Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 40 (2007) 160–173
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Co-clustering and visualization of gene expression data and gene ontology terms for *Saccharomyces cerevisiae* using self-organizing maps

Markus Brameier *, Carsten Wiuf

Bioinformatics Research Center (BiRC), University of Århus, DK-8000 Århus C, Denmark
Molecular Diagnostic Laboratory, Århus University Hospital, Skejby DK-8200 Århus N, Denmark

Received 2 March 2006
Available online 20 May 2006

Abstract

We propose a novel co-clustering algorithm that is based on self-organizing maps (SOMs). The method is applied to group yeast (*Saccharomyces cerevisiae*) genes according to both expression profiles and Gene Ontology (GO) annotations. The combination of multiple databases is supposed to provide a better biological definition and separation of gene clusters. We compare different levels of genome-wide co-clustering by weighting the involved sources of information differently. Clustering quality is determined by both general and SOM-specific validation measures. Co-clustering relies on a sufficient correlation between the different datasets. We investigate in various experiments how much GO information is contained in the applied gene expression dataset and vice versa. The second major contribution is a visualization technique that applies the cluster structure of SOMs for a better biological interpretation of gene (expression) clusterings. Our GO term maps reveal functional neighborhoods between clusters forming biologically meaningful functional SOM regions. To cope with the high variety and specificity of GO terms, gene and cluster annotations are mapped to a reduced vocabulary of more general GO terms. In particular, this advances the ability of SOMs to act as gene function predictors.

Keywords: Co-clustering; Clustering visualization; Clustering validation; Gene function prediction; Gene expression data; Gene ontology; *Saccharomyces cerevisiae* yeast; Self-organizing maps

1. Introduction

In recent years, DNA microarrays [10] have become the state-of-the-art method for gene expression analysis and for understanding the underlying regulatory mechanisms of the cell. This high-throughput technology enables the simultaneous monitoring of expression levels for thousands of gene fragments. An *expression profile* of a gene describes the development of its expression level over a time series of microarray experiments.

The analysis of such a huge amount of data is challenging. It often requires time-consuming searches through the literature and public databases which is rather difficult to carry out by manual inspection and across the whole genome. One way of accelerating data analysis is to use data-driven methods [10,14,2,25]. Typically, clustering approaches are applied to identify groups of genes with similar expression patterns and to find potentially co-regulated and, thus, biologically related genes. Among these, hierarchical clustering, k-means clustering, and self-organizing maps (SOMs) [17,25] are the most popular ones (see e.g., review by Quackenbush [21]).

The basic underlying assumption of gene expression analysis is that genes with similar expression profiles are more likely to have similar biological function. However, sequential analysis by first clustering genes using gene expression data only and then assigning biological function

* Corresponding author. Fax: +45 8942 3077.
E-mail address: brameier@birc.au.dk (M. Brameier).

1532-0464/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2006.05.001

to the clusters may be suboptimal in a sense that it does not necessarily provide the best possible grouping by biological function. It is easy to observe genes with similar expression profiles in the same cluster that do not share biological similarity and, vice versa, genes known to share similar functions which end up in different expression clusters. Conventional clustering algorithms based solely on expression data, may not handle such *borderline cases* (or biological noise in general) sufficiently.

One way to reduce arbitrariness and resolve ambiguities during gene expression clustering is to integrate a certain amount of prior biological knowledge from other data resources directly into the clustering algorithm. Using additional sources of information for clustering genes (*co-clustering*) is supposed to cluster borderline cases more correctly and to level out noise and errors. In doing so, gene clusters may be (1) better biologically defined, i.e., become more meaningful in the biological sense, and (2) more compact and better separated with respect to *both* data sources. Co-clustering is further motivated by a merging of clusters with rather similar expression profiles, i.e., by an implicit reduction of the cluster number.

Many entries in bioinformatics databases, including sequences or gene expression profiles, are associated with biological annotations from natural language. Human-made information is difficult to interpret computationally and to use for an automatic data analysis. Ontologies provide a standard mechanism for capturing biological knowledge in the form of a terminology. The Gene Ontology (GO) [12] is currently the most popular source for biological annotations.

Co-clustering of genes, in particular with respect to expression profiles and biological annotations, is not a well-investigated field of research yet. Hanisch et al. [15] map genes to components of biological networks (metabolic pathways) and derive a distance function from their position in the network. This distance is combined with a gene expression distance and incorporated into hierarchical clustering. One drawback is that the network has to be known which is usually not the case. Cheng et al. [3] introduce a co-clustering approach that is based on hierarchical clustering as well, but uses the sum of gene expression distance and a GO-based distance as a simple weighting. The authors demonstrate their method by means of a small-scale clustering example and found a better biological clustering of borderline cases manually by visual inspection. Speer et al. [23] incorporate biological knowledge into an evolutionary clustering algorithm by weighting the influence of gene expression and GO information equally. However, the authors do not provide evidence of how their co-clustering method compares to standard clustering.

This paper presents a co-clustering approach based on self-organizing maps. SOMs denote a probabilistic clustering method that imposes a neighborhood structure on the clusters. Our method combines the center-based clustering of standard SOMs with a representative-based clustering. The latter defines the notion of *functional centers* which are derived from genes (and their associated GO terms) in each cluster. The success of combined clustering strongly depends on how the distance functions of the different clustering objectives are integrated. We favor a two-level cluster selection where the nearest cluster according to GO distance is selected among the m best matching clusters according to gene expression distance. By adjusting parameter m, we compare different levels of co-clustering yeast genes with respect to both their expression profiles and their biological (GO) annotations,[1] ranging from pure expression-based to pure annotation-based clustering.

Automated large-scale validation of clustering is performed using measures that reflect general clustering qualities or more SOM-specific features. Questions of particular interest are (1) how much additional GO information has to be provided to produce the best clustering of genes and (2) how much GO information is already contained in the applied gene expression data. Both questions are addressed in the paper in several ways.

The idea of co-clustering relies on the assumption that there is at least a minimum of correlation between the multiple sources of information. By means of different analysis and visualization techniques, which are supposed to provide different insights into the data, we are trying to quantify and give an impression of the amount of mutual information in both databases.

Besides the clustering algorithm, appropriate visualization techniques are essential for the biological interpretation of gene clusters. Clusterings of expression profiles are often represented by, so called, *heat maps* [10]. The expression vectors of all genes in a cluster are arranged in a matrix such that each column is labeled by a certain time point. Each matrix position (expression value) is assigned a certain color which ranges from saturated red (gene up-regulated) to saturated green (gene down-regulated). One alternative representation that is applied here plots the *average* expression profile for each cluster [25]. Such a visualization is especially suited for SOMs to better reflect a potentially higher degree of relationship between adjacent clusters.

Several visual data mining tools, e.g., MappFinder [7] or NetAffx [4], are available that link information from gene expression data to the graph structure of the Gene Ontology. In contrast to such approaches, we utilize the SOM structure for a technique named *GO term maps* to visualize the functional information *in and between* gene expression clusters. Due to the high complexity of the Gene Ontology, which aims to classify genes according to their highest specificity, there may be many *different* GO terms per cluster. Therefore, all GO terms associated with a gene are first mapped to a more concise and general GO vocabulary, called GO Slim. The frequency distribution of these high-level annotations gives a clearer functional image of a SOM cluster and its functional relationship to other clusters than is possible with

[1] GO annotations include biological *processes* here in the first place, but will also be referred to as biological *functions* in a more general sense.

all thousands of different low-level GO annotations. This technique is especially interesting when clustering many genes and for genome-wide analyses.

Assigning function to newly discovered gene sequences is an important goal of biosciences today. Different (mostly) unsupervised [10,27] and supervised methods [2,14,18] have been proposed for predicting the function of unknown genes from their expression profile. The cluster neighborhood in SOMs may improve gene function prediction such that genes of unknown function may not only be associated with known functions of other genes from the same cluster—which biologically may not be the most correct one—but also with functions of genes in adjacent clusters. Compared to unstructured clustering, this offers a higher potential of revealing new relationships for both uncharacterized and characterized genes.

Besides introducing the different techniques outlined above, the major contributions of this paper may be summarized as follows:

(1) By co-clustering of GO annotations and gene expression profiles it is possible to achieve a better biological clustering of genes while still maintaining a comparatively high quality of gene expression clustering.
(2) GO similarity and gene expression distance are (at least weakly) correlated. This indicates a sufficient amount of common information in the applied datasets.
(3) By using a general cluster validity index to measure functional clustering qualities, relatively small differences between a pure gene expression-based clustering and a random clustering are revealed. This hints to a rather low amount of GO information in the gene expression data.
(4) Our GO term maps show higher functional similarities between neighboring clusters induced solely by a gene expression clustering. Larger cluster regions of similar function are clearly distinguishable and provide biologically reasonable insights, e.g., by clear separation or by overlapping. These regions give additional confidence in the correctness of the single cluster annotations as well as an impression of the GO information that is contained in the gene expression data.
(5) Finally, gene function predictions become much less ambiguous when using the higher-level cluster annotations from the GO term maps instead of all GO annotations of genes in a cluster.

2. Methods and material

2.1. Gene ontology

The Gene Ontology (GO) has become one of the most important ontologies in bioinformatics and is maintained by the Gene Ontology Consortium. It defines a consistent, controlled standard vocabulary independently from any biological species that enables researchers to query across different databases. Today the GO incorporates over 18,000 different terms.

The GO terms are structured hierarchically in a Directed Acyclic Graph (DAG) such that each term, i.e., biological class, represents a node, and each connecting edge represents a relationship between terms. Nodes are allowed to have multiple parents as well as multiple children. There are two kinds of relationships, *is-a* relations and *part-of* relations, meaning that a child class is either a *part-of* the parent class or *is-a* more specific variant. The closer a node (term) is to the root, the more general is its biological class.

The GO comprises three orthogonal taxonomies corresponding to three domains of molecular biology: *biological process* (BP), *molecular function* (MF), and *cellular component* (CC). These are represented by separate disconnected subgraphs of the root node.

2.2. Datasets

We test our method on the microarray dataset of Spellman [24,32] which contains 6178 yeast (*Saccharomyces cerevisae*) genes and four time series of experiments from different origins. Gene expression levels are given as log ratios of the measured level and a reference (control) level. We have used time series *cdc15* and *cdc28*, comprising 24 and 17 experiments, respectively. All expression values in the dataset are centered around mean value 0 across the time points.

Gene Ontology annotations have been extracted from the UniProt database [29]. We restrict ourselves to annotations from the largest of the three branches, the BP branch.

GO Slims are smaller subsets of more general GO terms. There are different GO Slims available for different genomes. We use a version for yeast including 32 different biological process annotations in total (see Supplementary Table 1). The applied dataset [31] from the *Saccharomyces* Genome Database (SGD) [30,9] contains a GO Slim term for each yeast gene products (protein or RNA) and each GO branch. Due to the graph structure of the GO, the GO terms for a single gene may map to multiple GO Slim terms from the same ontology. If two such GO Slim terms are related, only the child term is chosen. In most cases, there is only one GO Slim term for each gene.

2.3. Data preprocessing

All experiments documented in this paper are based on the original Spellman dataset (*cdc15* time series) which suffers least from missing values. The *cdc28* time series has been applied for control experiments only.

First, we selected genes with at least one GO term entry in the UniProt database which describes a biological process. The expression profile of each gene in the resulting total set of 2264 genes has at most 1 missing value. All

missing expression values have been replaced by 0 (mean value). Second, we applied a filter to remove genes with little variation in expression across the profile and kept only the top 50% when ranked according to standard deviation. The resulting 1130 *high variation* genes are randomly distributed on two disjunct subsets, forming training set and test set in the following experiments.

Further note that for almost all genes (2124) a GO Slim annotation exists in the applied dataset. In 1852 cases this term is unique while the rest of the genes is associated with two (or very rarely three) GO Slim terms.

2.4. Self-organizing maps

Self-organizing maps (SOMs) [17] represent a (non-deterministic) machine learning approach to clustering which applies an unsupervised learning scheme. Some features make them particularly interesting for clustering gene expression profiles. First, they impose a partial structure on the *clusters*, i.e., *non-empty* SOM locations, and are, thus, well suited to exploratory data analysis. Second, unlike hierarchical clustering and k-means clustering, which are both deterministic[2] and operate only locally, SOMs get less likely stuck in local minima and have a higher robustness and accuracy. Third, SOMs facilitate both visualization and interpretation of the clustering results. Gibbons and Roth [13] compared different clustering methods for different gene expression datasets and distances and found SOMs to be superior to both hierarchical clustering and k-means clustering when evaluating the clustering result by a GO score.

Algorithm 1 describes the basic principle behind self-organizing maps. SOMs define a structure on the set of clusters, e.g., a $N \times N$ grid topology, and, thus, define a distance between the cluster nodes. Neighboring nodes tend to represent related clusters. Each node holds a center vector from k-dimensional data space. Initially the centers are random and then iteratively adjusted during the training phase.

Algorithm 1. (*standard self-organizing map*)

(1) Choose an l-dimensional topology (usually $l \in \{1,2,3\}$) of cluster nodes.
(2) Initialize the k-dimensional center vector of each cluster randomly.
(3) *Training phase*:
   (a) For each data point $\vec{p}$ find the nearest center vector $\vec{c}_p$ in k-dimensional space according to a distance metric d.
   (b) Move $\vec{c}_p$ and all centers $\vec{c}$ within its local neighborhood (according to a radius r in the l-dimensional cluster structure) closer to $\vec{p}$: $\vec{c}_{t+1} := \vec{c}_t + \alpha \cdot (\vec{p} - \vec{c}_t)$.
   (c) After each epoch $t = 1, \ldots, t_{\max}$: learning rate $\alpha_{t+1} := \alpha_t - \Delta\alpha$ where $\Delta\alpha := \alpha_0/t_{\max}$, neighborhood radius $r_{t+1} := r_t - \Delta r$ where $\Delta r := r_0/t_{\max}$.
(4) *Application phase*: Assign each data point to the cluster with the nearest center vector. Result is clustering $C = \{C_1, \ldots, C_n\}$.

Each iteration involves randomly selecting a data point $\vec{p}$ and moving the closest center vector a bit in the direction of $\vec{p}$. Only distance metric d defined *on the data space* influences the selection of the closest cluster. Additionally, other centers are moved whose clusters lie in a local neighborhood *on the grid*. In this way, nearby points in high-dimensional data space tend to be mapped to close positions in the low-dimensional cluster topology. Learning rate α and neighborhood radius r are decreased after each *epoch*, i.e., every n iterations where n is the total number of data points.

2.5. Distance measures

The DAG structure of the GO provides a way of estimating the degree of similarity between two terms. Lord et al. [20] adapt a couple of similarity and distance measures—including Resnik [22], Lin [19], and Jiang and Conrath [16]—to the Gene Ontology. All measures use the fact that less frequently used terms are more specific. The *information content* of a GO term t is defined as the number of times this term *or* any of its child terms occur in a certain database (e.g., Swiss–Prot) and is expressed as a *GO term probability* $P(t)$.

As indicated already, terms can share multiple parents. In this paper, we use the *GO similarity* measure developed by Resnik [22] that uses only the information content of the common parents. Eq. (1) calculates the minimum probability among all parent terms (denoted $S(t_1, t_2)$) shared by the two query terms t_1 and t_2.

$$s_{\mathrm{GO}}(t_1, t_2) := -\ln \min_{t \in S(t_1,t_2)} P(t). \tag{1}$$

Other measures use both the information content of the shared parent terms *and* that of the query terms [16,19]. The GO similarity score can be transformed into a *GO distance*:

$$d_{\mathrm{GO}}(t_1, t_2) := \min\{s_{\mathrm{GO}}(t_1, t_1), s_{\mathrm{GO}}(t_2, t_2)\} - s_{\mathrm{GO}}(t_1, t_2) \tag{2}$$

with $d_{\mathrm{GO}}(t, t) = 0$. We identify the GO distance (GO similarity) between two genes g_1 and g_2 with the *minimum* distance (*maximum* similarity) among all pairs of GO terms (t_1, t_2) annotated to these genes:

$$s_{\mathrm{GO}}(g_1, g_2) := \max_{t_1 \in T_{g_1}, t_2 \in T_{g_2}} s_{\mathrm{GO}}(t_1, t_2), \tag{3}$$

$$d_{\mathrm{GO}}(g_1, g_2) := \min_{t_1 \in T_{g_1}, t_2 \in T_{g_2}} d_{\mathrm{GO}}(t_1, t_2), \tag{4}$$

where T_{g_i} is the set of all GO terms of gene g_i. Note that with this definition $d_{\mathrm{GO}}(g, g) = 0$.

[2] Apart from the randomized initialization in k-means clustering.

For clustering gene expression profiles, several different distance functions have been proposed, including Euclidean distance and Pearson Correlation Coefficient as the most popular ones. We use the Euclidean distance here which is argued to be superior with ratio-based data [13]. The *Euclidean distance* d_E between two genes g_1 and g_2 is identified with the distance of their expression vectors $\vec{g_1}$ and $\vec{g_2}$:

$$d_E(g_1, g_2) := \|\vec{g_1} - \vec{g_2}\| = \sqrt{\sum_{i=1}^{n} (g_{1i} - g_{2i})^2}. \quad (5)$$

In the following, Euclidean distance d_E will be also referred to as *structural distance*. As a counterpart, GO distance d_{GO} between genes [see Eq. (4)] will be refereed to as *functional* distance. Moreover, depending on whether gene distance $d = d_E$ or $d = d_{GO}$ in Eqs. (10) and (11) we will speak of *Euclidean index* I_E or *GO index* I_{GO}, respectively. The cluster distances defined by Eqs. (10) and (11) will be treated accordingly.

The structural distance between a gene g and a cluster C_i is reduced to the Euclidean distance between expression vector $\vec{g}$ and center vector $\vec{c_i}$ in the SOM:

$$d_E(g, C_i) := \|\vec{g} - \vec{c_i}\|. \quad (6)$$

The functional gene–cluster distance is defined by the average pairwise GO distance between gene g and genes that have been assigned to cluster C_i, excluding g if it is in C_i:

$$d_{GO}(g, C_i) := \frac{1}{|C_i \setminus \{g\}|} \sum_{g_j \in C_i \setminus \{g\}} d_{GO}(g, g_j). \quad (7)$$

We distinguish two basically different ways of combining distance functions in co-clustering. The first is a simple *linear combination* of the two distance functions, i.e., a weighted sum of d_E and d_{GO}:

$$d_w(g, C_i) := (1 - w) \cdot d_E(g, C_i) + w \cdot d_{GO}(g, C_i) \quad (8)$$

with $0 \leqslant w \leqslant 1$. The second variant is referred to as *two-level cluster selection*. Assume that gene g is supposed to be added to a clustering C.

(1) All clusters $C_1, \ldots, C_n$ are ranked in ascending order by their Euclidean distance $d_E(g, C_i)$ to gene g.
(2) Among the m top ranking clusters the one with minimum distance $d_{GO}(g, C_i)$ is selected.

With this approach the functionally closest cluster is selected only among clusters that are already structurally close (for not too large m). Moreover, cluster selection depends only on the distance rank, but not directly on the distance value as with weighting parameter w. Finally, one can integrate both controls into one. That is, among the m top ranking clusters, the one with minimum weighted distance d_w is selected. In doing so, $w < 1$ allows a better fine-tuning of the GO influence if already $m = 2$ turns out to be too high.

2.6. Clustering validation methods

Most cluster validation techniques rely on internal features only, since, in many cases, additional external information is not available. For a literature review of validation methods that are especially suited for gene expression clustering we refer to, e.g., Yeung et al. [28] and Famili et al. [11]. Typical problems that are addressed in this context include the optimum clustering algorithm, the optimum choice of the cluster number, and the identification of highest quality clusters from a clustering. The optimum *clustering quality* indicates that the distance of genes from the same cluster is minimum while the distance of genes from different clusters is maximum.

We are interested in global or *large-scale* clustering validation, including averages over single cluster qualities, to compare different levels of co-clustering using a SOM-based algorithm (see below). Azuaje [1] combined different versions of the Dunn validity index [8]—based on different combinations of inner cluster and inter cluster distances—into a framework to validate clusterings of gene expression data. In this work we apply validity index I in Eq. (9), a variant we derived from the popular Davies–Bouldin Index (DBI) [6]:

$$I(C) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n-1} \sum_{\substack{j=1 \\ i \neq j}}^{n} \frac{d_{inner}(C_i) + d_{inner}(C_j)}{d_{inter}(C_i, C_j)}, \quad (9)$$

where $C = \{C_1, \ldots, C_n\}$ is the given clustering, $d_{inner}(C_i)$ is the *inner* cluster distance of a cluster C_i, and $d_{inter}(C_i, C_j)$ denotes the *inter* cluster distance between two clusters C_i and C_j. The index is small if the clusters are compact and/or far distant from each other. Consequently, a small index indicates a good clustering. In the original definition of the Davies–Bouldin index, the second average is replaced by a maximum which puts more weight to single clusters. Moreover, d_{inner} is defined originally as the standard deviation of a gene expression vector from its cluster mean, while d_{inter} is the distance between two cluster means. Since it is difficult to identify cluster means for the GO distance (using $d = d_{GO}$ in Eqs. (10) and (11)) we simply calculate the *pairwise* gene distance or, more precisely, the *root mean square* (RMS) distance:

$$d_{inner}(C_i) := \left(\frac{1}{\binom{|C_i|}{2}} \sum_{g_1, g_2 \in C_i, g_1 \neq g_2} d(g_1, g_2)^2 \right)^{1/2} \quad (10)$$

and for $i \neq j$:

$$d_{inter}(C_i, C_j) := \left(\frac{1}{|C_i||C_j|} \sum_{g_1 \in C_i, g_2 \in C_j} d(g_1, g_2)^2 \right)^{1/2}. \quad (11)$$

In the result section we will, in addition, discuss and apply other, more specific cluster distances with some being defined particularly for self-organizing maps.

2.7. SOM co-clustering algorithm

In the following the center vector of each SOM cluster is referred to as the *structural center* while all genes (or a subset of these) that have been assigned to the cluster (and their associated GO terms) act as its *functional center*.

Co-clustering means here that a gene expression-based clustering and a Gene Ontology-based clustering are performed in parallel. In principle, Algorithm 2 combines the center-based clustering of standard SOMs (see Algorithm 1) with representative-based clustering. The structural part of this co-clustering method applies cluster means, the functional part uses cluster representatives.

Algorithm 2. (*SOM co-clustering*)

(1) Initialize the (structural) center vectors $\{\vec{c}\}^0$ of the $N \times N$ SOM randomly.
(2) Initialize the functional centers $C^0 = \{C_1, \ldots, C_n\}$ by distributing all (or a subset of the) genes randomly over the SOM locations.
(3) Train SOM using Euclidean distance d_E *and* GO distance d_{GO}:
    (a) Select each gene g exactly once per epoch and in random order.
    (b) Rank all clusters in ascending order by distance $d_E(g, C_i)$ [see Eq. (6)] between their functional center C_i and g.
    (c) Among the m top ranking clusters select cluster i with minimum weighted distance $d_w(g, C_i)$ [see Eq. (8)].
    (d) Update (structural) center vector $\vec{c}_i$ of nearest cluster i and all centers within its local neighborhood using expression vector $\vec{g}$ of gene g (see Algorithm 1).
    (e) Update functional center C_i with g (by moving it from its previously assigned center). If a maximum center size has been exceeded, remove a gene according to a certain selection criterion (optionally).
    (f) After each epoch $t = 1, \ldots, t_{max}$: Update settings for learning rate α and neighborhood radius r (see Algorithm 1). Update GO weight w and/or number of GO-compared clusters m (optionally).
(4) Apply SOM by assigning each vector to its nearest cluster (as done in Steps 3b and 3c). Result is clustering $C' = \{C'_1, \ldots, C'_n\}$.

The initialization of the functional centers in Step 2 is essential. Otherwise the GO distance to an empty SOM location may not be maximum and may have a high impact on the clustering result, especially on the cluster number.

In Step 3 (training phase) the ranking of clusters that are nearest to a gene g is solely based on the Euclidean distance between the SOM center vectors and expression vector $\vec{g}$. Intuitively, m allows vectors to be moved to clusters within another type of structural "neighborhood" that is based on their Euclidean distance to gene g (not on the SOM structure). Optionally, the two parameters, GO weight w and/or number of GO-compared clusters m, may be updated after each epoch in different possible directions. For the experiments documented in this paper, however, we only applied constant settings. Moreover, the GO influence will be controlled by varying parameter m only. Parameter w is always fixed to 1 and thus $d_w(g, C_i) = d_{GO}(g, C_i)$ [see Eqs. (7) and (8)].

Using always *all* genes that are assigned to a cluster as its functional center is feasible as long as the ratio of gene number n and SOM size N^2 is not too large. Otherwise, such representatives might be selected more specifically (Step 3e), already to limit the computational overhead. Instead of just selecting a center gene randomly or rejecting new genes after a certain maximum center size has been exceeded, one might, for instance, replace the gene with the *largest* GO distance to any other gene in the center. The final application phase (Step 4) is absolutely compulsory if clusters and functional centers are not identical.

3. Results

Since SOM clustering applies a non-deterministic algorithm, statistical evidence is given over multiple trials. All values or plots (excluding example maps) in this paper represent results over minimum 20 independent clusterings. In general this leads to a relatively low statistical standard error[3] (SE). Note that SOM clusterings from different runs cannot be compared position-wise. Similar clusters may be located at completely different SOM positions. Only the distances and relative positions of clusters on the grid structure are similar.

3.1. Gene expression—gene ontology correlation

A basic precondition of co-clustering is a sufficient correlation between the applied multiple sources of information. In our case, co-clustering is based on the assumption that similar expression profiles are more likely involved in similar biological processes. This requires sufficient correlation between gene expression data and GO annotations. In Lord et al. [20] correlation is demonstrated between three different GO similarity measures and sequence similarity. Wang et al. [26] have shown that this also holds true for the same set of GO similarity measures and Pearson correlation coefficients of gene expression profiles.

Fig. 1 plots the GO similarity [see Eq. (3)] against the (binned) Euclidean distance of expression profiles [see Eq. (5)]. Distances are calculated over all $\binom{2264}{2}$ gene pairs and for the 50% of genes with highest variation in expression values (see Section 2). In both cases correlation is significant, even though rather weak in general. In the latter case correlation is more pronounced ($r = -0.072$,

[3] The *standard error* is defined as standard deviation divided by $\sqrt{n}$ with n is the number of runs.

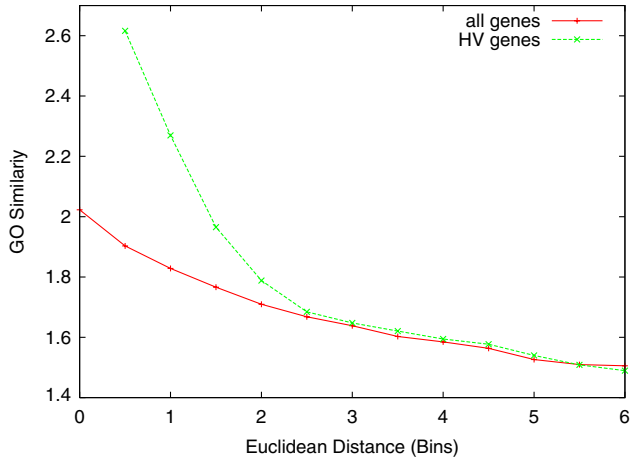

Fig. 1. GO similarity vs. Euclidean distance of expression profiles. GO similarity (Resnik [22]) vs. Euclidean distance of expression profiles. Clear correlation when using all gene pairs. Higher correlation for smaller Euclidean distances when including only genes with high variation (HV) in expression values. Euclidean distances are binned to level out noise.

$p < 10^{-7}$) than in the former case ($r = -0.059$, $p < 10^{-7}$), especially for smaller Euclidean distances. Apparently, profiles with low expression levels are more noisy.

3.2. Clustering validation

One goal of co-clustering is to obtain a better (*functional*) clustering of the GO terms without necessarily a correspondingly worse (*structural*) clustering of the gene expression data. This trade-off, i.e., the influence of the GO information, is controlled solely over the number of GO-compared clusters m in the following (assuming always $w = 1$ in Algorithm 2). For a $N \times N$ SOM, a pure GO-based clustering corresponds to $m = N^2$, while a pure expression-based clustering corresponds to $m = 1$. Intermediate settings $1 < m < N^2$ mean co-clustering.

Tables 1 and 2 show results of 8×8 SOM clusterings. Cluster validity indices I_{GO} and I_E (see Section 2) measure how well the (combined) clusterings preserve the quality of the GO information or the expression information, respectively. In general, a smaller index value indicates a better clustering. By comparing *relative difference* factor δ in both Tables, one can see that the GO validity index I_{GO} decreases relatively more with m than the Euclidean validity index I_E increases. This indicates that at least to a certain degree the clustering quality is relatively less reduced in terms of the Euclidean distance than it is improved in terms of the GO distance.

Table 1
GO validity index and mean cluster distances

m	I_{GO}	δ	d^{GO}_{inner}	d^{GO}_{inter}
1	1.78	1.00	4.5	4.9
2	1.59	1.12	4.0	5.0
8	1.11	1.60	2.9	5.2
32	0.61	2.92	1.6	5.4
64	0.40	4.45	1.1	5.5
r	2.00	—	4.9	4.9

GO validity index I_{GO}, mean inner cluster distance d^{GO}_{inner}, and mean inter cluster distance d^{GO}_{inter}. 8×8 SOM clustering using different numbers of GO-compared clusters m (during training). Smaller index value implies better clustering. Standard errors (SEs) $\leqslant 0.01$ for index values and $\leqslant 0.02$ for distance values. Relative difference δ compared to standard gene expression clustering ($m = 1$). Index value about 10% higher in random clustering r than for $m = 1$.

Table 2
Euclidean validity index and mean cluster distances

m	I_E	δ	d^{E}_{inner}	d^{E}_{inter}
1	1.16	1.00	2.5	4.7
2	1.22	1.05	2.7	4.6
8	1.38	1.19	3.0	4.6
32	1.64	1.41	3.6	4.5
64	1.81	1.56	4.0	4.4
r	2.00	—	4.4	4.4

Euclidean validity index I_E, mean inner cluster distance d^{E}_{inner}, and mean inter cluster distance d^{E}_{inter}. SEs similar to those of Table 1.

Tables 1 and 2 further show the effect of m on the inner and the inter cluster distance, each *averaged* over all cluster pairs. The clustering quality improves if d_{inter} increases while d_{inner} decreases [see Eqs. (10) and (11)]. As one might expect, for the GO variant of both distances, a better clustering quality is achieved with larger m, and for the Euclidean variant with smaller m.

In both cases, inter cluster distances change less than inner cluster distances which is partly a side-effect from the *pairwise* distance that is calculated between genes.

Pure GO-based SOM clustering (using $m = 64$) performs slightly better than random clustering (r in Tables 1 and 2) in terms of the Euclidean index. Correspondingly, a pure gene expression-based clustering ($m = 1$) performs better than a random clustering when comparing the GO indices. In both cases, the relative difference between the validity indices is about 10%. We cannot directly conclude, however, that this percentage reflects already all the common information that is in both data sources (see also Sections 3.7 and 4). In any case, combined clustering configurations ($1 < m < 64$) may be found where both GO index and Euclidean index are clearly smaller than their random counterparts (always 2).

Another indicator for the common information content in the two databases is the difference between the inner and the inter cluster distance. While this difference is zero in a random clustering, there is a small, but statistically highly significant ($p \ll 0.01$) difference when comparing the GO equivalents for minimum m (see Table 1) and the Euclidean equivalents for maximum m (see Table 2).

What is the optimum trade-off? To answer that question we plot the sum of the normalized index values over different configurations for parameter m in Fig. 2 Both index values, I_{GO} and I_E, are scaled to the same range ([0,1]) before summation. The optimum configuration occurs

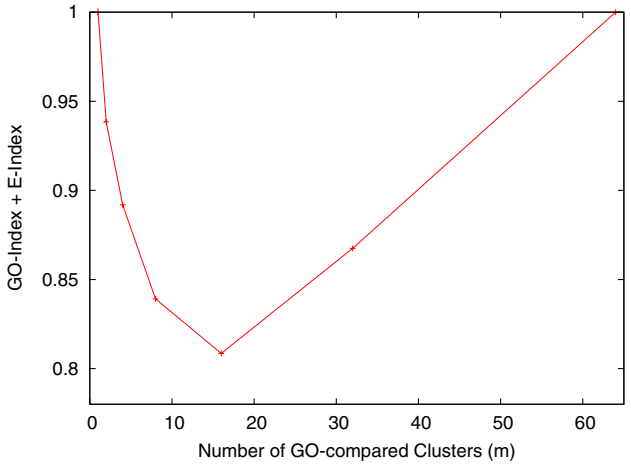

Fig. 2. Sum of GO index and Euclidean index over number of GO-compared clusters m. Sum of indices $I_{GO} + I_E$ over number of GO-compared clusters m. GO validity index I_{GO} and Euclidean validity index I_E scaled to same range ([0,1]) before summation. Minimum means best co-clustering, i.e., best trade-off in terms of both objectives. Optimum configuration for m is 16 here.

when this sum is minimum (here with $m = 16$). This is supposed to produce the best co-clustering, i.e., the best trade-off in clustering quality between the gene expression clustering, on the one hand, and the GO term clustering, on the other hand.

One has to note, however, that the optimum setting always depends on the applied validation and distance measures, i.e., on the definition of clustering quality. The DBI—like most other cluster validity indices—ignores the SOM structure and does not measure how far the cluster neighborhoods are preserved. It is also not appropriate to compare clusterings of *different* data using the same SOM, e.g., during training and testing. Both cases require other measures that will be introduced further below.

3.3. Merging of clusters

Co-clustering with additional functional information allows the merging of structurally similar SOM clusters. This is not practiced by combining full clusters explicitly. Instead, the GO information helps to decide between clusters whose expression profiles are similarly close to the profile of a newly assigned gene (see Algorithm 2). At least to a certain extend this is supposed to improve the biological meaning of gene groups (1) by avoiding too small cluster sizes and (2) by leveling out wrong and missing information in the data sources.

The merging effect is documented in Supplementary Table 2. The number of clusters, i.e., non-empty SOM positions, converges to a certain minimum with increasing m, while clusters become larger on average and more different in size. The cluster number drops down to *42* for maximum m, but is reduced most for $m \leqslant 8$ already. Apparently, the influence of the GO information on this clustering quality grows quite non-linear in m.

Another benefit of merging is that the final clustering depends less on the chosen size of the SOM, i.e, the *maximum* possible number of clusters (*64*). A certain self-adaptation of the cluster number by the clustering algorithm is especially helpful when dealing with gene expression data. For this type of data a general recommendation for the choice of the cluster number can hardly be given [13].

3.4. GO slim mapping

The Gene Ontology provides the most detailed information available by annotating gene products to the most granular GO term(s). For example, if a gene product is localized in the *perinuclear space*, it will be annotated to that particular term, but not necessarily to the parent term *nucleus*, too. In many cases, like the functional analysis of gene expression clusters, it is useful to have a higher level view of the Gene Ontology. The biological annotations of (all genes in) a cluster are less ambiguous and much more human-interpretable if a smaller subset of high-level GO terms (GO Slim) is used instead of all GO terms, including very many low-level terms. Therefore, more specific GO terms have to be mapped to more general GO terms. Such a GO Slim mapping does not only hide details but also levels out noise.

Supplementary Table 1 contains a frequency distribution of the genes in the dataset over 32 different GO Slim terms (biological processes in yeast). For the following experiments we reduce training set and test set to genes which may be assigned to exactly one term. In total 995 genes out of the 1130 high variation genes are annotated with exactly one GO Slim term (see also Section 2).

Note that the GO Slim subset is used only for visualization purposes while still the full Gene Ontology is used for computing the GO distance in Algorithm 2.

3.5. Clustering visualization and interpretation

To visualize the quality of a gene clustering in terms of the gene expression information, on the one hand, and the Gene Ontology information, on the other hand, we apply two different representations of SOMs.

Fig. 3 gives an example of a 6×6 SOM clustering with each cluster's content represented by its mean expression profile, i.e., the simple average over the expression vectors of all genes in a cluster plotted over the vector positions (time points). For the same clustering (of training genes) Fig. 4 uses the mapping of genes to a single higher level GO Slim term. Each cluster position of such a *GO term map* holds a frequency distribution of terms. The different GO terms are ranked by their absolute number. A term is only used here as a cluster annotation (and printed) if it occurs more than once in its cluster.

This example represents a typical result of a self-organizing map if trained with a pure Euclidean distance function ($m = 1$). While the expression profiles are most diverse over the whole SOM, profiles of neighboring clusters in

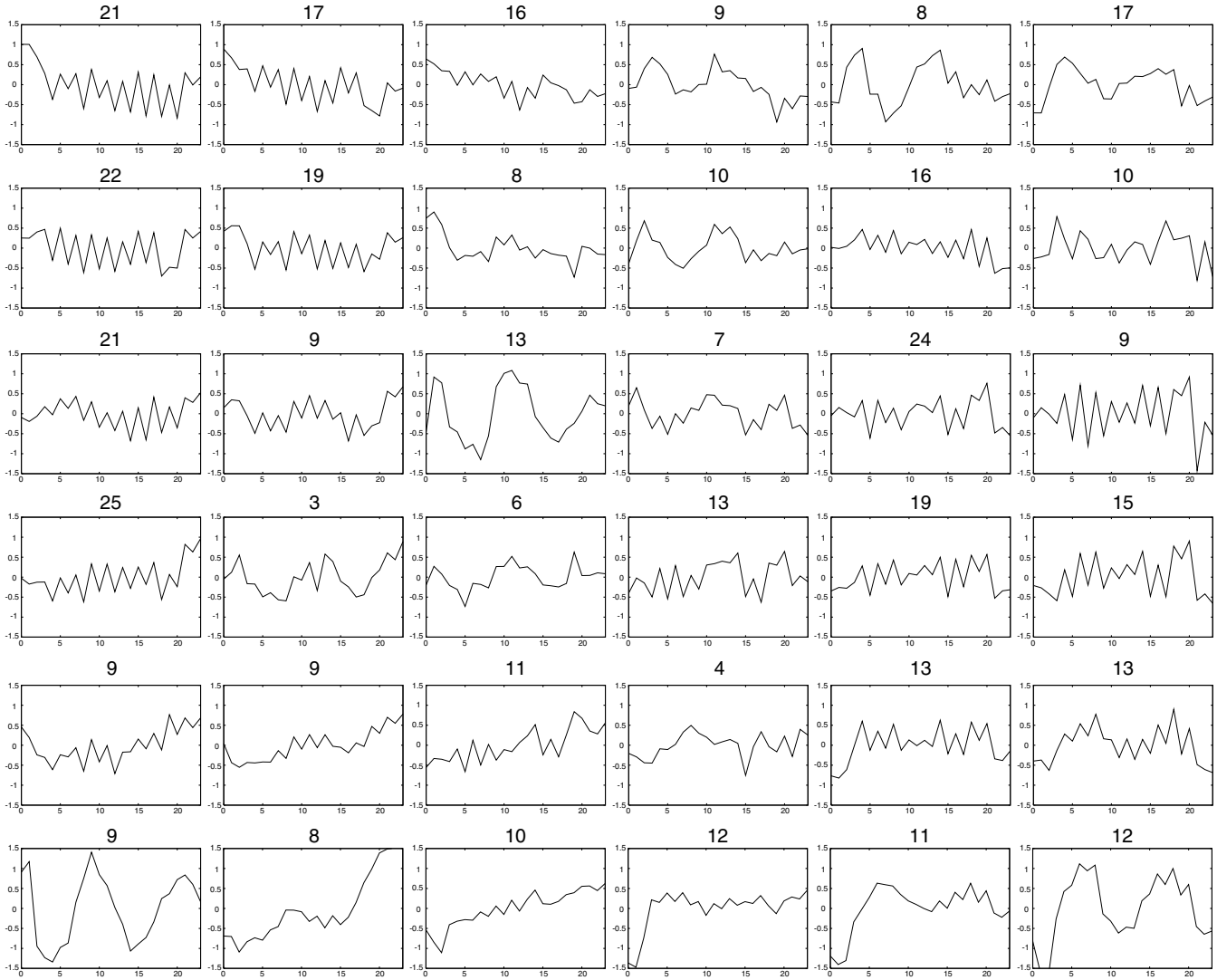

Fig. 3. Gene expression map. Example 6×6 SOM clustering (training data) with mean expression profiles of clusters (over time points) and cluster sizes. Standard gene expression-based clustering ($m = 1$). Clearly higher similarities between (directly) neighboring clusters. Genes regularly distributed over SOM.

Fig. 3 are clearly more similar than profiles of more distant clusters. The given cluster sizes demonstrate that genes are distributed quite regularly over the SOM locations. Even though the differences in cluster size may be still large, empty SOM locations are not very frequent (see also Section 3.3).

The corresponding clusters in Fig. 4, by comparison, reveal relatively low (maximum) frequencies of GO Slim terms and several different terms per cluster. Nevertheless, noticeable higher functional similarities may be observed between annotations in neighboring SOM locations. Note that for a pure expression-based clustering, these are induced solely over correlations to structural similarities. Besides, the cluster structure gives additional confidence in the correct clustering of a term if its location in the SOM is close to other clusters holding the same term.

The cluster structure of the GO term map allows us to get a better idea of the amount of biological information that is in the gene expression data. In doing so, closely related biological terms do not necessarily have to fall into the same cluster (see also Section 3.8). Instead, their clusters may be just closely located on the SOM grid. In the example map the upper left corner is clearly dominated by genes annotated with *RNA metabolism*. The SOM region of these clusters overlaps—fully or at least partly—with other connected regions whose genes take part in dependent biological processes. These include *transcription*, *ribosome biogenesis*, and *DNA metabolism* here. It is interesting to note that some of the cluster regions, e.g., *RNA metabolism*, express very similar functional patterns in Fig. 3. Other less directly related processes like *transport* fall into more distinct regions of the map, in this case the bottom right corner. Thus, neighborhood relations between different SOM locations demonstrate clearly that structurally similar regions have similar function.

	1	2	3	4	5	6
1	ribosome biogenesis 7 RNA metabolism 6 conjugation 2 protein biosynthesis 2 transcription 2 transport 2	RNA metabolism 4 ribosome biogenesis 3 transcription 2	transcription 3 ribosome biogenesis 3 amino acid metabolism 2 protein catabolism 2		amino acid metabolism 2	protein modification 3 cell wall organization 2 RNA metabolism 2 meiosis 2 amino acid metabolism 2 response to stress 2
2	DNA metabolism 4 RNA metabolism 4 protein modification 3 transcription 2	DNA metabolism 5 protein biosynthesis 3 RNA metabolism 3 protein modification 2 response to stress 2	protein modification 2	cytoskeleton organization 4	protein modification 4 transport 2 cell homeostasis 2	transport 2
3	protein biosynthesis 4 vesicle-mediated transport 2 organelle organization 2 cell cycle 2 RNA metabolism 2	RNA metabolism 3	DNA metabolism 5 cell wall organization 2	lipid metabolism 2 DNA metabolism 2	transport 5 response to stress 4 DNA metabolism 2 organelle organization 2	transport 3 meiosis 2
4	protein biosynthesis 10 protein catabolism 4 amino acid metabolism 2 transport 2		protein biosynthesis 2	transport 2 vitamin metabolism 2	transport 4 protein modification 3 organelle organization 2 cytoskeleton organization 2	transport 7 protein modification 2
5		protein biosynthesis 3 response to stress 2 carbohydrate metabolism 2	response to stress 3 protein biosynthesis 2 amino acid metabolism 2		cellular respiration 3 lipid metabolism 2 transport 2	transport 6 protein biosynthesis 3
6	cell wall organization 3 cell cycle 2	conjugation 2	protein biosynthesis 3 transport 2	electron transport 3 transport 2 cellular respiration 2	transport 3 amino acid metabolism 3 cellular respiration 2	transport 6 response to stress 2

Fig. 4. GO term map. Example 6×6 SOM clustering (training data) with GO Slim terms of each cluster ranked by absolute frequency. Standard gene expression-based clustering ($m = 1$). Same SOM clustering as in Fig. 3. Higher functional correlations between (directly) neighboring clusters. Terms with frequency 1 not printed.

3.6. Co-clustering visualization and generalization

In the Supplementary material[4] two additional example clusterings are shown[5]: a combined clustering using both the Euclidean distance and the GO distance ($m = 4$) and a pure GO-based clustering ($m = 36$). For each clustering there are shown four different figures, including a gene expression map and a GO term map for both training data and test data.

For larger values of m, there are more different expression profiles assigned to a cluster, causing the average profile to flatten out over the time points. Moreover, the average expression profiles become more similar between (arbitrary) clusters. On the other hand, higher numbers of GO Slim terms and fewer different terms per cluster may be observed. The neighborhood relations between adjacent clusters, however, become weaker in *both* gene expression maps and GO term maps. All these developments reach a peak level for maximum m (see Supplementary Figures 5 and 11). In a pure GO-based clustering, higher (structural and functional) correlations between adjacent clusters may only be random effects.

Positive features—smaller inner and larger inter cluster distance—are to be found in a co-clustering for *both* clustering objectives. To what extent these appear depends on the co-clustering level, i.e., the configuration of m. In Supplementary Figure 3 cluster profiles are only slightly less diverse (globally) compared to Fig. 3 while local neighborhoods are still preserved. Instead, functional cluster annotations are biologically more clearly defined, i.e., more genes with the same GO Slim term fall into the same clusters, when comparing Supplementary Figure 9 with Fig. 4. Functional relations between (direct) cluster neighbors are, however, relatively weaker in this example and better preserved for smaller $m > 1$.

To understand why the functional neighborhood relations of SOM clusters do not become stronger for larger m, one should keep in mind, that our co-clustering approach does not include a direct neighborhood adaptation for the GO terms. Actually, a functional neighborhood may be established only indirectly over correlations to the structural neighborhood.

Test data is assigned to the self-organizing map using the same m value as during training. When comparing the expression profiles of training clusters and test clusters at equal SOM positions, these are obviously most similar in pure gene expression-based clusterings (compare Supplementary Figures 1 and 2). Noticeable functional similarities exist between the corresponding GO Slim maps (see also Section 3.8). It is important to realize that these are induced only indirectly over the structural similarities.

The opposite case may be observed in pure GO-based clusterings where training terms and test terms are most similar for same SOM locations (see Supplementary Figures 11 and 12). Even though the gene expression information is not used for clustering here, there are still noticeable similarities between the corresponding mean expression maps.

Using the GO distance exclusively does not necessarily lead to an absolutely perfect clustering of GO Slim terms. Genes annotated with the same GO Slim term, like *transport*, do not necessarily fall into the same cluster (see Sup-

[4] Due to space limitations we refer to some Supplementary figures here.
[5] Supplementary Figures 1 and 7 are identical to Figs. 3 and 4.

plementary Figure 11). The main reason is that the distance between two genes is measured as the (minimum) pairwise distance between *all* their GO terms [see Eq. (4)] and not directly between their GO *Slim* terms. Considering the fact that (especially lower-level) GO term annotations may be ambiguous and erroneous in databases, it is interesting that the result comes so close to a perfect clustering of higher-level terms.

3.7. SOM validation

While the above observations are made on the basis of examples, they will now be confirmed by averaging results over multiple SOM clusterings. Therefore, in this section we define distances that operate directly on the different feature maps.

Fig. 5 visualizes the mean ranking of (different) GO Slim terms in a cluster by their absolute frequency (number). For each rank the number of terms is averaged over all clusters and clusterings, ignoring the fact that different terms may have this rank in different clusters. If the GO influence is configured larger (using m) the number of terms increases, especially at rank 1, and less different terms fall into a cluster. Pure gene expression clustering ($m = 1$) turned out to be slightly better than pure random clustering (r) in this sense (see also below).

Tables 3 and 4 summarize results for the following three distance variants all based on a certain cluster distance d.

- *Position-wise* distance d_{pos} is defined as the average distance over all pairs of test and training clusters at the *same* SOM position.

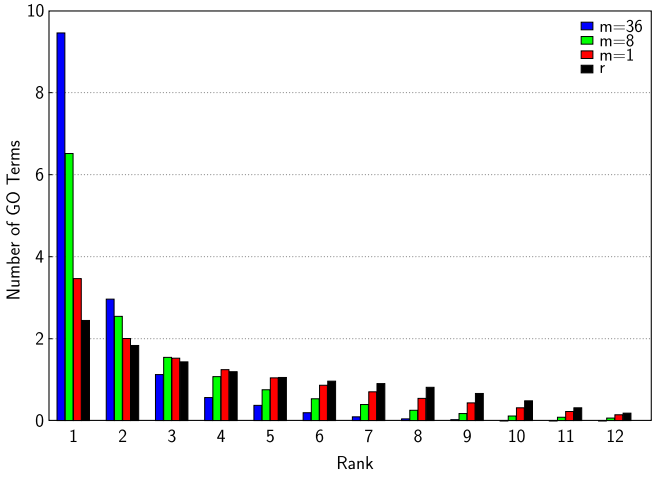

Fig. 5. Mean ranking of GO Slim terms in a (training) cluster by absolute number over all clusters and clusterings, ignoring different term identifiers at the same rank. Rankings shown for different numbers of GO-compared clusters m and random clustering r. Higher (maximum) numbers of GO Slim terms and less different terms occur in clusters with larger m. Very similar distribution in test clusters (not shown).

Table 3
Euclidean-based SOM validation

m	$d^{\text{E}}_{\text{pos}}$	$d^{\text{E}}_{\text{inter}}$	$d^{\text{E}}_{\text{neigh}}$
1	0.65	2.65	1.66
2	0.75	2.64	1.67
8	0.90	2.42	1.66
36	1.00	1.34	1.27
r	1.23	1.02	1.02

Cluster distance d^{E} defined as Euclidean distance between *mean* expression profiles. *Position-wise* distance d_{pos} between test and training clusters at *same* location in 6×6 SOM. Corresponding distance to random test clustering r. *Inner* cluster distance d_{inner} of either clustering is constantly 0. *Global inter* cluster distance d_{inter} between *different* training clusters (similar for testing). *Local inter* cluster or *neighborhood* distance d_{neigh} between *directly* neighboring cluster positions. Corresponding distances for random training clustering r. SEs $\leqslant 0.02$ for d_{pos} and $\leqslant 0.03$ for the other distance values.

Table 4
GO-based SOM validation

m	$d^{\text{GO}}_{\text{pos}}$	$d^{\text{GO}}_{\text{inter}}$	$d^{\text{GO}}_{\text{neigh}}$
1	0.54	0.81	0.65
2	0.48	0.84	0.76
8	0.31	0.90	0.92
36	0.16	0.91	0.90
r	0.77	0.75	0.75

Cluster distance $d^{\text{GO}}(C_i, C_j)$ defined as the proportion of GO Slim terms from a cluster C_i that do *not* occur in a cluster C_j (maximum distance 1). Only terms with frequency 2 or higher are considered. SEs $\leqslant 0.02$ for d_{pos} and $\leqslant 0.01$ for the other distance values.

- *(Global) inter* cluster distance d_{inter} denotes the average distance over all pairs of *different* clusters (from the same clustering).
- *(Local) neighborhood* distance d_{neigh} includes only directly neighboring SOM locations, instead. Two clusters are considered as *direct neighbors* if their SOM coordinates have maximum Euclidean distance $\sqrt{2}$. Each cluster may have 8 direct neighbors at the most.

While d_{inter} is independent of neighborhood relations between clusters, d_{pos} and d_{neigh} exploit specific qualities of the SOM structure. d_{pos} measures the generalization ability of a SOM, i.e., how far the self-organizing map has learned to cluster (unknown) data that is different from the training data. The *difference* between d_{neigh} to d_{inter} reflects how much more similar two neighboring clusters are on average than two arbitrary clusters.

In Table 3 cluster distance d is defined as the Euclidean distance between the *mean* expression profiles of two clusters. In Table 4 $d(C_i, C_j)$ is the proportion of (different) GO Slim terms in cluster C_i that do *not* occur in cluster C_j (non-symmetric distance). It means that only the Boolean distance is calculated between GO Slim terms ($0 =$ identical, $1 =$ different), the GO distance [defined by Eq. (2)] is not taken into account. Moreover, this definition is

independent of the number of (different) terms and their frequencies in a cluster (apart from the fact that only terms with frequency 2 or higher are considered). This minimum threshold might be selected higher with higher numbers of genes per cluster.

Both cluster distances are closely related to the feature maps used for visualization, in contrast to the cluster distance defined on the basis of pairwise gene distances. [see Eqs. (10) and (11)].

Results of both tables may be summarized as follows. First, the Euclidean-based distance d^{E}_{pos} increases between training and test clusters (in Table 3) together with the number of GO-compared clusters m, while the GO-based counterpart d^{GO}_{pos} decreases (in Table 4). Not surprisingly, the generalization ability of the SOM is best with one or the other measure if its influence during training is maximum. In both tables relative differences between the training clustering (for $m = 36$ in Table 3 and $m = 1$ in Table 4) and a random test clustering have been found to be significantly larger.

Second, average Euclidean-based distance d^{E}_{neigh} in Table 3 documents clearly higher similarities of training clusters in a local SOM neighborhood. The difference to the (global) inter cluster distance d^{E}_{inter} hardly changes up to $m = 8$ at least. As reported in Section 3.2, this shows again that co-clustering affects the clustering quality less in terms of the gene expression data than the GO terms (see below). When clustering with the GO distance only ($m = 36$) this difference (almost) disappears because there is no active adaptation of the Euclidean neighborhoods anymore. Nevertheless, both inter cluster distances are clearly smaller in a pure random clustering.

GO-based distances d^{GO}_{inter} and d^{GO}_{neigh} in Table 4 show a difference of about 16% for standard clustering ($m = 1$). This means an almost twice as high cluster similarity in direct neighborhoods. The difference shrinks to about 8% already for $m = 2$ and is not visible for $m \geqslant 8$ anymore. As argued in the previous section, functional relations between clusters are only induced indirectly over structural relations. The latter become weaker if the influence of the gene expression information is reduced.

All experiments include comparisons between pure gene expression-based (GO-based) clustering and random clustering using only GO-based (Euclidean-based) distances. In general, *relative* differences are higher for the SOM-related cluster distances used here than for the cluster distance used in Tables 1 and 2. This reveals a higher amount of mutual information that is contained in the gene annotations and the gene expression data.

3.8. Function prediction

Trained SOMs may be applied as explicit function predictors such that genes of *unknown* function are classified based on the structural objective only, i.e., the gene expression information. In doing so, the existing GO annotations of the destination clusters may be associated with these genes. Obviously, this requires that the common functional profile of all genes in a cluster is clearly defined and not too diverse.

In the previous section GO Slim maps have been used to measure the generalization ability of SOMs in terms of (co-)clustering. Test genes are assigned to a trained SOM using the same m value as during training. Then test and training clusters are compared position-wise by calculating the probability (d_{pos}) with which a GO Slim term in a test cluster does *not* appear in the corresponding training cluster.

In contrast to that, test genes are assigned here *without* using their GO annotations, i.e., m is always 1 during testing. Thus, only results concerning pure expression-based clustering apply here as well. An average rate of 46% common GO Slim terms in test clusters and training clusters is reported ($d_{pos} = 0.54$ in Table 4). The best prediction rate is 55% (out of 20 SOM clusterings). When comparing training clusterings with random test clusterings, chances drop down to 23 percent ($d_{pos} = 0.77$ in Table 4) on average and 27% at best.

What one has to consider here is that prediction quality is measured based on GO terms at the *same* cluster position only, not including terms in neighboring clusters. The neighborhood structure of SOM clusters is able to reveal functional correlations even for test genes that do not exactly fall into the perfect (GO-closest) training cluster. By searching the direct cluster neighborhood in such a case, function prediction becomes less precise but is still possible, unlike other methods, e.g., unstructured k-means clustering. Of course, the prediction quality also depends on the distribution of terms over training genes and test genes.

Although gene function prediction, by definition, is based on the expression data only, it may still profit from co-clustering, at least if SOMs are *trained* with comparatively small settings for parameters m (>1) and w (not documented). In this way, the (structural) center vectors of the SOM are not altered too much in favor of a better functional annotation of genes and clusters.

4. Discussion and conclusions

In addition to the discussions of single results, this section brings results together and draws more general conclusions. This includes also other experimental experiences we made which could not be documented in detail here.

All experiments in this paper have been performed, in addition, with an equally-sized training set selected from the *full* gene set, including low variation genes. Compared to using high variation genes only, results are similar in principle, but may differ in detail. A lower correlation with GO distance for smaller gene expression distances has already been reported in Results. Changes become most obvious for gene expression maps and GO term maps. Mean expression levels differ less over the samples as well as between clusters. Even though higher GO term similarities are still measurable between neighboring clusters, they are less clearly visible. That is, both structural and

functional neighborhood relations are weaker when not excluding low variation genes.

The *cdc28* time series in the Spellman dataset [24] (see Section 2) is identical to the popular Cho data [5] (except for renormalization). This data has been used to verify that our results (based on the *cdc15* time series) do not depend too much on experimental conditions. In general, results turned out to be surprisingly similar if based on the *cdc28* data.

We analyzed the amount of functional (GO) information that is contained in gene expression data and gene expression clusterings. First, correlation between GO distance and Euclidean distance has been found to be significant, especially when selecting genes with a higher variation in expression values. Second, pure Euclidean-based clustering has been compared with random clustering on the basis of pure GO-based measures and visualizations. On the one hand, differences have been found relatively small when using general cluster validity indices (and cluster distances) to measure traditional clustering qualities—small inner and large inter cluster distance. This alone would hint to a rather low amount of GO information in the applied gene expression data. Higher relative differences to random clustering are revealed by cluster distances defined on the SOM feature maps. On the other hand, clearly higher functional similarities between neighboring SOM clusters are induced by a pure structural clustering. This becomes clearly evident by visual inspection of the GO Slim maps (see Fig. 4 for a representative example). On average, an almost twice as high GO similarity was calculated compared to more distant clusters in the SOM.

The influence of parameter m in our co-clustering algorithm, i.e., the m closest clusters in terms of the Euclidean distance which are compared in terms of the GO distance, may be summarized as follows. In general, a higher m value leads to a worse structural, but a better functional clustering. However, at least for moderate m values a better functional clustering (of GO annotations) is possible without a correspondingly worse structural clustering (of gene expression profiles). That is, the clustering quality seems to be relatively less affected in terms of the Euclidean distance than is true for the GO distance. This has been found when measuring traditional qualities—based on a low inner cluster distance and a high inter cluster distance—and when comparing distances between arbitrary and adjacent clusters in the SOM structure. In the latter case, the functional neighborhoods get lost with a smaller m than is the case for the structural neighborhoods.

The closest cluster neighborhoods occur with the worst clustering quality in terms of the GO distance, i.e., the largest inner cluster the smallest inter cluster distances. This is a consequence, at least for the most part, from the lack of an explicit neighborhood adaptation of the functional centers in Algorithm 2, as this exists for the structural centers. Such a mechanism would require that the same term is added to the functional center of both the closest cluster *and* some of its neighbors. Hence, there is not only a trade-off between the influences of different clustering objectives to solve, but between different clustering qualities, too.

As already mentioned in Section 3.2, there is no general optimum setting for co-clustering parameter m. This depends not only on the qualities that are defined by the applied distance and validation measures but also on the quality and noise level of the applied datasets. By clustering genes with a lower variation in expression levels the same clustering quality has been found with a larger m value than is required for higher variation genes. Obviously, the m selected clusters will be more diverse in the latter case. It should be further noted that the influence of the GO information on some clustering qualities is quite non-linear in m. In general, setting $m = 8$ has been found to induce already a quite high GO influence with our data. The different analysis techniques presented in the paper may be used to identify the best setting of m for other gene expression data.

A slight influence of GO information by a small m value (e.g., $m = 2$ and possibly $w < 1$) might be recommended, if the goal of co-clustering is to achieve a better gene expression clustering, i.e., to find groups of genes that are co-regulated in the first place. This is supposed to change the clustering of relatively few genes with borderline profiles only and not necessarily to level out larger errors or noise. Instead, if the purpose of co-clustering is to obtain the most correct grouping and separation of genes according to both clustering objectives—ignoring the (functional) cluster neighborhood—a larger m value may be a better choice, as this has been suggested by the validity index applied here (see Fig. 2).

Besides clustering, databases of gene annotations may be completed by deriving SOM-based prediction models to detect functions of unknown (non-annotated) genes. In doing so, the precision of predictions is improved significantly by using a limited selection of GO terms only. The granularity and the number of these terms may be made dependent on a lower, i.e., less general, GO level than has been done here.

Larger functional regions which are imposed by the neighborhood structure of the SOM clusters give additional confidence in both the correctness (enrichment) of existing cluster annotations and the newly predicted gene functions. Moreover, the positions of the functional regions on the SOM towards each other and the overlaps between them have the potential of revealing new relationships between and hypotheses about genes and gene clusters. We found clear separations as well as overlaps of functional regions in GO Slim maps to be biologically reasonable.

Acknowledgments

CW is supported by the Danish Cancer Society. MB has been supported in part by the Aarhus University Research

Foundation. The authors thank Robert M. MacCallum for permission to use his Perl code of a self-organizing map.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2006.05.001.

References

[1] Azuaje F. A cluster validity framework for genome expression data. Bioinformatics 2002;18(2):319–20.
[2] Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 2000;97(1):262–7.
[3] Cheng J, Cline M, Martin J, Finkelstein D, Awad T, Kulp D, et al. A knowledge-based clustering algorithm driven by Gene Ontology. J Biopharm Stat 2004;14(3):687–700.
[4] Cheng J, Sun S, Tracy A, Hubbell E, Morris J, Valmeekam V, et al. NetAffx Gene Ontology mining tool: a visual approach for microarray data analysis. Bioinformatics 2004;20(9):1462–3.
[5] Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, et al. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 1998;2(1):65–73.
[6] Davies DL. Bouldin, D.W. A cluster separation measure. In: IEEE trans. on pattern analysis and machine intelligence (PAMI); 1979. 1 (4): pp. 224–7.
[7] Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol 2003;4:R7.
[8] Dunn JC. Well separated clusters and optimal fuzzy partitions. J Cybern 1974;4:95–104.
[9] Dwight SS, Balakrishnan R, Christie KR, Costanzo MC, Dolinski K, Engel SR, et al. Saccharomyces Genome Database: underlying principles and organisation. Brief Bioinform 2004;5(1):9–22.
[10] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. PNAS Genet 1998;95:14863–8.
[11] Famili A, Liu G, Liu Z. Evaluation and optimization of clustering in gene expression data analysis. Bioinformatics 2004;20(10): 1535–1545.
[12] Gene Ontology Consortium. (2000) Gene Ontology: tool for the unification of biology. Nat Genet. 25(1):25–9.
[13] Gibbons FD, Roth FP. Judging the quality of gene expression-based clustering methods using gene annotation. Genome Res 2002;12:1574–81.
[14] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531–7.
[15] Hanisch D, Zien A, Zimmer R, Lengauer T. Co-clustering of biological networks and gene expression data. Bioinformatics 2002;18(90001):145–54.
[16] Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. of Int. Conf. on Research in Computational Linguistics (ROCLING), 1997.
[17] Kohonen T. Self-organizing maps. Berlin: Springer; 1997.
[18] Lægreid A, Hvidsten TR, Midelfart H, Komorowski J, Sandvik AK. Predicting Gene Ontology biological process from temporal gene expression patterns. Genome Res 2003;13(5):965–79.
[19] Lin D.An information-theoretic definition of similarity. In: Proc. of Int. Conf. on Machine Learning, 1998.
[20] Lord PW, Stevens R, Brass A, Goble CA. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 2003;19(10):1275–83.
[21] Quackenbush J. Computational analysis of microarray data. Nat Rev Genet 2001;2(6):418–27.
[22] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proc. of Int. Joint Conf. on Artificial Intelligence (IJCAI), 1995, pp. 448–53.
[23] Speer N, Spieth C, Zell A. A memetic co-clustering algorithm for gene expression profiles and biological annotation. In: Proc. of Congress on Evolutionary Computation (CEC), 2004, vol. 2, pp. 1631–8.
[24] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 1998;9(12):3273–97.
[25] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999;96(6):2907–12.
[26] Wang H, Azuaje F, Bodenreider O, Dopazo J. Gene expression correlation and Gene Ontology-based similarity: an assessment of quantitative relationships. In: Proc. of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2004, pp. 25–31.
[27] 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(3):255–65.
[28] Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. Bioinformatics 2001;17(4):309–18.
[29] Universal Protein Resource (UniProt) <http://www.ebi.uniprot.org>.
[30] Saccharomyces Genome Database (SGD) <http://www.yeastgenome.org>.
[31] GO Slim annotations <ftp://ftp.yeastgenome.org/yeast/data_download/literature_curation/go_slim_mapping.tab>.
[32] Spellman expression dataset <http://sgdlite.princeton.edu/download/yeast_datasets/expression/cellCycle.pcl>.