the method finds informative clusters from several levels of the cluster tree and indicates that the clusters found could not have been obtained by simply cutting the cluster tree. Results were also used in the comparison of cluster trees from different clustering methods.

Conclusion: The presented method should facilitate the exploratory analysis of big data sets when the associated categorical data is available.

Background

Some of the most common methods used for clustering and visualization of gene expression data are the hierarchical agglomerative clustering methods [1] where the data points and/or the clusters are repetitively joined in a hierarchical fashion. Initial analysis of hierarchical cluster tree often relies on cutting the tree at some level or inspecting the sorted gene list based on the tree. Cutting the cluster tree at a certain level and analyzing only the resulting clusters will miss information that can be present at the other levels of the hierarchical cluster tree. Analysis of sorted gene lists on the other hand usually involves the usage of short descriptions of the gene function and a lot of manual labour. Clusters cut and obtained gene lists leave several questions unanswered: Is there some common feature to genes included in a cluster and what information is presented by each of the clusters in the cluster tree?

Here a novel and simple method that should facilitate the analysis of cluster trees is proposed based on the existing categorizations of genes to functional gene classes obtained from databases. The presented work used the categorizations of the genes according to biological process, molecular function and cellular localization available from Saccaromyches Genome Database (SGD, [2]). The functional, complex and component categorizations from Munich Information center for Protein Sequences (MIPS, [3]) were also used. These classifications enable the selection of important clusters for interpretation of the data. As a by-product the co-regulated genes from the same gene class give strong support to actual regulation of biological system presented by a gene class.

Method compares all the clusters in the cluster tree with all the classes using a measure that is similar to a commonly used hypergeometric distribution-based p-value measure [4] and looks for optimal correlation of the gene classes and the clusters from different tree branches. As a result it selects the best scoring clusters from varying levels of the cluster tree and also presents the information on what were the associated gene classes. This directs the analysis to biologically most significant clusters. The obtained clusters are also visualized on top of the cluster tree enabling an overview of distribution of different enriched functional classes. Visualization is also shown using only those clusters that were associated with protein synthesis demonstrating the analysis of clusters that are involved in the same function. Cluster tree visualization was also used as a starting point for the analysis of two clusters having enriched the same gene class to see if they are far apart in the cluster tree by accident.

The list of interesting clusters was also tested for comparison of different clustering methods. A surprise from this comparison was that the method had picked out identical clusters from the results of different clustering methods and some clusters were identical in all of the three clustering results. Such observation increases the reliability of those clusters. This method adds to the repertoire of algorithms available for analysis of microarray data.

Results

Search of optimally correlating clusters

Preprocessing steps for gene expression data and the collection of gene classes were done as described in the methods section. Genes in the gene classes are classified to the same class on basis of common function, common localization etc. Average, complete, and Ward's clustering method were used as the clustering methods. Before visualization the resulting cluster trees were sorted for more optimal resulting order using method described in supplementary text [see additional file 10]. Sorting was done to keep the similar gene expression clusters closer to each other. The following text uses the term 'child cluster' on all the clusters that are part of the current cluster and 'parent cluster' on all the clusters that include the current cluster (see fig. 1).

After clustering the next step was to calculate log-p-value (equation 4 in supplementary text [additional file 10]) based on correlation measures for each cluster-class pair. This calculation was observed to be the bottleneck in the analysis and it was speeded by excluding clusters that included less than 5 genes. These small clusters were considered to be of little importance in the exploratory analysis step, but this depends on the size of the data set. The exclusion step filtered on average 60 % of the cluster nodes from the cluster tree. The biggest log-p-value (most significant positive cluster-class correlation) and the class from which it was obtained were stored for each cluster. For the excluded clusters log-p-value zero (p-value = 1) was given as the best score.

The searching method for previous best scoring clusters (BSC) takes a cluster tree and information on enrichment of each gene class in each cluster as an input (see fig 1.) and stores the most enriched functional class for each cluster (shown with a blue and red colour in the fig. 1). The method repeats a process where it:

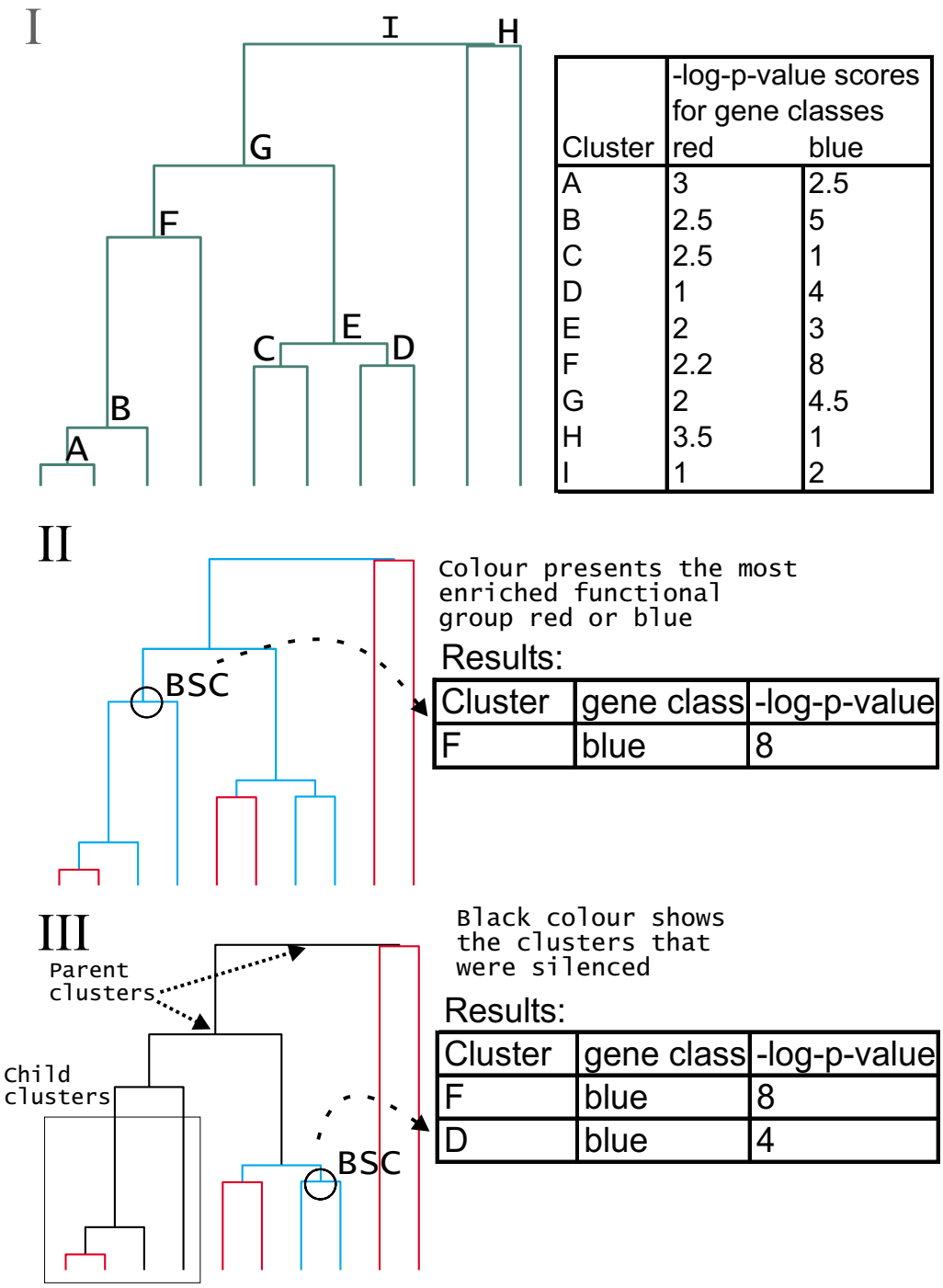
i) looks for the biggest correlation score from the cluster tree (BSC in fig 1.ii)

ii) includes it in the results

iii) replaces the value for highest scoring class with zero in BSC (result from first step) and in all of those child and parent clusters of BSC that had the same class with the best score (fig. 1.iii).

The last step is called 'silencing' and it is required to discard repeated hits to the same branch of cluster tree. At the same time also the other best-scoring classes that correlate 0.9 (calculated with equation 1) or more with the silenced class are also silenced among the BSC's child and parent clusters. Note that the silenced classes can still be reported in other parts of the cluster tree and other classes that did not have strong correlation with silenced class can be reported from the BSC's cluster tree branch. Method section explains more information on the used method.

As a condition for ending the search process a p-value limit for the observed correlation was looked for by calculating the Bonferroni-corrected log-p-value (p = 0.05) for both categorizations using the number of classes as a Bonferroni-correcting factor. This stopping rule reported 175 clusters with SGD classes and 116 clusters with MIPS classes from an average cluster tree (both shown in table



Cluster	-log-p-value scores for gene classes red	blue
A	3	2.5
B	2.5	5
C	2.5	1
D	1	4
E	2	3
F	2.2	8
G	2	4.5
H	3.5	1
I	1	2

Cluster	gene class	-log-p-value
F	blue	8

Cluster	gene class	-log-p-value
F	blue	8
D	blue	4

Figure 1

Principle of the searching method for Best Scoring Clusters I. Figure parts show the work flow of the search method. As an input, the method takes the hierarchical cluster tree and calculated correlation values for each cluster in the cluster tree and each functional class (red and blue). II. Cluster tree from part I is shown with colour visualization referring to the most enriched gene class (blue or red) in each cluster based on the table shown in the previous part. Colouring is done to a U-shaped profile created by two joined clusters that create each cluster. Here the best result is obtained from the cluster F marked BSC with the blue class. III. Black colour presents clusters that will be silenced (omitted from search) after BSC is located. Note that the small red cluster is not silenced as the blue functional class is not the best enriched functional class and that the blue class is not silenced from other parts of the cluster tree. After silencing, the results from clusters B, G and I will not be taken into consideration any more and so the next BSC will be cluster D.

Table 1: 50 best clusters from average method with MIPS classes

Ordinal number	Most enriched MIPS gene class	Log-p-value	Bonf. corr. log(p)	Observed	Clust. size	Class size
1	cytoplasmic ribosomes	102.5676	99.7780	78	159	108
2	respiration chain complexes	61.0379	58.2483	28	41	31
3	mitochondrion	57.3513	54.5617	75	142	305
4	AA metabolism	55.4905	52.7009	75	259	171
5	AA biosynthesis	55.4849	52.6953	59	244	98
6	mitochondrial ribosomes	46.2851	43.4955	32	101	46
7	rRNA transcription	33.5036	30.7140	46	274	99
8	cytoplasmic ribosomes	25.8871	23.0975	15	15	108
9	nucleosomal protein complex	25.2346	22.4450	8	8	8
10	26S proteasome	23.1373	20.3477	12	21	28
11	glycolysis and gluconeogenesis	21.1692	18.3796	10	13	28
12	cytoplasmic ribosome large subunit	20.5629	17.7734	11	12	63
13	purine ribonucleotide metabolism	20.4238	17.6342	9	9	33
14	mitochondrion	19.3933	16.6038	19	24	305
15	pher. resp., mating-type determination, sex-specific proteins	18.8355	16.0460	13	15	140
16	fungal cell differentiation	18.066	15.2764	17	20	318
17	phosphate metabolism	17.3187	14.5291	9	15	29
18	lipid, FA. and isoprenoid biosynthesis	16.8961	14.1065	14	32	88
19	cytoplasmic ribosome large subunit	16.6398	13.8502	9	10	63
20	cell cycle and DNA processing	16.6338	13.8442	51	153	493
21	26S proteasome	15.6397	12.8501	8	13	28
22	aminoacyl-tRNA-synthetases	14.2599	11.4660	7	10	28
23	nucleus	13.8439	11.0544	119	475	687
24	mitochondrial ribosomes	13.1286	10.3389	7	9	46
25	carbon compound and CH. transporters	12.373	9.5834	15	235	31
26	pher. resp.	12.1225	9.3329	6	8	32
27	carbon compound and CH. utilization	12.1195	9.3299	24	114	192
28	rRNA processing	12.0528	9.2632	12	59	60
29	metabolism of energy reserves	11.7646	8.9750	11	98	30
30	nitrogen and sulfur metabolism	11.5355	8.7459	6	7	49
31	unclassified proteins	11.3345	8.5449	34	79	656
32	alpha, alpha-trehalose-phosphate synthase	11.2183	8.4287	4	10	4
33	carbon compound, CH. transport	11.0861	8.2965	15	300	30
34	splicing	11.001	8.2114	25	378	74
35	heavy metal ion transporters	10.9979	8.2083	7	34	19
36	translation	10.8811	8.0915	6	8	50
37	cation transporters	10.5607	7.7711	11	75	48
38	F0/F1 ATP synthase (complex V)	10.1474	7.3579	4	5	12
39	lipid, FA. and isoprenoid metabolism	9.8666	7.0770	7	8	159
40	cytoplasmic ribosome large subunit	9.7246	6.9350	5	5	63
41	subcellular localization	9.51	6.7204	265	562	1885
42	homeostasis of metal ions	9.5024	6.7128	10	67	50
43	AA degradation (catabolism)	8.9569	6.1673	5	12	25
44	glycine decarboxylase	8.8106	6.0210	3	5	4
45	ribonucleoside-diphosphate reductase	8.8106	6.0210	3	5	4
46	ER	8.8061	6.0165	72	1613	131
47	mitochondrial inner membrane	8.6293	5.8397	5	5	103
48	cytoskeleton	8.3735	5.5839	6	11	81
49	ER lumen	7.9589	5.1694	4	18	9
50	energy	7.3865	4.5969	5	5	181

The table shows the ordinal number of the each obtained cluster shown in the figure 2, functional class that was most enriched in the cluster, and the $\log_{10}$-p-value for the most enriched class with and without Bonferroni-correction. Table shows also number of class members in the cluster (observed), size of the cluster (clust. size), and size of the functional class. The five most enriched functional classes can be found from table 4 [additional file 1]. Notice clusters that include only or almost only functional class members (8, 9, 12, 13, 15, 19 etc.) and the clusters that include the whole or almost the whole functional class (2, 9, 32 etc.). An especially good example is cluster 9 as it includes all the histone complex members and only histone complex members. Also it is worthwhile to notice the extremely small obtained p-values (for example, log-p-value 12 would refer to p-value 10^{-12}). Abbreviations are: AA, amino acid; metab, metabolism; pher. resp. pheromone response; CH, carbohydrate; FA, fatty-acid. More details of these results are shown in table 4 [Additional file: 1].

Table 2: Grouping of obtained functional classes

Colour in figures	Functional group (MIPS)	Member clusters from table 1	Functional group (SGD/GO)	Member clusters from additional file 1.
Blue	cytoplasmic ribosome proteins	1, 8, 12, 19, 40	cytoplasmic ribosomes	1, 11, 19, 27, 30
Red	respiration	2	energy (mitochondria)	8, 14, 15
	AA metabolism	4, 5	AA metabolism	7, 9
Red	mitochondria	3, 6, 14, 24, 38, 47	mitochondrial ribosomes	4, 6 22, 41
Blue	rRNA	7,28		
Blue			ribosome biogenesis	2, 3, 5
Blue	translation	36	translation	44
Green	nucleosomal complex	9	nucleosome	16
	proteasome	10, 21	proteasome	13, 28, 39
Green	nucleotide metabolism	13, 45	nucleotide metabolism	25
Green	cell differentiation, pher. resp.	15, 16, 26	mating response (shmoo tip)	49
	phosphate metabolism	17	acid phosphatase/vacuole	50
	lipid metabolism	18, 39	phospholipid metabolism	35, 37
			steroid metabolism	12
Green	cell cycle	20	cell cycle	18, 21
Blue	tRNA	22	tRNA ligase	33
Green			nucleus/RNA and nucleotide processing	10
Green	nucleus	23		
Red	CH metabolism, energy	11, 27, 29, 33, 50	energy (CH metabolism)	17, 31, 46, 47
	nitrogen and sulfur metabolism	30	sulfur metabolism	20, 34
			nitrogen starvation	36
			vitamin metabolism	38, 42, 48
	unclassified	31	unknown	40
Red	CH. transport	25, 33	CH transport	43
Blue	mRNA processing	34		
Red	cation transport and homeostasis	35, 37, 42	kation transport	24, 26, 29, 32
	cytoplasm	41		
	AA degradation (/cell wall)	43, 44		
	ER	46, 49	ER	23
	cytoskeleton	48		

Different functional class groups obtained with both MIPS and SGD classification using the first 50 clusters. Groups are rough approximate categorizations giving an overview of what functional classes were obtained showing the ordinal numbers for the member clusters. The matching groups for MIPS and SGD classifications are shown on the same line and similar groups are placed to near by rows. The first column shows the colour of the clusters in cluster tree visualizations when it is not black. New abbreviation is ER, endoplasmic reticulum. Other abbreviations are explained in the text for table 1. Functional grouping is based mainly on table 4 [see additional file 1].

4 [additional file 1]). The results are given in their estimated order of goodness (their order according to their log-p-values) so new p-value limits can be set. A big number of obtained clusters might be a result of too strong requirement for correlation between two functional classes when silencing them (limiting value of 0.9), which probably allows reporting of similar functional classes from the same tree branch. It can also simply signal multiple strong correlations between the data set and gene classes pointing to the vast amount of information in the data set.

Results from the search of optimally correlating clusters

The first 50 clusters, resulting from the mainly analyzed average method with MIPS categorizations are shown in table 1 with their most enriched gene class. For a better view on associated gene functions, additional files show the five most enriched gene classes for each reported cluster. The most enriched classes include protein complexes, cellular components and functional classes showing that all the used classifications were among the most enriched classes. Additional files show all the obtained clusters for each method. One small reported cluster worthwhile mentioning is cluster 9 in MIPS results (table 1). This cluster includes all 8 histone proteins and no other proteins. An identical cluster to it was also obtained from two other clustering methods and also when using SGD classification. Indeed the strong co-regulation of histones has been reported earlier [1] but not in the analysis of this data set. Other similar findings are: cluster 2 that includes 28 out of 31 respiratory complex genes; clusters 8, 12 and 19 including either exclusively or almost exclusively ribos-

omal proteins; and cluster 13 that includes exclusively nucleotide metabolism genes.

A more detailed analysis of the associated functions with the selected clusters showed that there was redundancy among the associated functions. To highlight this, a rough manual grouping of these obtained functional classes shown in table 2 was done. Grouping gives a broad overview of obtained functions among the clusters shown in table 1. Also three even larger functional class groups were created for visualization purposes. These are explained in text for table 2. Clusters in table 1 are also visualized with the cluster tree in fig. 2 (also shown in additional file 4.). BSCs are marked in figure 2 by colouring the line formed by the two joined clusters that created each BSC. Clusters in the cluster tree and in the tables are linked to each other by the ordinal numbers from table 1. Visualizations were done with both the MIPS classification (fig. 2) and SGD classification (fig. 3 and additional file 5.) but after observing that the results were in most parts very similar, analysis concentrated on MIPS categorization. Similarity of the results from two classifications can be also seen in the results of other methods (figures 8 – 13 in additional file 6, 7, 8, 9). All the visualizations show the top 50 clusters.

Using information from table 1 and table 4 [additional file 1] allows the analysis of the distribution of functional classes in the cluster tree in fig. 2. Starting from the upper region the first reported cluster is phosphate metabolism connected cluster 17. In the results with SGD classification (table 4 in additional file 1 and fig. 3), the same cluster (shown with number 50) is associated with acid phosphatase and vacuole. The next reported area down is the area of MIPS cluster 20 with the SGD result clusters 18 and 21 at the same area. MIPS cluster 20 and SGD cluster 18 are associated with cell cycle and SGD cluster 21 with DNA replication. Even though the two database classifications do not give exactly the same results they still point to the same functionality in the same cluster tree branch. Cluster 41 was associated with cytoplasm localized genes giving little information on function but besides this the bigger child cluster (cluster 46 in SGD results) had carbohydrate catabolism groups in second, third, and fourth position when the enriched gene classes were sorted with log-p-values. This demonstrates the benefits of analyzing also other gene classes from a sorted list of enriched gene classes. The rest of the tree includes clear analysis when looking at larger clusters like clusters 23 and 34 associated with nucleus (and mRNA splicing) and clusters 4 and 5 associated with aminoacid metabolism (7 and 9 in SGD results). Clustering of protein synthesis associated genes is shown by cluster 1 associated with cytoplasmic ribosomes and cluster 7 associated with rRNA (clusters 1, 2, 3 and 5 SGD results). Energy associated clusters were presented in neighbouring branches. Clusters 3 and 6 enrich mitochondrial ribosomal proteins and mitochondrial genes (4 and 6 in SGD results), the second cluster 2 represents respiratory chain complex (8 in SGD results) and third C-compound metabolism and energy reserve metabolism are enriched in cluster 27 and 29 (47 and 31 in SGD results). Also the respiratory-associated cation transporters that were reported in the original article [5] are present in the clusters 37 and 42 next to cluster 3 and in addition the clusters enriching carbohydrate transport functions (25 and 33 in MIPS results and 43 in SGD results) are next to energy associated clusters. As an overall view, the energy group clusters shown in red are placed into the lower area and the protein synthesis clusters shown in blue are in the middle area of the tree. The green group included cell cycle, cell differentiation, nucleus and nucleosomal complex (see table 2). This is more functionally heterogeneous than the other two and it is also equally more scattered in the cluster tree. Note that the energy clusters 11, 14 and 24 (17, 22 and 41 in SGD results) separate from the rest of the energy clusters and similarly protein synthesis clusters 19 and 36 (27, 30, 44 in SGD results) separate apart from most of the protein synthesis clusters and would have easily gone unnoticed in a manual analysis. Similar colouring is used also in the cluster trees shown in figures 10 – 13 [see additional files additional file 6, 7, 8, 9] from other clustering methods.

Table 3 compares results of the method presented here to a manual analysis [5]. Comparison shows only the results from the analysis of average clustering results with both MIPS and SGD classes against all the reported gene clusters from the earlier analysis of the data set where few clusterings with different cut-off settings were done. The results obtain most of the previously reported functional classes. In addition to analyzing all the BSCs at the same time it is also informative to concentrate on one group of clusters enriching functionally connected gene classes to see how they have been placed in the cluster tree. As an example protein synthesis-associated clusters are shown on top of the cluster tree in fig. 4. Visualization shows two clusters 12 and 19 that are associated with the same gene class (ribosomal protein complex, large sub-unit). This raises the question whether the clusters form actually one cluster that is accidently split in clustering procedure or if they are really two differently regulated clusters. In order to answer this, the expression profiles of the genes in the clusters were plotted in fig. 5. showing that the clusters have very different expression profiles.

Analysis of results using randomization

An important issue in the analysis is that the reported p-values do not give the exact probabilities for the observed enrichment as the parallel monitoring of several gene classes and several clusters at the same time causes the

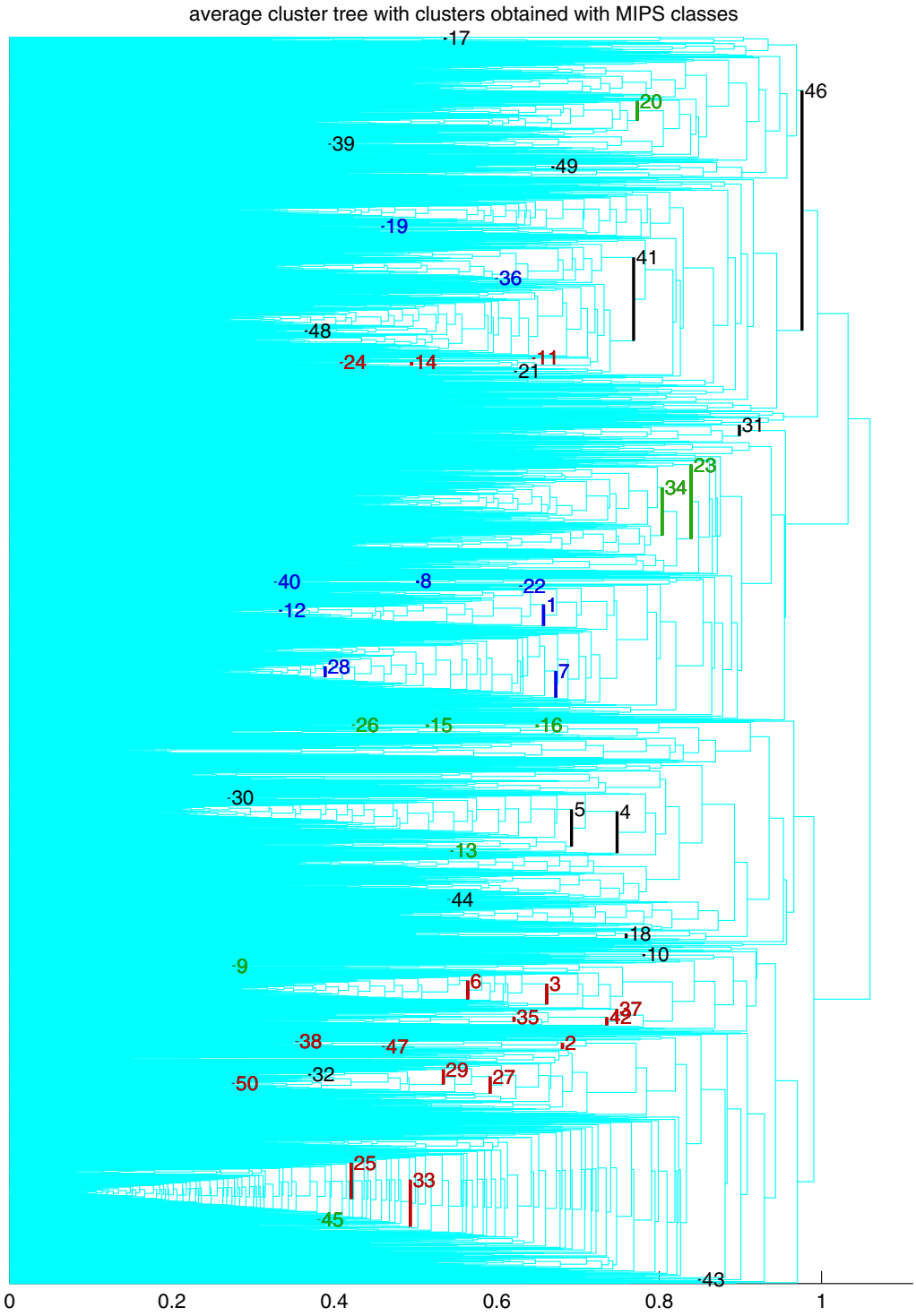


Figure 2

BSCs selected from average clustering with MIPS classification showing an overview of enriched gene classes. Average methods cluster tree with the 50 top BSC clusters shown in table 1. Clusters are linked to table 1 by the ordinal numbers in table 1. Three coloured cluster groups are: blue, protein synthesis; red, energy and carbohydrate metabolism; green, cell cycle, differentiation and growth, nucleus, chromosome structure and mRNA processing. The rest of the clusters are black. Coloured groups are shown in detail in table 2 (MIPS classes). The figure is shown also as a postscript file for separate zooming and printing in figure 8 [additional file 4]. The tree is analyzed more in detail in the text.

Figure 3

BSCs selected from average clustering with SGD classification showing an overview of enriched gene classes. Average methods cluster tree with the 50 top BSC clusters obtained with SGD classes are presented for comparison with fig. 2. Clusters are linked to table 4 [additional file 1] by the ordinal numbers in the data file. Three coloured cluster groups are the same as in figure 2 and they are shown in detail in table 2 (SGD/GO classes). Notice that although the clusters are placed differently than in fig. 2, similar functional classes are often associated to the same cluster tree branches. The tree is analyzed more in detail in the text. The figure is shown also as a postscript file for separate zooming and printing in figure 9 [see additional file 5].

Table 3: Correlation of results with previous analysis

Hughes et al. Results	Corresponding results with	
	MIPS	SGD
groups in fig. 1B		
PAU gene family		
RNR2,3,4		
stress and CH metabolism (both)	29,32	47
AA biosynthesis	4,5	7,9
PKC/Calcineurin		
Ergosterol		12
mitochondrial function	3,24, 38	6,8,14
Mating	15, 17, 26	[71]
additional groups from text		
cell wall	43	[88]
mitochondrial ribosome	14,24	3,6,22
iron/cation homeostasis	35, 37, 42	24,26,32
protein synthesis	1, 8, 12	1,2, 3, 5

Comparison of obtained results to ones observed by Hughes et al. [5]. The comparison uses groups that Hughes et al. presented in their article's figure 1 and the additional ones that were only mentioned in the text. The table presents the easily spotted corresponding clusters obtained with both SGD and MIPS classification reporting the ordinal numbers for the corresponding clusters. Brackets [] point to clusters that are outside the top 50 clusters or present weaker correlation to previous results. Results were obtained with text search from table 4 [additional file 1]. Note that all the analysis results from the previous analysis work from several gene clustering results with different cut-offs were pooled and compared against one cluster tree showing that similar clusters to earlier results can be obtained. Additional abbreviation is PKC responsive to protein kinase C. Other abbreviations are explained in the texts for tables 1 and 2.

probabilities to be bigger than what is reported. To estimate the real probabilities, the maximum log-p-values from all the clusters were analyzed using randomized connections between the genes in the cluster data set and the genes in class data set to see if similar results could be obtained. Only five different randomizations with each cluster tree were tested as each of the cluster trees already included over 5000 measurements. A distribution of results from one randomization is shown in fig. 6A. Randomizations were different for each clustering method.

Randomizations were kept separate to see how similar results were between the randomizations. From each of the randomizations different percentiles were monitored (see fig 6B). Reported percentiles (and also maximum values from each distribution) were much smaller than the log-p-values analyzed from the real data pointing that the observed log-p-values reported from data set must really come from true correlation between classes and clusters.

It was observed that the different clustering methods produce very similar results in the random analysis. The only difference that seems significant is the difference between two classifications, SGD and MIPS. This encouraged testing to see if the results from two classifications could be brought on the same level using Bonferroni-corrections. The resulting 99th percentile values (equal to p-value 0.01) are shown in fig. 6C showing that it produces satisfactory correction to the differences between two classifications and that the resulting log-p-values are on average 0.8 (p-value ~ 0.15 when p-value = 0.01 should equal here to log-p-value 2). So using Bonferroni-correction gives a larger probability than the real probability values making the used stopping criteria (Bonferroni corrected p = 0.05) in the BSC searching quite conservative.

Comparison of results from different clustering methods

With the randomized results from different methods there was little overall difference in the percentiles between the different methods when the same classification is used. There was also similarities in the obtained functions included in the gene lists indicating that the different clustering methods probably found the same local regions enriching the same gene class. These observations encouraged the cluster level comparisons of the results between the different clustering methods using the selected clusters.

A rough comparison was based on the selection of the correlating cluster pairs (using eq. 5.) from the results of two

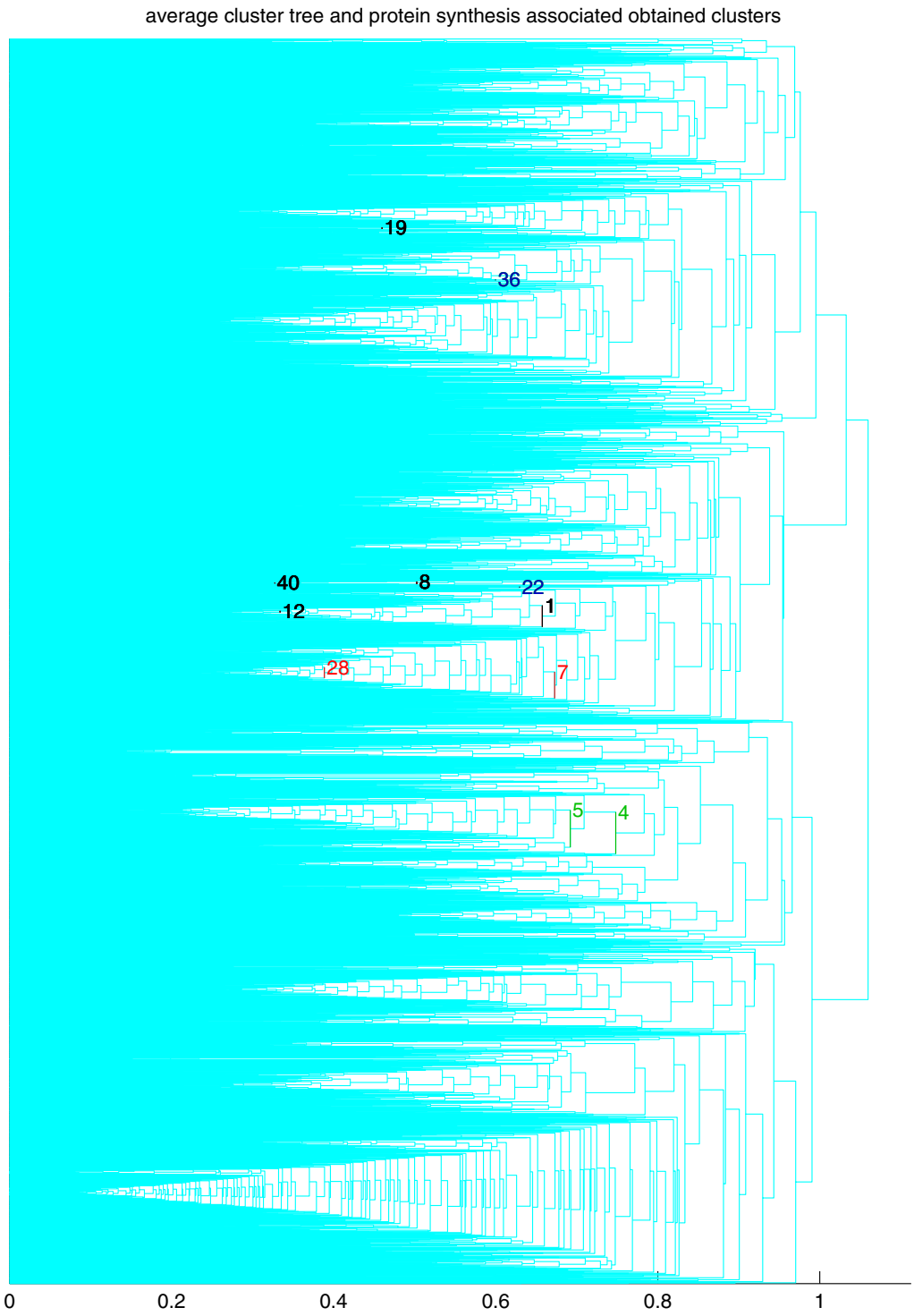


Figure 4

Visualization of the protein synthesis associated MIPS clusters MIPS clusters associated with protein synthesis are visualized. This enables the analysis of heterogeneity of the group of functional classes in question. Colouring of the clusters was also used marking cytoplasmic ribosomal protein clusters (1,8,12,19,40) with black, ribosomal biogenesis clusters (7,8) with red and clusters of translation genes (36) and tRNA genes (22) with blue. Additionally, the amino acid synthesis clusters are presented with green colour. Note the two clusters, 12 and 19, that are placed in separate branches of the cluster tree although they are associated with the same protein complex in table 1.

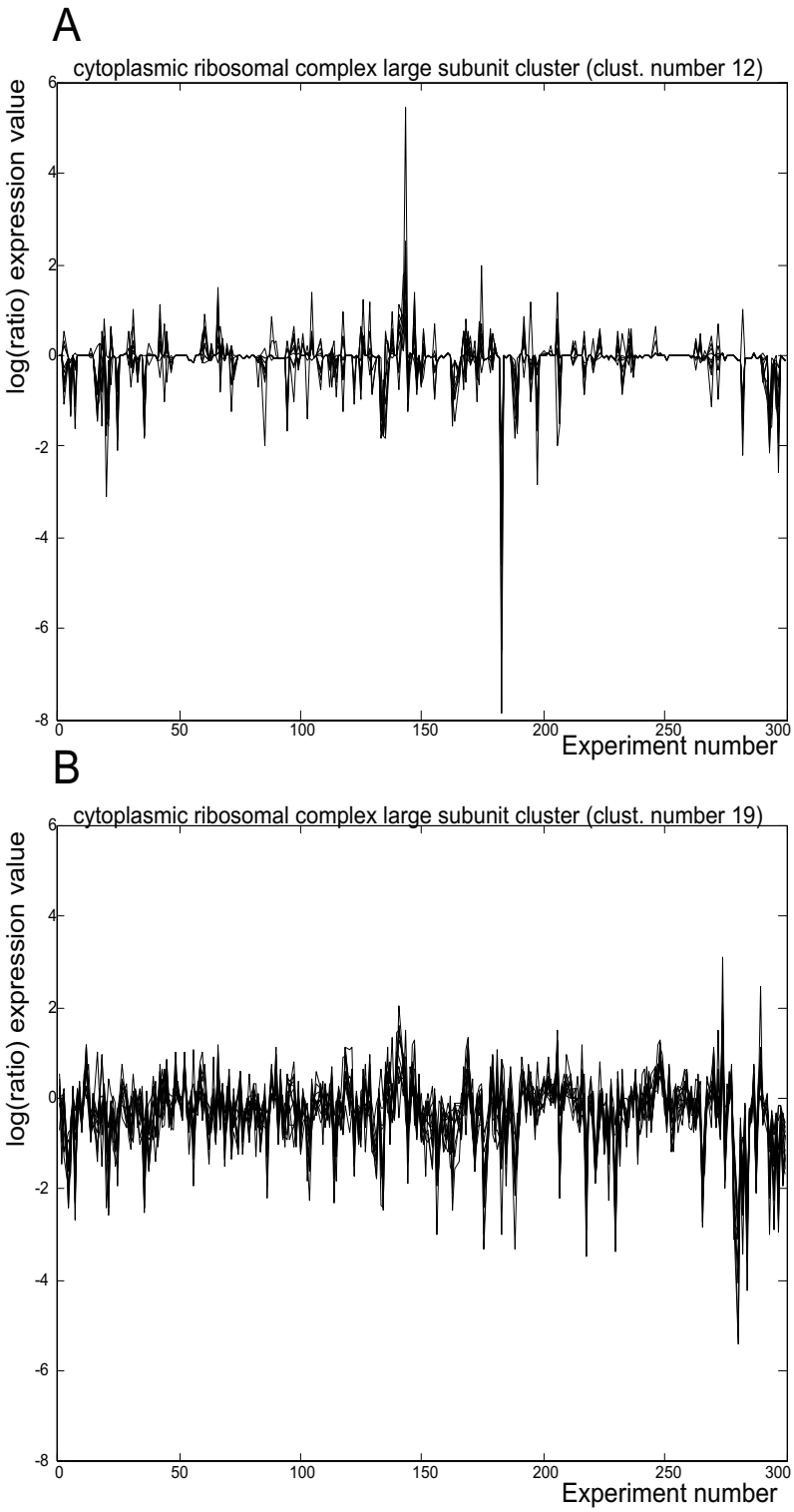


Figure 5

Expression profiles for two clusters enriching ribosomal complex large subunit genes A. Expression profiles for genes from the cluster 19 are shown. Y-axis shows the log(ratio-value) and X-axis presents the treatments in their original order. B. Expression profiles for genes from the cluster 12 are shown. Axes are similar to previous picture. Notice the strong difference in overall expression profiles as the cluster 12 has more regulation over all treatments whereas the cluster 19 has only few treatments showing regulation.

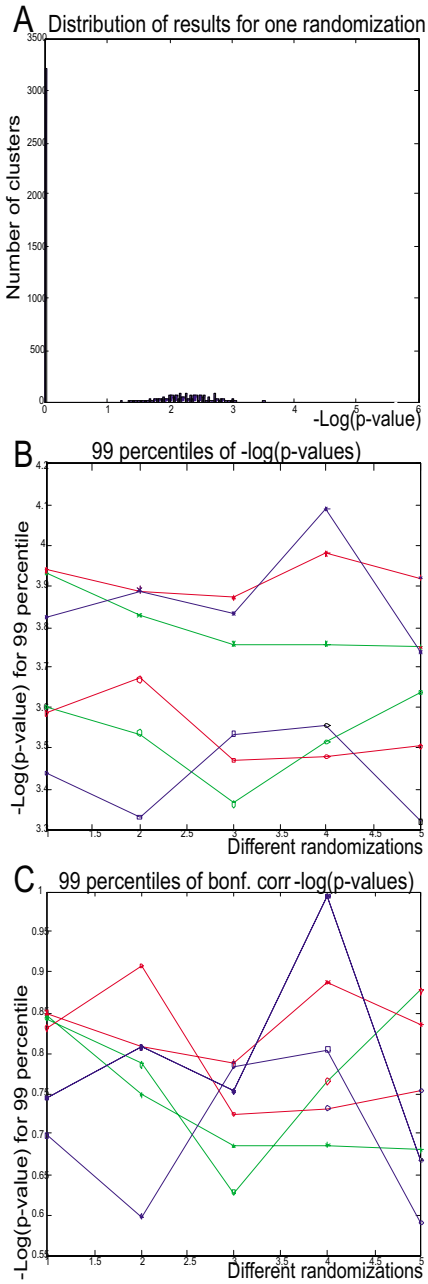


Figure 6

Randomization results. A. Histogram of different log-p-value results for the highest scoring cluster-class pair in each of the clusters from one randomization for average results with MIPS classes are shown. A peak of zero results from clusters with too small size as they are given the value 0 automatically (see results). B. 99th percentile log-p-values from different randomizations for all methods (red = complete, green = Ward, blue = average) and for both the SGD (marked with '*', three higher profiles in the plot) and MIPS (marked with 'o', three lower profiles in the plot) classifications. Note that the correct value for 99th percentile would be 2. C. Previous 99th percentile values after Bonferroni-correction. Different results are marked similarly as before.

methods for the comparison of log-p-values. Comparisons were done with both the MIPS and SGD results. Resulting-lμ(p-values) are compared in the scatter plots in fig. 7. Although the results show many almost identical clusters between all the methods, there seems to be a trend among the clusters with stronger enrichment that shows bigger negative log-p-values for average linkage and for Ward's method when compared with complete linkage. On the other hand, Ward's method and average show little difference when compared.

Although the first idea was to see if the clustering methods could be sorted according to their performance, a detailed analysis showed that in the comparison of the methods, there are clusters that were found with stronger enrichment from the 'less well' performing method. This raised the question of whether a combination of the results selecting or reporting the cluster that performed better would be possible, enabling a parallel analysis of a couple of clustering results from different clustering methods. The combination of the results was done simply by taking the mainly analyzed results from the average method and adding a note referring to another methods gene list cluster number whenever another method showed a better enrichment with a correlating cluster (see table 4 [additional file 1], columns A and V).

An unexpected result was observed when analyzing the correlating clusters between the gene lists. The compared gene lists had identical clusters and some of them were identical in each of the three gene lists as for example cluster 9 (cluster of histone proteins) from the results with MIPS classification. These clusters were also marked with a note (see table 4 [additional file 1], columns B and W) as there is a smaller probability that they are clustering artefacts. Also 'weaker' correlations like the ones bigger than 0.9 could be used similarly to look for clusters with strong similarity in the clustering results.

Discussion

This article presents a method to analyze hierarchical cluster trees resulting in candidate clusters for analysis. Clusters obtained here presented very strong correlations with the gene classes with many cases where almost all of the functional class was included in the cluster and/or most or all of the members of the cluster belonged to the reported gene class. The obtained clusters enable an overview of the distributions of functional classes, similar to work shown in [6], with the difference that the reported functional classes are automatically selected on the basis of statistical significance of their enrichment. Visualization also shows that the obtained clusters could not be found by simply cutting the cluster tree. Results showed many gene classes found in the original analysis and also many classes that were not reported in the earlier analysis. We omitted the

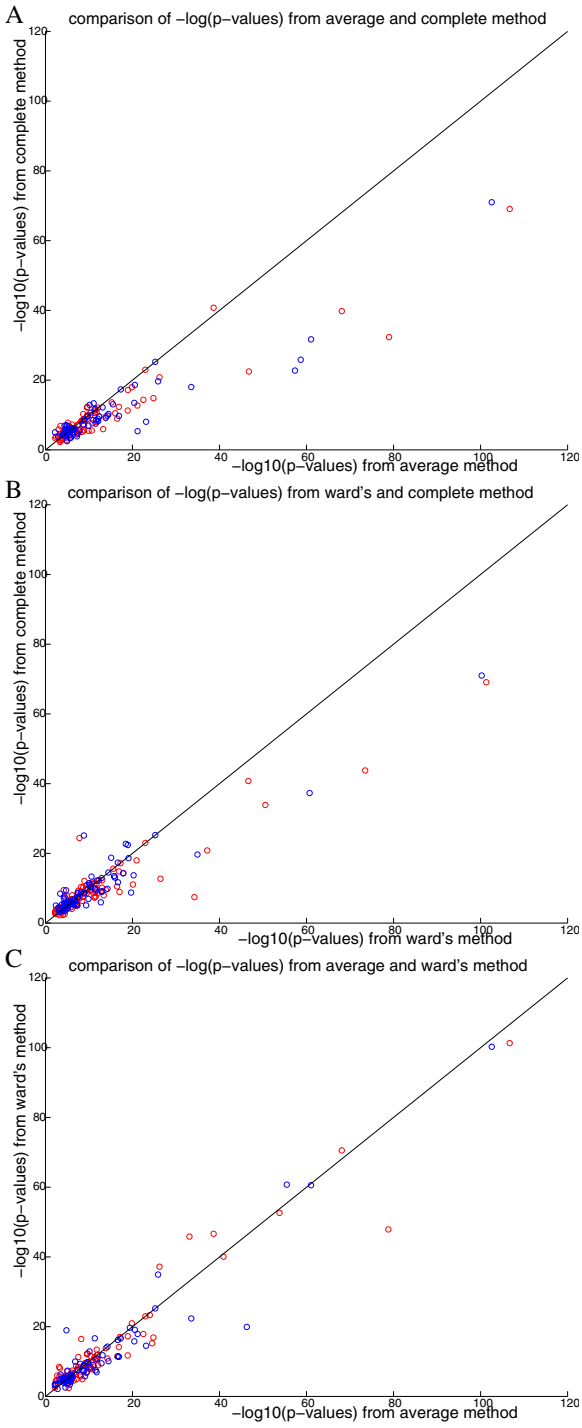


Figure 7

Comparison of cluster trees using correlating clusters. A. Comparison of negative $\log_{10}$-p-values for correlating clusters between average (X-axis) and complete linkage (Y-axis) cluster trees. The figure also includes x = y line to ease the comparison. Notice that although many clusters show quite similar results the clusters with bigger negative log-p-values show bigger log-p-values in average linkage method. Red circles show comparison of results obtained with SGD classes and the blue circles show the same comparison obtained with MIPS classes. B. Comparison of negative $\log_{10}$-p-values for correlating clusters between Ward's method (X-axis) and complete linkage (Y-axis) cluster trees. Note the similarity to the scatter in part A. C. Comparison of negative $\log_{10}$-p-values for correlating clusters between average linkage (X-axis) and Ward's method (Y-axis) cluster trees. Here the scatter differs from other scatters (A and B) with clusters showing more scatter along the x = y line.

groups from fig. [2] from Hughes et al. [5] from comparison as the groups included the clusters obtained from the control data set. Still the results even found a correlation to phosphate metabolism cluster that was in previous figure [2] even though the data set was different.

The results presented the log-p-values for class enrichment in clusters. The drawback in the calculated log-p-values was that they do not take into account the multiple classes and clusters present during the testing. Therefore the stopping criteria for the search of BSCs used Bonferroni-corrected log-p-values with the number of classes for correction. Actually the Bonferroni-correction should have taken the number of clusters also into account but the results showed that already the number of classes corrected the log-p-values more than was needed, making the stopping criteria quite conservative. This was expected as the Bonferroni-correction is optimal for classes with no correlation between them, whereas here, the classes have strong correlation structure explaining a too strong correction. Bonferroni correction does not affect the order of clusters found.

Clustering and visualization methods are forced to do some damage to the relationships in the original data. Therefore the analysis of the results should pay attention to the reliability of the obtained results. This was demonstrated with a functional class associated with cytoplasmic ribosomal proteins where visual analysis was done to see if one functional class was split by accident into two clusters. Expression profiles showed very different regulation pointing to possibility that the functional class includes differently regulated sub-parts. The other similar case was observed with proteasome complex genes.

The method relies on the gene classes used so their selection is important. Also the used gene classes should be based on observations other than gene expression as this would cause direct dependencies between the class data set and the expression data set (circular logic). The addition of gene classes that end up performing badly does not effect the performance of the method as it only monitors the best performing gene classes. Here the functional categories have been used to select the clusters. As the quality of the expression data has been often questioned, one could use functional categories as a quality control of the expression data. The reliability of individual gene expression profiles increases if the genes with similar functionality present similar expression profiles, as they support the assumption that the observed regulation is real and not caused by measurement error. Still these methods are unable to detect such cases where only one rate-limiting enzyme is regulated from the pathway and therefore this does not totally replace manual analysis.

A custom created ontology or categorization [7] or Swiss-Prot key words (Lahteinen et al., unpublished results; Knuuttila et al., in press) could be used with or instead of the database categories. Also the strongest correlation with up-stream sequence patterns from genomic sequence or the correlation with protein-protein interaction data clusters could be analyzed similarly. Moreover the protein sequence clusters could be looked for correlation with functional groups or protein folds. Disease sample expression profiles such as cancer tissue samples could be analyzed with the presented method using the associated information from sample medical records to guide the analysis. Note also that the presented correlation measure enables a simultaneous analysis of significant positive correlation (class enrichment) and significant negative correlation (class under-representation) as is shown in fig. [3] in supplementary text [see additional file 10]. Although negative correlation was not analyzed here it could reveal important details in some situations.

Three different hierarchical clustering methods were used here in parallel to produce different cluster trees and these were compared with each other. The aim was to test selected BSCs for comparison and combination of the clustering results by looking for correlating clusters between different clustering results. A similar comparison is often done manually by analyzing the obtained gene lists to see what results seem most biologically reasonable. Gibbons et al. [8] and Oja et al.[9] also proposed similar ideas for comparison of clustering results, different preprocessing steps, and metrics. The obtained results show that all the methods have mainly almost equally well performing clusters. Still the differences are strongest in the comparison of the complete linkage and average linkage method and almost equally strong between complete linkage and Ward's method whereas Ward's method and the average linkage seem to produce equally good results.

The reported results propose a joint analysis concentrating on the Ward's method and on the average but the work here concentrated on the average in order to save space. One way to perform a joint analysis of two cluster trees would be to combine the results using the information on correlating clusters. Here every cluster in the analyzed results of the average method was flagged when a correlating cluster with a more significant result was found. This enables the selection of better performing cluster for the analysis from the other methods results when such cluster is available. The presented combination was similar to parallel analysis that was presented by Wu et al. [4] for classification purposes. An added plus, with this kind of combination, would be that the analysis could actually select the less well performing method if it has created a better correlation with those gene classes that are the most interesting. A similar correction could be done by select-

ing only the classes of most interest as the input for the selection of clusters.

The performed comparison was not obviously optimal as it concentrated on the performance with similar clusters found by two compared methods. A more thorough comparison should also include an analysis of gene classes that were discovered by one method but not by the other. Also, another issue that was left out from this analysis was the level of variation that could be seen in between the results of two equally well performing clustering methods or when the same method is used on the same data set with noise added to the data. These issues were considered to be outside this article's scope and the aim is to emphasize the idea of cluster level comparison and combination of the results presented here showing one further application for BSC lists.

It would have been interesting to compare similarly other different clustering methods than just hierarchical methods. This would have been questionable as a hierarchical clustering includes more cluster candidates to choose from than for example a k-means clustering and so the probability of observing strong gene class correlation in a cluster tree would have been bigger than in k-means clusters. Still we have done similar work with Self-Organizing Maps (Knuuttila et al., in press) where several clusters obtained with SOM were similarly compared using log-p-values. These methods are under further development.

Even though some of the uses for BSC lists have been already presented, there are options for further analysis. The literature reports several methods for classification (and for semi-supervised clustering) that need a starting point similarly to the k-means algorithm in clustering (See [10][11] for example). Normally one selects the whole group of class members for definition of the starting point but this would fail if the class forms few separate clusters in the data set. One could consider the observed BSCs as a better starting point in the classification analysis. For example, assuming that the gaussian distribution would be suitable for modelling the cluster, one could start the modelling of the distribution of genes that belong to a certain class using mean and variance of the obtained cluster. After this, some local optimum searching algorithm can be used to optimize the cluster in question for the classification of the most enriched class in the starting cluster. This could create a useful combination with a hierarchical clustering as some of the possible errors created by the clustering procedure could be corrected at the later step. Also, even though the classification method would not need a starting point, the information of the enriched gene classes could be used to select classes taken to the classification. This could save enormous efforts in testing with hundreds of gene classes.

Conclusions

The article presented a method for analysis of hierarchical clustering that looks for clusters showing strong correlation with functional gene classes and displays it's power in the exploratory stage of the analysis. The results showed functional gene groups that were obtained earlier in manual analysis of the data and also many gene groups that were not found in manual analysis. The method was tested with randomization showing that the observed results could not have been obtained by randomized classes. The method should be applicable also to other data mining problems having big data sets and categorical data available.

Methods

Pretreatment of the gene expression data

The whole expression data set was first downloaded from Rosetta Inpharmaceutics web site http://www.rii.com. The data set represented expression of over 6000 yeast genes in 300 mutated yeast strains [5]. The actual data sets included a $\log_{10}$-value of ratio of expression genes in a mutated strain vs. a normal control strain. The data set also included the estimated error for $\log_{10}$(ratio-value) and p-value for observed expression by a random process. The data included two separate files of measurements. Both had the same actual log-ratio expression measurements but the error and probability estimates were different. The bigger of the two error estimates was always selected similarly to original analysis [5]. Besides the main data set, a similar separate control data set with over 60 control vs. control measurements for the same genes was also downloaded.

Log-ratio measurements in the data set were scaled with their selected error estimates similarly to the original analysis and to other previous work [9]. In some cases, error might have been approaching infinity or was reported as Not-a-Number (NaN), the actual log(ratio-value) was missing or was reported as NaN and/or the p-value for log(ratio-value) was estimated to be 1. In all of these cases the error-normalized log(ratio-value) was replaced with zero. Replacement with zero is based on the idea that the NaN values have massive error resulting to a value close to 0 after error scaling.

Pearson correlation was selected as the distance measure similar to the original analysis. Use of pearson correlation requires the filtering of datapoints with too small variance. Here two unique steps were done for filtering. First the amount of signal (variance of expression measurements) was compared to combined variance of signal and noise (error in expression measurements) for every gene. Second, the obtained results were compared between mutation data set and two control data sets. Control data sets were used as negative controls. As a cut-off for the low

variance genes the bigger value for 95th percentile limit from two control data sets was selected (see supplementary text [additional file 10] for detailed text and figure).

Collection of gene classes

Two yeast databases were selected (MIPS and SGD) and used in parallel. Functional classification, complex classification and component classification were downloaded from MIPS and biological process classification, molecular function classification, and cellular component classification from SGD. A search obtained 616 gene classes from the MIPS database. As the SGD database uses the massive Gene Ontology (GO) classification only those GO classes were selected to analysis that included more than 4 genes from the expression data set resulting in all together 1321 gene classes from SGD.

The resulting classes create separate binary categorizations with complex correlations as genes usually belong to many of these classes. Classes also create hierarchical tree structures (MIPS) or directed acyclic graphs (SGD) where smaller classes with more detailed information are nested inside bigger broader classes. All of these levels of functional classifications were used during the analysis. MIPS database also included genes that were classified to functional classes on basis of sequence similarities. These were classified to groups of strong similarity, similarity, and weak similarity. Only the ones with strong similarity were included to categories.

Measure for correlation between a cluster and a gene class

Here the comparison of clusters was based on the selected gene classes that clusters are expected to group together. A cluster was considered to be worth selecting when it grouped together unexpected number of members of some gene class. Therefore a measure that takes into account the correlation between cluster members and class members was needed. The measure should be independent of gene class size and cluster size as the data set includes clusters and gene classes of varying size. Also the probability of observing a similar enrichment as a result of random clustering should be taken into account. The aim was to use a measure that can monitor both significant increase (class is under-represented, cluster and class have negative correlation) and significant decrease (class has been enriched to cluster, cluster and class have positive correlation) of class members. Fisher exact test (see Additional file: 10) is often used to test this.

Fisher exact test was modified in two ways: Instead of the original p-value a logarithm of the p-value was taken to highlight the differences between very small p-values. Also the cases of positive and negative correlation were separated from each other with a definition of a sign according to whether correlation of cluster and class were either negative or positive. Resulting log-p-value was defined as -log(p-value) when the correlation was positive (resulting in a positive value) and as log(p-value) when the correlation was negative (resulting in negative value). This measure and it's development are described more in detail in the supplementary text [additional file 10]. Figure [3] in supplementary text shows that measure is capable of reporting simultaneously both significant negative and significant positive correlation. Still in the current analysis, only the positive correlation was taken into account, so the method is here identical to one used in the literature before (see for example [4]).

Method for finding optimally class-enriching clusters

The main idea for enhancing the analysis of cluster tree is that at some point in each branch of the cluster tree, there is a cluster that represents an optimal enrichment of the gene class that is most enriched in that part of the cluster tree. As one expects this cluster to have the best score with the used correlation measure, it is called the Best Scoring Cluster (BSC). Child nodes that are part of this BSC represent clusters that are not big enough to be optimal. Their probability from random clustering of gene classes is bigger and their analysis would result in a too complicated image of the function of the enriched gene class in the data set. BSC has also parental clusters that include the optimal cluster. These will include also genes around the optimal cluster and can therefore cause erroneous conclusions. Also, if the clusters are used as a basis of classification of genes in the style of guilt-by-association methods, too small clusters might miss some of the genes that should be classified into the class in question. Too big clusters would on the other hand include too many genes that should have been excluded.

The method used first calculates a correlation measure for each class in each cluster. After this the method concentrates on the class with the best score in each cluster. From these results the cluster with the best score is selected. Other well scoring clusters with almost identical results can be usually found among the child-nodes and parent-nodes of this BSC but these are not considered any more interesting as the cluster with the best result already captures the correlation with the functional class in question. Therefore, BSC search method replaces the value of the best scoring functional class with zero (log-p-value zero equals p-value 1) from all the parental and child nodes of the BSC that had the same best scoring functional class as in the already selected cluster (see fig. 1). This is called silencing. Next, the BSC search method looks for new optimal cluster with the highest result among highest scoring functional classes of clusters. As a result of the silencing step where the method placed the minimum value to previous found optimum cluster and to its parent clusters and child clusters one should not get the same

cluster branch again. The only exception is when the cluster branch includes varying enriched functional classes as the method still takes all those clusters among previous parental and child clusters that had a different best scoring functional class than the one in the BSC normally into account. When the new BSC is found method again replaces the values for found class in the dendrogram branch with zero as before and starts the search again.

After testing the method by filtering only one functional class from the branch where a BSC was located, it was observed that in several cases the method obtained results with very similar classes from clusters that were close to each other in the cluster hierarchy. These classes were considered different although there might be a very strong similarity in between them. Therefore the simultaneous filtering of strongly correlating gene classes was added to the method. Correlation was calculated between all the gene class pairs as a correlation of Bernoulli distributions:

$$Corr(A,B) = \frac{Cov(A,B)}{D(A)D(B)} = \frac{E(AB) - E(A)E(B)}{\sqrt{D^2(A)D^2(B)}} = \frac{P(A = 1 \cap B = 1) - p_A p_B}{\sqrt{p_A(1-p_A)p_B(1-p_B)}} \qquad (\text{eq.}1)$$

Here A and B refer to different functional classifications being analyzed and p_A refers to P(A = 1) (meaning the probability that class A equals 1), p_B to P(B = 1) and P(A = 1 ∩ B = 1) refers to probability that both A and B are 1. This measure corresponds to how well the membership of another gene class can be predicted on the basis of the other. Now every time the method filtered a gene class from a certain branch of the hierarchical tree, it also filtered all the other functional classes with correlation more than or same as 0.9 with the filtered gene class.

The search process for BSC is repeated until one of the conditions for quitting is fulfilled. Two conditions have been used in tests for ending the cycle. One is a threshold number of selected clusters and the other is too weak best correlation to be considered significant. The latter was thought to be more suitable as it does not limit the number of selected clusters but only requires that the probability of obtaining a resulting enrichment from a random cluster should be small. The cutoff value can be estimated for example using Bonferroni correction with the p-values similarly to the analysis.

Comparison of results from different cluster trees

On the basis of the observed similar results from randomization with different methods, it seemed logical to compare the optimal clusters from different methods. The idea behind the comparison is that the list of found clusters presents the most important details from the cluster tree. It was also observed that the lists of optimal clusters are in most parts pretty similar (meaning that one observes the same enriched functional classes approximately in the same order in the lists). It is also simpler to compare a hundred clusters to another hundred clusters than the comparison of the whole cluster trees with thousands of clusters to each other and therefore the benefits of simplification are considered larger than the drawbacks.

The comparison was done between two clustering results by comparing the log-p-values for the most enriched gene class in clusters in pairs. Cluster pairs were selected on basis of the correlation of cluster members between the clusters using equation 1 to calculate how much clusters from different methods correlate. The obtained correlation tells the probability that the member of one cluster belongs also to another cluster and the non-member of one cluster is also similarly excluded from the other cluster and it should not be mixed with the correlation of expression profiles. The resulting correlations were calculated into a matrix format with one list of clusters presented by columns and the other by rows.

After this procedure, clusters from the first list that correlate most with clusters from the second list and the same in opposite (looking clusters from second list that that correlate most with clusters from the first list) were looked for. If the same pair was found when the matrix was analyzed both column and row-wise that pair was included in the analysis. So these are simply the correlation matrix elements that are maximum values for both their row and their column. Next, the log-p-values of the reported best scoring functional classes were compared between these two correlating clusters by visualizing the obtained results as is shown in fig. 7 to see on which side of the x = y line the results lay.

Additional material

Additional File 1

All the BSCs found from average cluster tree with SGD and MIPS classes. Table 4 presents all the BSCs found from average cluster tree in a table format. The table presents separately the results obtained with SGD classification (starting column A) and MIPS (starting from column W) on the first sheet showing them in the estimated order of significance. The five best scoring gene classes are presented for each BSC showing the name of the gene class with the original and the Bonferroni corrected log-p-value. The table also shows a flagging for clusters (first column of the results) having correlating clusters with better log-p-value in the other methods results. It also includes flagging for all the clusters that were observed to be identical in one or in both of the other results (second column of the results).

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-5-32-S1.xls]

Additional File 2

All the BSCs found from complete cluster tree with SGD and MIPS classes. Table 5 presents all the BSCs found from the complete cluster tree. The table presents separately the results obtained with SGD and MIPS (starting from column W) classification on the first sheet showing them in the estimated order of goodness. The five best scoring gene classes are shown for each BSC.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-32-S2.xls]

Additional File 3

All the BSCs found from Ward's method cluster tree with SGD and MIPS classes. Table 6 presents all the BSCs found from Ward's method cluster tree. The file presents separately the results obtained with SGD and MIPS (starting from column W) classification on the first sheet showing them in the estimated order of goodness. The five best scoring gene classes are shown for each BSC.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-32-S3.xls]

Additional File 4

50 best BSCs from average method obtained with MIPS classes visualized on top of the average cluster tree. Figure 8 shows the visualization of the 50 best BSCs obtained with MIPS classes shown in table 1 and in table 4 [see additional file 1]. This is also shown in figure 2. BSCs are linked to tables with ordinal numbers. Colour coded clusters are blue, protein synthesis; red, energy, mitochonrial genes and carbo-hydrate metabolism and green, cell cycle, differentiation and growth, nucleus, chromosome structure and mRNA processing.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-32-S4.eps]

Additional File 5

50 best BSCs from average method obtained with SGD classes visualized on top of the average cluster tree. Figure 9 shows the visualization of the 50 best BSCs obtained with SGD classes shown in table 4 [see additional file 1]. This is also shown in figure 3. BSCs are linked to tables with ordinal numbers. Colour coded cluster groups are similar to figure 8 [see additional file 4].
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-32-S5.eps]

Additional File 6

50 best BSCs from complete method obtained with MIPS classes visualized on top of the complete cluster tree. Figure 10 shows the visualization of the 50 best BSCs obtained with MIPS classes from table 5 [see additional file 2]. BSCs are linked to table with ordinal numbers shown in table 5. Colour coded cluster groups are similar to figure 8 [see additional file 4].
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-32-S6.eps]

Additional File 7

50 best BSCs from complete method obtained with SGD classes visualized on top of the complete cluster tree. Figure 11 shows the visualization of the 50 best BSCs obtained with SGD classes from table 5 [see additional file 2]. BSCs are linked with ordinal numbers shown in the table 5. Colour coded cluster groups are similar to figure 8 [see additional file 4].
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-32-S7.eps]

Additional File 8

50 best BSCs from Ward's method obtained with MIPS classes visualized on top of the Ward's method cluster tree. Figure 12 shows the visualization of the 50 best BSCs obtained with MIPS classes from table 6 [see additional file 3]. BSCs are linked with ordinal numbers shown in the table 6. Colour coded cluster groups are similar to figure 8 [see additional file 4].
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-32-S8.eps]

Additional File 9

50 best BSCs from Ward's method obtained with SGD classes visualized on top of the Ward's method cluster tree. Figure 13 shows the visualization of the 50 best BSCs obtained with SGD classes from table 6 [see additional file 3]. BSCs are linked with ordinal numbers shown in the table 6. Colour coded cluster groups are similar to figure 8 [see additional file 4].
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-32-S9.eps]

Additional File 10

Description of filtering of low-variance genes, sorting method for cluster tree and cluster-class correlation
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-32-S10.pdf]

Acknowledgements

PT would like to thank Eero Castren, Sami Kaski and his group members, Garry Wong and his group members and Marko Vauhkonen for comments, Jari-Pekka Ikonen from Comsol inc. Finland for matlab support and Jarmo Pirhonen from CSC for many matlab visualization tips. GW also corrected many parts of the text. Markus Storvik, Juha Knuuttila and Kari Mauranen have also contributed by reading and commenting on this article. This research was supported by Finnish Academy in parts by grant 50059.

References

1. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95:**14863-14868.
2. Weng S, Dong Q, Balakrishnan R, Christie K, Costanzo M, Dolinski K, Dwight SS, Engel S, Fisk DG, Hong E, Issel-Tarver L, Sethuraman A, Theesfeld C, Andrada R, Binkley G, Lane C, Schroeder M, Botstein D, Michael Cherry J: **Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins.** *Nucleic Acids Res* 2003, **31:**216-218.
3. Mewes HW, Frishman D, Gruber C, Geier B, Haase D, Kaps A, Lemcke K, Mannhaupt G, Pfeiffer F, Schuller C, Stocker S, Weil B: **MIPS: a database for genomes and protein sequences.** *Nucleic Acids Res* 2000, **28:**37-40.

4. Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, Altschuler SJ: **Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters.** *Nat Genet* 2002, **31:**255-265.
5. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102:**109-126.
6. Nikkila J, Toronen P, Kaski S, Venna J, Castren E, Wong G: **Analysis and visualization of gene expression data using Self-Organizing Maps.** *Neural Netw* 2002, **15:**953-966.
7. Kontkanen O, Toronen P, Lakso M, Wong G, Castren E: **Antipsychotic drug treatment induces differential gene expression in the rat cortex.** *J Neurochem* 2002, **83:**1043-1053.
8. Gibbons FD, Roth FP: **Judging the quality of gene expression-based clustering methods using gene annotation.** *Genome Res* 2002, **12:**1574-1581.
9. Oja Merja, Nikkila Janne, Toronen Petri, Wong Garry, Castren Eero, Kaski Samuel: **Exploratory clustering of gene expression profiles of mutated yeast strains.** *Computational and Statistical Approaches to Genomics* Edited by: Wei Zhang and Ilya Shmulevich. Boston, MA, Kluwer; 2002:65-78.
10. Segal E, R. Yelensky, Koller D: **Genome-wide discovery of Transcriptional Modules from DNA Sequence and Gene Expression.** *Bioinformatics* 2003, **19 Suppl. 1:**i273-82.
11. Sinkkonen J, Kaski S, Nikkila J: **Discriminative clustering: Optimal contingency tables by learning metrics.** *Lect Notes Artif Int* 2002, **2430:**418-430.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral