Associative clustering for exploring dependencies between functional genomics data sets

Samuel Kaski, *Senior Member, IEEE,* Janne Nikkilä, Janne Sinkkonen, Leo Lahti, Juha Knuuttila, and Christophe Roos

- S. Kaski and J. Nikkilä are with University of Helsinki, Department of Computer Science, P.O. Box 68, FI-00014 University of Helsinki, Finland. E-mails: samuel.kaski@cs.helsinki.fi, janne.nikkila@hut.fi. During part of the work they were with Helsinki University of Technology.
- J. Sinkkonen and L. Lahti are with Helsinki University of Technology, Neural Networks Research Centre, P.O. Box 5400, FI-02015 HUT, Finland. E-mail: janne.sinkkonen@hut.fi, leo.lahti@hut.fi.
- J. E. A. Knuuttila is with Neuroscience Center, P.O. Box 56, FI-00014 University of Helsinki, Finland. E-mail: Juha.Knuuttila@helsinki.fi.
- C. Roos is with Medicel Oy, Huopalahdentie 24, FI-00350 Helsinki, Finland. E-mail: christophe.roos@helsinki.fi.

Abstract

High-throughput genomic measurements, interpreted as co-occurring data samples from multiple sources, open up a fresh problem for machine learning: What is in common in the different data sets, that is, what kind of statistical dependencies there are between the paired samples from the different sets. We introduce a clustering algorithm for exploring the dependencies. Samples within each data set are grouped such that the dependencies between groups of different sets capture as much of pairwise dependencies between the samples as possible. We formalize this problem in a novel probabilistic way, as optimization of a Bayes factor. The method is applied to reveal commonalities and exceptions in the expression of organisms, and to suggest regulatory interactions, in the form of dependencies between gene expression profiles and regulator binding patterns.

Index Terms

Biology and genetics, Clustering, Contingency table analysis, Machine learning, Multivariate statistics

I. Introduction

Assume two data sets with *co-occurring* samples, that is, samples coming in pairs $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x}$ belongs to the first set and $\mathbf{y}$ to the second set. In this paper both $\mathbf{x}$ and $\mathbf{y}$ are gene expression profiles or other multivariate real-valued genomic measurements about the same gene. The general research problem is to find *common properties* in the set of pairs; statistically speaking, the goal is to find statistical dependencies between the pairs.[1]

In this paper we search for dependencies expressible by clusters. The standard unsupervised clustering methods, reviewed for gene expression clustering for instance in [32], aim at finding clusters where genes have similar expression profiles. Our goal is different: to cluster the $\mathbf{x}$ and the $\mathbf{y}$ separately such that the dependencies between the two clusterings capture as much as possible of the statistical dependencies between two sets of clusters. In this sense

[1] The fundamental difference from searching for differences between data sets [18], where the relative order of the samples within the two sets is not significant, both sets are within the same space, and the goal is to find differences between data *distributions*, is that our data are paired and we search for commonalities between the pairs of *samples* that can have different variables (attributes) and different dimensionalities.

the clustering is *associative*; it finds associations between samples of different spaces. The research problem will be formalized in Section II.

The problem of searching for common properties in two or more paired data sets differs from classic machine learning problems, commonly categorized into unsupervised and supervised. Supervised learning targets at finding classes (in classification) or predicted values of a variable (in regression). In probabilistic terms the goal is to build a good model for the distribution $p(\mathbf{y}|\mathbf{x})$ while in the kind of dependency modeling discussed in this paper the goal should be symmetric. Basic unsupervised learning, on the other hand, is symmetric in a trivial sense: All variation of one variable—be it $\mathbf{x}$, $\mathbf{y}$, or the combination $(\mathbf{x}, \mathbf{y})$—is modeled, and there is no mechanism for separating between-data-set variation from within-data-set variation. Common to both kinds of learning, and indeed to all machine learning, is model fitting: A model parameterized by $\boldsymbol{\theta}$ is fitted to the data.

A different kind of problem, to be addressed in this paper, is modeling only the variation in $\mathbf{x}$ and $\mathbf{y}$ that is *common* to both variables. In other words, we search for *dependencies* between the $\mathbf{x}$ and $\mathbf{y}$. This symmetric goal has traditionally been formalized as maximizing the dependency between two representations, $\hat{\mathbf{x}} \equiv \mathbf{f}_x(\mathbf{x}; \boldsymbol{\theta}^x)$ and $\hat{\mathbf{y}} \equiv \mathbf{f}_y(\mathbf{y}; \boldsymbol{\theta}^y)$, of $\mathbf{x}$ and $\mathbf{y}$, respectively. A familiar example is canonical correlation analysis [24] where both the $\mathbf{f}_x$ and $\mathbf{f}_y$ are linear projections, and the data are assumed to be normally distributed. This idea has been generalized to nonlinear functions [4], and to finding clusters of $\mathbf{x}$ informative of a nominal-valued y [3], [37]. It has been formalized in the information bottleneck framework [40], [44], resulting in efficient algorithms for two nominal-valued variables [35], [41].

Symmetric dependency modeling with non- or semiparametric methods (such as clustering) is a natural way of formalizing the search for commonalities in co-occurring data sets, when one is not able or willing to postulate a detailed parametric model *a priori*. Such situations are common in modern data-driven functional genomics: Microarray-based high-throughput measurement techniques make it possible to test broad hypotheses, related, for example, to organism-wide differences in response, or to functions of a gene over a range of organisms. Mining the data, stored in community-resource databanks, for new hypotheses is fruitful as well. In data mining the search for dependencies between data sets is a considerably better-defined target than the common, unsupervised search for clusters and other regularities.

We study two cases of symmetric dependency modeling: search for regularities and differences in expression of orthologous genes in different organisms, and search for regulatory interactions between expression and transcription factor binding patterns. More generally, we argue that once a research goal can be dressed into a search for dependencies between data sets, our approach is a well-defined middle ground between purely hypothesis-driven research, for which hypotheses must be available, and purely exploratory research, where the task is often ill-defined.

Analogically to the two linear projections in canonical correlation analysis, we use two sets of clusters as the representations in the dependency search. Clusters are more flexible than linear projections, and they have a definite role in exploratory data analysis, that is, in "looking at the data": Clustering reveals outliers, finds groups of similar data, and simply compresses numerous samples into a more manageable and even visualizable summary. Clusters and other kinds of unsupervised models are of particular importance as the first step of microarray data analysis, where data are often noisy and even erroneous, and in general not well-known *a priori*.[2]

For microarray data, the existing dependency-searching techniques have two deficiencies. First, mutual information, the dependency measure that they maximize, is defined for probability distributions which in turn need to be estimated from samples. The separate estimation stage with its own optimality criteria will introduce uncontrollable errors to the models. The errors are negligible for asymptotically large data sets but non-negligible for many real-life sets. We will directly define a dependency measure for data instead of distributions, and justify it by combinatorial and Bayesian arguments. For asymptotically large data sets the dependency measure becomes mutual information, and can therefore be viewed as a principled alternative to mutual information for finite data sets.

The second shortcoming has been that the models are not applicable to symmetric dependency clustering of *continuous* data. While a trivial extension of existing continuous-data methods may seem sufficient, a conceptual change is actually required. Existing finite-data formulations either maximize the likelihood $p(\mathbf{y}|\mathbf{x})$ of one data set, say $\mathbf{y}$, given $\mathbf{x}$, or

[2] This very legitimate and necessary use of clustering in the beginning of the research process should not be confused with the widespread use of clusterings as a general-purpose tool in all possible research tasks, which could better be solved by other methods.

maximize the symmetric joint likelihood for $p(\mathbf{x}, \mathbf{y})$. Neither of these approaches is dependency modeling: Conditional models are asymmetric, while joint density models represent all variation in $\mathbf{x}$ and $\mathbf{y}$ instead of common variation, and therefore do not even asymptotically reduce to mutual information. A solution we present in this paper is to use a hypothesis comparison approach which translates to a Bayes factor cost function.

Bayesian networks, used also as models of expression regulation [16], [36], are models for the joint density of all data sources. In these models the structure of dependencies between variables is, at least to some extent, fixed in advance. To a degree dependencies can be learned from data, but learning is hard and data-intensive. Our approach complements Bayesian networks in two ways. First, it is more exploratory and assumption-free because no dependency structure is imposed, except the one implied by cluster parameterization and division of the data set. Second, as joint distribution models Bayesian networks represent not only the common variation between the data sets but partly also the unique variation within each data set. In this sense, the representations they produce are compromises for the task of modeling the between-set variation.

From the biological perspective, the advantages of clustering by maximizing dependency between two sources of genomic information are at least two-fold. First, the new problem setting makes it possible to formulate new kinds of hypotheses about the dependency of the sources, not possible with conventional one-source clusterings. Such hypotheses are sought in the orthologous genes application in Section V. Second, mining for regularities in the common properties of two data sets is a more constrained problem that mining for any kinds of regularities within either of them. Hence, assuming the sets are chosen cleverly, the results are potentially better targeted. Our hypothesis is that there will be less false positives in the discovered regulatory interactions when expression and transcription factor binding are combined in a dependency maximizing way, compared to one-source clusterings. We will study the interactions in Section VI.

II. Associative clustering

The abstract task solved by associative clustering (introduced in the preliminary paper [39]) is the following: cluster two sets of data, with samples $\mathbf{x}$ and $\mathbf{y}$, each separately, such that (i) the clusterings would capture as much as possible of the dependencies between pairs

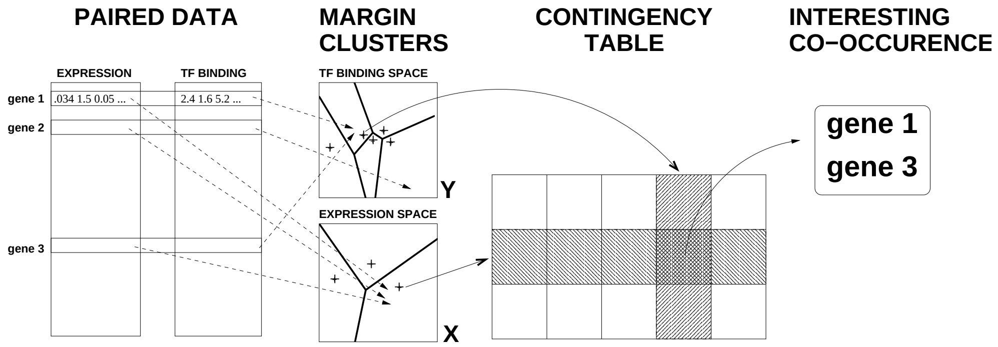


Fig. 1. Associative clustering (AC) in a nutshell. Two data sets are clustered into Voronoi regions. The Voronoi regions are defined in the standard way as sets of points closest to prototype vectors, but the prototypes are not optimized to minimize a quantization error but by the AC algorithm. In this example, the data sets are gene expression profiles and transcription factor (TF) binding profiles. A one-to-one correspondence between the sets exist: Each gene has an expression profile and a TF binding profile. As each gene falls to a TF cluster and to an expression cluster, we get a contingency table by placing the two sets of clusters as rows and columns, and by counting genes falling to each combination of an expression cluster and a TF cluster. Rows and columns, that is, the Voronoi regions defined within each data set respectively, are called *margin clusters*, while the combinations corresponding to the cells of the contingency table are called *cross clusters*. *Associative clustering* by definition finds Voronoi prototypes that maximize the dependency seen in the contingency table. Voronoi regions are representations for the data sets just as the linear combinations are in canonical correlation analysis. In both cases, dependency between the two parameterized representations is maximized. Maximization of dependency in a contingency table results in a maximal amount of surprises, counts not explainable by the margin distributions. The most surprising cross clusters with a very high or low number of genes potentially give rise to interesting interpretations. Reliability is assessed by the bootstrap.

of data samples $(\mathbf{x}, \mathbf{y})$, and (ii) the clusters would contain (relatively) similar data points. The latter is roughly a definition of a cluster.

Figure 1 gives a brief overview of the method. For paired data $\{(\mathbf{x}_k, \mathbf{y}_k)\}$ of real vectors $(\mathbf{x}, \mathbf{y}) \in \mathbf{x} \times \mathbf{y} \subset \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, we search for partitionings $\{V_i^{(x)}\}$ for $\mathbf{x}$ and $\{V_j^{(y)}\}$ for $\mathbf{y}$. The partitions can be interpreted as clusters in the same way as in K-means; they are Voronoi regions parameterized by their prototype vectors $\mathbf{m}_i$. The $\mathbf{x}$ belongs to $V_i^{(x)}$ if $\|\mathbf{x} - \mathbf{m}_i^{(x)}\| \leq \|\mathbf{x} - \mathbf{m}_{i'}^{(x)}\|$ for all i', and correspondingly for $\mathbf{y}$.

A. Bayes Factor for Measuring Dependency between Two Sets of Clusters

The dependency between two cluster sets, indexed by i and j, can be measured by mutual information if the joint distribution p_{ij} is known. If only a *contingency table* of co-occurrence frequencies n_{ij} computed from a finite data set is available, mutual information computed from the empirical distribution would be a biased estimate. A *Bayes factor*, to be introduced below, has the advantage of properly taking into account the finite size of the data set while still being asymptotically equivalent to mutual information. Bayes factors have classically been used as dependency measures for contingency tables (see, e.g., [20]) by comparing a model of dependent margins to another model for independent margins. We will use the classical results as building blocks to derive an optimizable criterion for associative clustering; the novelty here is that the Bayes factor is *optimized* instead of only being used to measure dependency in a fixed table. The categorical variables defining the rows and columns of the contingency table are defined by the Voronoi regions. They are parameterized by the cluster prototypes which are optimized to maximize the Bayes factor.

The Bayes factor compares two alternative models, one describing a contingency table where the margins are dependent and the other a table with independent margins. The clusters are then tuned to make the dependent model describe the (contingency table) data better than the independent model, which can be interpreted as maximizing dependency.

In general, frequencies over the cells of a contingency table can be assumed to be multinomially distributed. The model M_I of *independent margins* assumes that the multinomial parameters over cells are outer products of posterior parameters at the margins: $\theta_{ij} = \theta_i \theta_j$. The model M_D of *dependent margins* ignores the structure of the cells as a two-dimensional table and samples cell-wise frequencies directly from a table-wide multinomial distribution θ_{ij}. Dirichlet priors are set for both the margin and the table-wide multinomials.

Maximization of the Bayes factor

$$BF = \frac{p(\{n_{ij}\}|M_D)}{p(\{n_{ij}\}|M_I)}$$

with respect to the margin clusters then gives a contingency table where the margins are maximally dependent, that is, the table is as far from the product of independent margins as possible. In associative clustering, the counts are influenced by the parameters of the Voronoi regions. The BF is maximized with respect to these parameters.

After marginalization over the multinomial parameters, the Bayes factor takes the form (derivation in the technical report [38])

$$BF = \frac{\prod_{ij} \Gamma(n_{ij} + n^{(d)})}{\prod_i \Gamma(n_{i\cdot} + n^{(x)}) \prod_j \Gamma(n_{\cdot j} + n^{(y)})} \, , \tag{1}$$

where $n_{i\cdot} = \sum_j n_{ij}$ and $n_{\cdot j} = \sum_i n_{ij}$ express the margins. The hyperparameters $n^{(d)}$, $n^{(x)}$, and $n^{(y)}$ arise from Dirichlet priors. We have set all three hyperparameters to unity, which makes the *BF* equivalent to the hypergeometric probability classically used as a dependency measure of contingency tables. For large data set sizes N the logarithmic Bayes factor approaches mutual information of the distribution $p_{ij} = n_{ij}/N$ with margins $p_i = n_{i\cdot}/N$ and $p_j = n_{\cdot j}/N$ [38]:

$$\frac{1}{N}\log BF = \sum_{i,j} p_{ij}\log\frac{p_{ij}}{p_i p_j} - \log N + 1 + \mathcal{O}\left(\frac{1}{N}\log N\right) = \hat{I}(I,J) - \log N + 1 + \mathcal{O}\left(\frac{1}{N}\log N\right) \;, \tag{2}$$

where $\hat{I}(I,J)$ is the mutual information between the categorical variables I and J having cluster indices as their values.

B. Optimization of AC

The Bayes factor (1) will be maximized with respect to the Voronoi prototypes. The optimization problem is combinatorial for hard clusters, but gradient methods are applicable after the clusters are smoothed. Gradients are derived in a technical report [38]. An extra trick, found to improve the optimization in the simpler case where one of the margins is fixed [27], is applied here as well: The denominator of the Bayes factor is given extra weight by introducing constants $\lambda^{(\cdot)}$. A choice of $\lambda^{(\cdot)} > 1$ introduces to the cost function a regularizing term that for large sample sizes approaches margin cluster entropy, and thereby in general favors solutions with uniform margin distributions.

The smoothed *BF*, here denoted by *BF'*, is then optimized with respect to the cluster prototypes $\{\mathbf{m}\}$ by a conjugate-gradient algorithm (for a textbook account see [2]). We have

$$\begin{aligned}\log BF' = &\sum_{ij}\log\Gamma\left(\sum_k g_i^{(x)}(\mathbf{x}_k)g_j^{(y)}(\mathbf{y}_k) + n^{(d)}\right)\\ &- \lambda^{(x)}\sum_i \log\Gamma\left(\sum_k g_i^{(x)}(\mathbf{x}_k) + n^{(x)}\right) - \lambda^{(y)}\sum_j \log\Gamma\left(\sum_k g_j^{(y)}(\mathbf{y}_k) + n^{(y)}\right) \;,\end{aligned} \tag{3}$$

where

$$g_i^{(x)}(\mathbf{x}) \equiv Z^{(x)}(\mathbf{x})^{-1}\exp\left(-\|\mathbf{x} - \mathbf{m}_i^{(x)}\|^2/\sigma_{(x)}^2\right) \;,$$

and similarly for $g^{(y)}$. The $g(\cdot)$ are the smoothed Voronoi regions at the margins. The $Z(\cdot)$ is set to normalize $\sum_i g_i^{(x)}(\mathbf{x}) = \sum_j g_j^{(y)}(\mathbf{y}) = 1$. The parameters σ control the degree of smoothing of the Voronoi regions.

The gradient of $\log BF'$ with respect to an X-space prototype $\mathbf{m}_i^{(x)}$ is

$$\nabla_{\mathbf{m}_i^{(x)}} \log BF' = \frac{1}{\sigma_{(x)}^2} \sum_{k,i'} (\mathbf{x}_k - \mathbf{m}_i^{(x)}) g_i^{(x)}(\mathbf{x}_k) g_{i'}^{(x)}(\mathbf{x}_k) \left(L_i^{(x)}(\mathbf{y}_k) - L_{i'}^{(x)}(\mathbf{y}_k) \right) ,$$

where

$$L_i^{(x)}(\mathbf{y}) \equiv \sum_j \Psi \left(\sum_k g_i^{(x)}(\mathbf{x}_k) g_j^{(y)}(\mathbf{y}_k) + n^{(d)} \right) g_j^{(y)}(\mathbf{y}) - \lambda^{(x)} \, \Psi \left(\sum_k g_i^{(x)}(\mathbf{x}_k) + n^{(x)} \right) ,$$

and for $\mathbf{y}$ accordingly. In the gradient, $\Psi(\cdot)$ is the digamma function.

In summary, the optimization of AC proceeds as follows: (i) Parameters $\{\mathbf{m}^{(x)}\}$ and $\{\mathbf{m}^{(y)}\}$ are independently initialized by choosing the best of several (here: three) K-means runs initialized randomly. (ii) On the basis of experience with other data sets, we choose $\lambda^{(\cdot)} = 1.2$. (iii) Parameters $\sigma_{(\cdot)}$ are chosen by running the algorithm for half of the data and testing on the rest. (iv) The $\{\mathbf{m}^{(x)}\}$ and $\{\mathbf{m}^{(y)}\}$ are optimized with a standard conjugate gradients algorithm, using $\log BF'$ as the target function. Gradients of the $\mathbf{m}$-parameters plugged into the algorithm are shown above. The reported results are from cross-validation runs.

In one-margin optimization with clusters in the other margin fixed, the smoothing trick performs equivalently to or better than simulated annealing [27]. Also note that smoothing is for optimization only: Results are evaluated with *BF*, which translates to having crisp clusters.

C. Uncertainty in clustering

Our use of Bayes factors is different from their traditional use in hypothesis testing, cf. [20]. In AC we do not test any hypotheses but maximize the Bayes factor to explicitly find dependencies. This leaves the uncertainty of the solution open.

A widely used "light-weight" (compared to posterior computation) method to take into account the uncertainty in clustering is bootstrap [12], [21]. As in [29], we use bootstrap to produce several perturbed clusterings. We wish to find cross clusters (contingency table cells) that signify dependencies between the data sets and are reproducible.

Reproducibility of the found dependencies will be estimated from the bootstrap clusterings as follows.

First, we define what we mean by a significantly dependent cross cluster within a given AC-clustering. The optimized AC model provides a way of estimating how unlikely a cross cluster is, *given that the margins are independent*. For this purpose several (1000 or more) data sets of the same size as the observed one are generated from the marginals of the contingency table (i.e., under the null hypothesis of independence). The cross clusters with the observed amount of data more extreme than that observed by chance with probability 0.01 or less (Bonferroni corrected with the number of cross clusters), are defined to be *significantly dependent cross clusters*.

Next, the two criteria, dependency and reproducibility, will be combined by evaluating how likely it is for each gene pair to occur within the same significantly dependent cross cluster in bootstrap (this is analogous to [29]). The result, interpreted as a similarity matrix, will finally be summarized by hierarchical clustering.

Please note that we do not expect to find dependencies for all genes in the whole data sets, since with noisy genomic data that would hardly be possible. In other words, we are interested in finding the most dependent, robust *subsets of the data*. This is exactly what the final gene clusters from bootstrapped, most dependent cross clusters provide.

D. Extremity of the clusters

In the yeast case studies we evaluate which cross clusters are exceptional by their expression or TF binding profile. For determining the extremity of the observed within-cluster profiles, for each of them 10,000 random sets of genes were first sampled, each of the same size as the cluster under study. We then computed within-cluster average profiles for the observed cluster as well as for the simulated ones. A part of the observed profile was denoted as extreme if it was lower or higher in value than all the simulations.

III. Reference methods

First we need a baseline method to give a lower bound for the results. For AC, it should not optimize the dependency of the clusters but only perform conventional clustering, while being as similar to AC as possible in other respects. In this work the baseline method will be independent K-means clusterings in both data spaces, since K-means is also prototype-based clustering for continuous data like AC. For more detailed description and references of K-means see for example [7].

We compare AC to the information bottleneck (IB) methods [17], [44]. The main problem with IB in our setting is the continuous nature of our data: IB works on nominal-valued data. Here we discretize the data first by K-means, resulting in a new algorithm called here K-IB. For discrete data, the closest alternative to AC among information bottleneck methods would be symmetric two-way IB [17]. Our sequential implementation is based on [40].

We first quantize the vectorial margins $\mathbf{x}$ and $\mathbf{y}$ separately by K-means, without paying attention to possible dependencies between the two margins. This results in two sets of margin partitions which span a large, sparse contingency table that can be filled with frequencies of training data pairs $(\mathbf{x}_k, \mathbf{y}_k)$. The number of elementary Voronoi regions is chosen by using a validation set. In the second phase, the large table is compressed by standard IB to the desired size by aggregating the atomic margin clusters. In this stage, joins at the margins are made with the symmetric sequential algorithm [40] to explicitly maximize the dependency of margins in the resulting smaller contingency table.

The final partitions obtained by the combination of K-means and IB are of a very flexible form, and therefore the method is expected to model the dependencies of the margin variables well. As a drawback, the final margin clusters will consist of many atomic Voronoi regions, and they are therefore not guaranteed to be particularly homogeneous with respect to the original continuous variables ($\mathbf{x}$ or $\mathbf{y}$). Interpretation of the clusters may then be difficult. Our empirical results support both the good performance of K-IB and the non-localness of the resulting clusters.

IV. Validation of Associative Clustering

A. *Demonstration with artificial data*

Figures 2 and 3 demonstrate two key properties of AC with as simple artificial data sets as possible.

The clusters focus on modeling those regions of the margin data spaces, that is, those subsets of data, where the co-occurring pairs $\mathbf{x}$ and $\mathbf{y}$ are dependent. This is clearly visible as the high-density area of cross clusters in Figure 2.

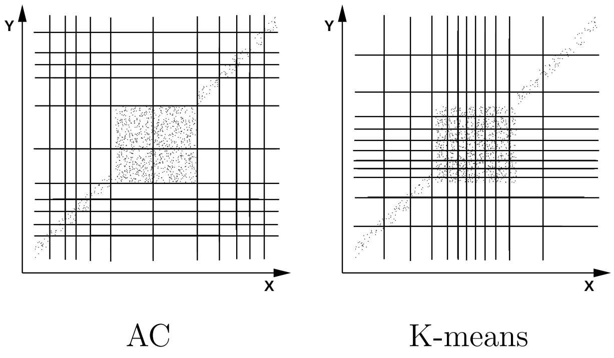


Fig. 2. Associative clustering concentrates on dependent subsets of data. Here both margin spaces, denoted by **X** and **Y**, are 1-dimensional, and the figure shows a scatterplot of the data (dots on the plane where **X** and **Y** are the axes). Cluster borders in the **X**-space are shown with the vertical lines and cluster borders in the **Y**-space with horizontal lines. The resulting grid of so-called cross clusters then corresponds to the contingency table; the number of dots within each grid cell gives the amount of data in a contingency table cell. The AC cells are sparse in the bulk of independent data in the middle and denser on the sides where the **X** and **Y** are dependent. K-means, in contrast, focuses on modeling the bulk of the data in the middle. (For this data set AC has lots of local maxima.)

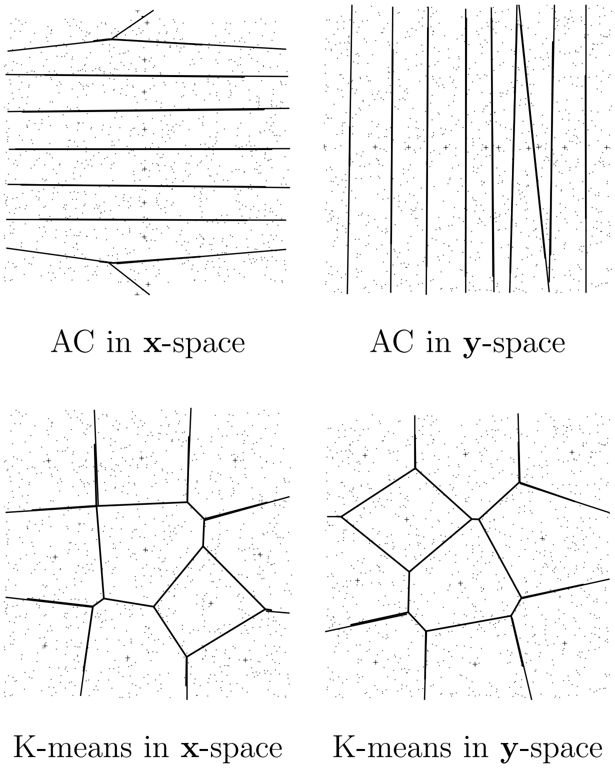
AC neglects variation that is irrelevant to the dependencies between $\mathbf{x}$ and $\mathbf{y}$. In Figure 3, the AC clusters have effectively become defined by only the relevant one of the two dimensions. By contrast, standard clustering methods, such as K-means, model variation in both dimensions.

B. Validation of bootstrapped AC analysis with real data

Especially in bioinformatics it is often challenging to test new methods since there rarely exists any ground truth, that is, known correct answers. We validated the (bootstrapped) AC approach by searching for dependencies between data sets containing known, real-world duplicate measurements that should be more dependent than random pairs.

Expression profiles of orthologous man-mouse gene pairs with unique LocusIDs were derived from a public source [43] (`http://expression.gnf.org/data_public_U95.gz` http://expression.gnf.org/data_public_U74.gz) using the HomoloGene [46] database and Affymetrix annotation files. The expression measurements include 46 human and 45 mouse arrays covering a wide range of tissues and cell-lines. For 21 of the tissues, expression values were available for both species.

We have derived two different data sets from the original data: (1) a larger one for this



AC in **x**-space

AC in **y**-space

K-means in **x**-space

K-means in **y**-space

Fig. 3. Associative clustering focuses on modeling the variation that is relevant to dependencies between the data sets. Both of the margin spaces are here 2-dimensional, and the data has been constructed such that the vertical dimension of the **x**-space is dependent on the horizontal dimension in the **y**-space. All other variation is uniform noise. Lines are approximate cluster borders (Voronoi borders), and the small crosses are the prototype vectors. Associative clustering neglects the irrelevant variation in both margin spaces and models the relevant, dependent variation. In contrast, K-means, as all purely unsupervised clusterings, models all the variation including noise.

validation study, with known ground truth in the form of naturally multiplicated genes, and (2) a smaller one for the actual analysis without any multiplicated genes (presented in Section V).

Due to technicalities related to the Affymetrix oligonucleotide array platform, in the original data sets [43] one gene (LocusID) may have multiple expression profiles. In the verification data set these profiles were considered as independent samples, resulting in a total of 4500 gene expression profile pairs. These "duplicate orthologous genes," representing the same sequence-level similarity between the species, should co-occur in the significantly dependent cross clusters (see Section II-C) more often than randomly chosen orthologous genes, and, since AC should model dependencies more effectively than K-means, also more often than in the cross clusters produced by K-means.

The validation study was carried out by exactly the same procedures as we will use in the rest of the experiments of the paper, to validate the setting.

The number of clusters was chosen to be such that each cross cluster would on average contain roughly 10 data points. For the verification set this translates to 19 clusters in both margin spaces. We sampled 100 bootstrap data sets, computed AC for each, got 100 different contingency tables, and from these we computed a similarity matrix for the genes as described in Section II-C.

The optimization parameter σ was chosen by leaving half of the data for validation.

We then tested with a rank sum test whether the similarity distribution of the known duplicates is different from the similarity distribution of all the other genes. In AC the known duplicates turned out to co-occur unexpectedly frequently in dependent cross clusters (rank sum test; $p < 2.2 \times 10^{-16}$).

Compared to K-means, AC detected connections of the multiple ortholog profiles statistically significantly more often (sign test, $p < 0.001$). These two results support the validity of AC in finding dependent subsets of data better than standard unsupervised clustering.

V. Experimental results: Dependencies between man and mouse

Functions of human genes are often studied indirectly, by studying model organisms such as the mouse. An underlying assumption is that so-called orthologous genes, that is, genes with a common evolutionary origin, have similar functional roles in both species. Exploration of dependencies (regularities and irregularities) in functioning of orthologous genes helps in assessing to which extent this assumption holds. In practice, gene pairs are defined as putative orthologs based on sequence similarity, and we seek for regularities and irregularities in their expression by associative clustering.

Exceptional level of functional conservation of an orthologous gene group may indicate important physiological similarities, whereas differentiation of function may be due to significant evolutionary changes. Large-scale studies on orthologous genes may ultimately lead to a deeper understanding of what makes each species unique. (For related approaches, see, e.g., [6], [9], [11], [14], [30]).

A. Data and experiments

In the original data [43], multiple expression profiles may correspond to one gene. In Section IV-B they were used for validating the methods, whereas in this section we use a single representative profile for each gene. The profiles corresponding to a same gene are averaged after discarding weakly correlating ($r < 0.65$) profiles of the same gene, when multiple measurements from incomplete or potentially non-specific probe sets are available. This results in a set of 2818 orthologous gene pairs with unique LocusIDs.

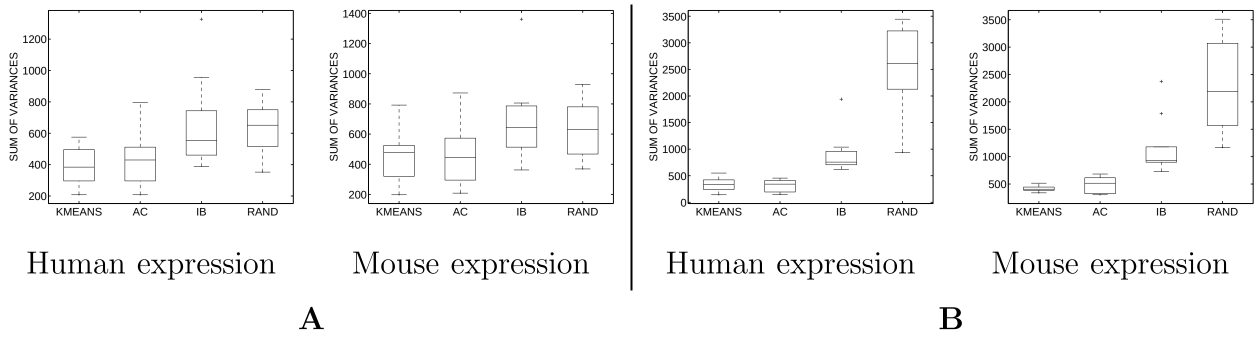
B. Quantitative comparisons of the methods

A dependency-maximizing clustering method should (i) find dependencies and (ii) represent the results as homogeneous clusters. We compared AC to a baseline method that does not search for dependencies at all, that is, separate K-means for both mouse and man, and to symmetric IB following a discretization with K-means (see Section III). The both σ:s of AC and the number of initial K-means clusters for IB were chosen using a validation set as in Section IV-B.

AC produced significantly more dependent clusters than standard K-means clustering (10-fold cross-validation, paired t-test with d.f.=9; $p < 0.001$). All methods were run in each fold from three different intializations, of which the best result according to each method's own cost function was selected. Averaged log-BF costs were -52.9 and -115.8 for AC and K-means, respectively. However, cluster homogeneity was not significantly reduced by focusing on dependency modeling (at the $p < 0.05$ significance level). Differences of the methods in cluster homogeneity have been visualized in Figure 4.

K-IB produced significantly ($p < 0.001$) more dependent clusterings (log-BF=10.24 on average over cross-validation folds) than AC and K-means. On the other hand, cross clusters from AC studies are significantly more homogeneous than those of K-IB and random clustering ($p < 0.002$). The measure of homogeneity (actually dispersion) was the sum of intra-cluster variances.

In summary, as expected, AC extracts more dependencies than K-means and the clusters are more homogeneous (and hence easier to interpret) than those of K-IB. K-IB is a good method for searching for dependencies if homogeneity is not essential.



Human expression Mouse expression

A

Human expression Mouse expression

B

Fig. 4. Dispersion of **A** margin clusters and **B** cross clusters in mouse-man studies. AC produces clusters that are comparable to K-means, whereas the clusters of K-IB are more dispersed (significantly in **B**). RAND is a kind of an upper limit for cluster dispersion, obtained by randomly assigning samples to clusters.

C. Biological results: findings of mice and men

Bootstrapped AC produces a similarity matrix for the genes, computed from the co-occurrence frequencies of genes in the AC cross clusters. The matrix is in this section summarized with simple hierarchical clustering, and a set of most homogeneous gene clusters is extracted by cutting the dendrogram at a specific cut-off level and discarding genes belonging to clusters smaller than 3 genes.

As the most reliable dependencies, produced by a high cut-off, are expected to be relatively trivial findings of similar behavior of orthologous genes in mouse and man, we set the threshold lower to include some unexpected findings as well. The (arbitrary) cut-off limit was set to include clusters with average co-occurrence frequency larger than 80% (of the bootstrap samples). This resulted in 139 orthologous gene pairs in 31 clusters.

1) Overall regularities in ortholog expression: Many orthologous genes are expected to be functionally similar, and similarity can, at its simplest, be measured by correlation. Weak correlation of expression of orthologous genes suggests differentiated gene function (or heavy noise), whereas strong correlation is an indication of functional conservation. To some extent, a global trend exists in our data: Median correlation of expression profiles of orthologous man-mouse gene pairs in the common 21 tissues is 0.33. It is expected that this trend dominates the AC analyses concerning unexpectedly common expression trends (large cross clusters) as well. Indeed, the more similar (highly correlating) the expression profiles of an orthologous gene pair are, the more often it tends to be located in an unexpectedly large cross cluster. This was measured by correlating the occurrence frequency with the correlation between the orthologs, and the resulting correlation coefficient $r = 0.41$ suggests that AC indeed is capable of detecting the simple tendency of the orthologs to depend linearly.

Weakly or negatively correlating orthologs are the other extreme; they are kinds of outliers and tend to be located in exceptionally small cross clusters. Expression similarity correlates negatively ($r = -0.38$) with frequency of occurrence in small cross clusters.

2) General functional trends of dependent genes: Orthologous genes are often functionally similar, although some deviation may have occurred in the course of evolution. Orthologous gene groups with exceptional functional conservation could be expected to be of a specific importance for species survival.

Such a cross-species feature is likely to contribute to dependencies in the data, and should be detected in AC analyzes. A straightforward approach to study such functional trends is to check enrichment of Gene Ontology (GO) [1] categories among the most dependent genes.

The most enriched GO categories among the genes showing remarkable dependency (average co-occurrence level $\geq 80/100$, minimum cluster size 3) were ribosomal categories (all findings having EASE score with the conservative Bonferroni correction < 0.05 are listed; EASE [23] is a program that annotates the given gene list based on GO and calculates various statistics for it). The three most significantly enriched GOs, for both species, were cellular component categories "cytosolic ribosome (sensu Eukarya)" and "ribosome," and the molecular function category "structural constituent of ribosome." Also the biological process "transmission of nerve impulse," was enriched for both species. For human, also the "eukaryotic 48S initiation complex," "cytosolic small ribosomal subunit (sensu Eukarya)," "small ribosomal subunit," and "synaptic transmission" categories were enriched.

Dependency structure of data is mostly explained by genes from these categories. A natural explanation for the enrichment of ribosomal functions in large cross clusters is that they often require coordinated effort of a large group of genes, and function in cell maintenance tasks that are critical for species survival. High conservation of such genes has been suggested also in earlier studies (see, e.g., [26]). The current result is an additional indication of exceptional conservation of ribosomal genes and of their crucial role for the cellular functions of an organism.

By contrast, enrichment of the "transmission of nerve impulse" category is somewhat

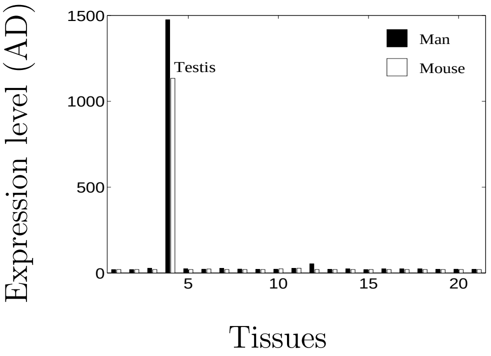


Fig. 5. Average expression profiles of the genes within the cluster showing the highest correlation between mouse and man. Only the 21 tissues which were measured for both species are shown for clarity. No genes were expressed ($AD < 200$) in the remaining tissues. Tissue list is in the Appendix.

surprising and worth more careful studies. It is interesting to note that such genes seem to contribute more to commonalities in the data than genes with other conserved functions. No straightforward biological explanation for this phenomenon could be found so far.

3) Examples of finer-scale regularities: Minor regularities are revealed by the individual clusters. In addition to conserved expression, AC can potentially reveal orthologs with functional deviation.

We used median correlation as a rough measure to order the clusters, and picked two clusters: one with the highest (suggesting preservation of function) and one with the lowest (suggesting differentiation of function) median correlation as examples.

The cluster with the highest median ortholog correlation contained three genes with strongly testis-specific expression (LocusID pairs 8852-11643, 11055-53604, 1618-13164; Fig. 5). Literature studies confirmed that the function of these genes is related to reproduction. Disturbances in the function of the last gene are known to cause infertility although its functions are otherwise not well known.

Although the presence of strongly correlated orthologs in the most dependent clusters of the two species is not surprising as such, the strong relationship of the three genes suggests a possibly unknown functional link.

The clusters having salient regularities suggest interactions: The gene products may have physical interaction, they may share a common pathway, or they may otherwise be responsible of similar biological functions. Even correlated expression within a single species is

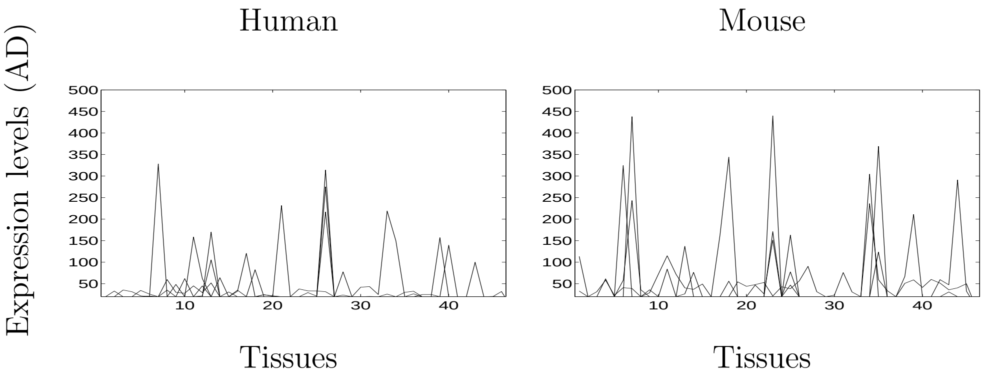


Fig. 6. Expression profile plots of the genes in the cluster with weakest median correlation between the orthologs. Since the correlation is low, no immediate relationships are visible. The cluster is very reliable, however, and hence the orthologs probably share some unexpected higher-order dependency.

known to be a valuable cue for such interactions (see, e.g., [8], [13], [19]), and preservation of co-expression in evolution is an even stronger hint. Moreover, such “conserved correlations” have also been suggested to be useful in confirming orthologous relationships between genes [15].

Low between-species correlation in a cluster with five genes suggests differentiated gene function (Fig. 6). Three of the genes are known to be related to embryonic development, and three are transcription factors. We were not able to find an interpretation for the cluster from the literature. It is reliable, however, and hence potentially interesting; the genes were clustered together in an exceptional cross cluster in over 80 out of 100 bootstrap samples. Our data is from adults, in which the embryonic genes may have unknown functions.

4) Functionally exceptional orthologs: Outliers, that is, genes having peculiarities in their function, can be sought by computing how often they end up in an unexpectedly small cross cluster in the bootstrap. Such genes are comparatively rare; only 1.5% of the orthologs end up in an exceptionally small cross cluster with a frequency of $\geq$ 50%. Such exceptional orthologs tend to correlate weakly or negatively, and potentially hint at differentiated gene function. Note that AC takes more than correlation into account as only 3 of the 43 found orthologs are among the 43 most weakly correlating orthologs. Hence, these exceptional genes could not have been found based on the correlation analysis alone.

Enrichment of certain GO categories among such exceptional orthologs would indicate functionalities that are more often differentiated between species. Interestingly, closest to significant enrichment were the “secretion” category with its subcategory “protein secretion,” and the “signal transduction” category with subcategories of “cell communication,” “signal transduction,” and “cell surface receptor linked signal transduction” for human, and “cell communication,” and “G-protein coupled receptor protein signaling pathway” for mouse. These categories have EASE score of < 0.05 without Bonferroni correction. With Bonferroni correction, the enrichment is not significant, however.

To some extent the secretion categories above could be related to the overall signaling phenomena. The protein secretion category fits well into this picture since many of these signaling pathway initiators are in fact secreted molecules. For example, G protein pathways include a variety of extracellular agents like hormones, neurotransmitters, chemokines, and local mediators that all are systemically secreted molecules [33]. From the relative abundance of such orthologs among those with exceptional functionality we may derive a hypothesis of their role in species divergence.

The most extreme gene (LocusIDs 998 and 12540 for human and mouse, respectively) occurs in an exceptionally small cluster in ≥ 80 of the 100 bootstrap iterations. The expressions in man and mouse correlate negatively (-0.47) in this case and the ortholog is exceptional already as such. The human gene is only expressed in neuronal tissues, whereas the mouse gene is more generally expressed (Fig. 7). Such outliers may be either real functional differences in the species or measurement errors. Whichever the reason, the detection of the outlier was useful.

Groups of orthologous genes with a similar but exceptional functional relationship would be more reliable findings than individual outliers. Unfortunately, co-occurrence of orthologous gene pairs in exceptionally small cross clusters is rare. The two cases with the most frequent co-occurrence in small cross clusters have a frequency of 45 out of 100 bootstrap iterations. It is interesting to note that in both cases (Fig. 8) mouse genes are only weakly or not at all expressed in the 21 tissues common to the organisms. In the first case the mouse and human genes are known to be related to translational regulation. Differences in the expression levels might hint at differentiation in the translational mechanisms. In the second case, the human genes (Protein tyrosine kinase 2 and Glia maturation factor, LocusID-pairs 5747-14083 and 2764-63985) are expressed specifically in neuronal tissues and

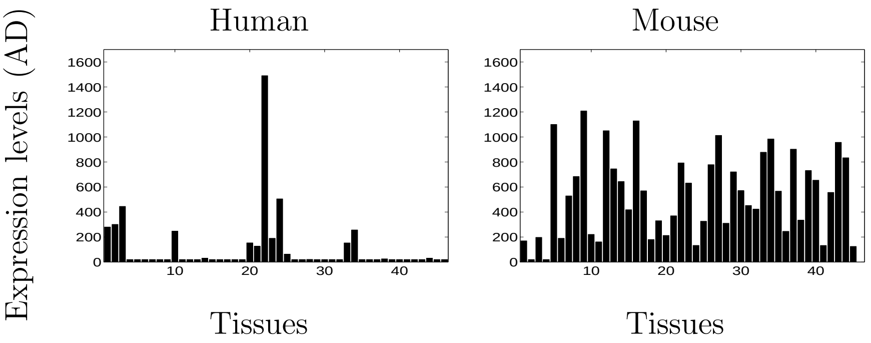


Fig. 7. The most strongly exceptional outlier gene, detected based on its most frequent occurrence in an unexpectedly small cross cluster. LocusIDs 998 and 12540 for human and mouse, respectively.

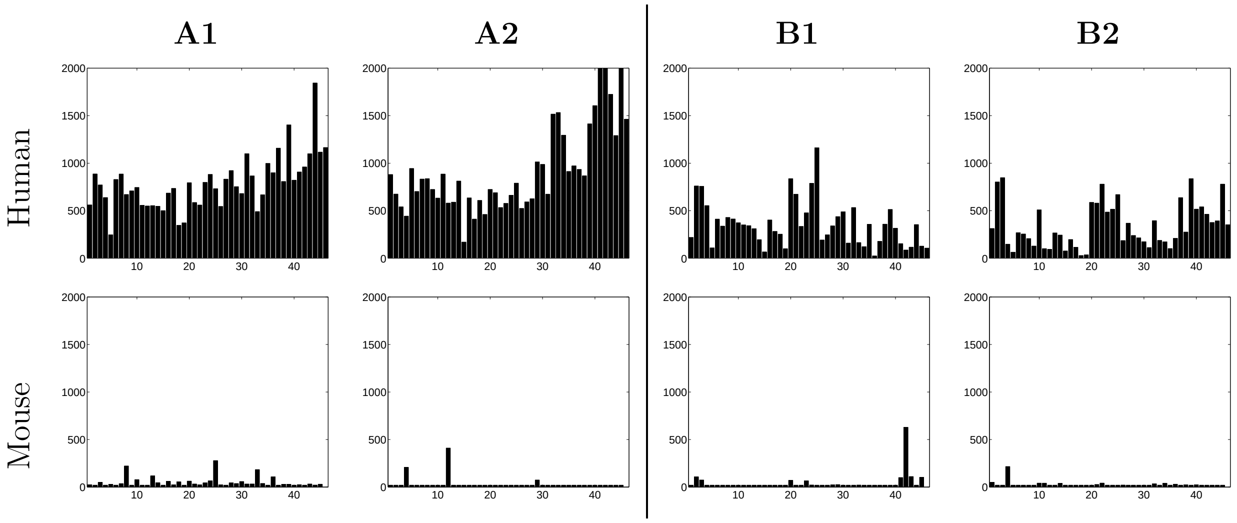


Fig. 8. Two examples (**A** and **B**) of frequently co-occurring and exceptional "clusters" of gene pairs. (They co-occurred frequently in exceptionally small cross clusters). Gene expression profiles belong to human-mouse LocusID pairs **A1** 10438-57316, **A2** 7458-22384 and **B1** 5747-14083, **B2** 2764-63985.

are known to participate in the regulation of growth and differentiation of neurons.

D. Summary

In summary, AC reproduced known findings and performed as expected in comparison with alternative methods. Although this case study is technically interesting and completely new, its biological implications are not yet as convincing as in the second one (Section VI).

From the man-mouse orthologs we found clusters of highly conserved orthologs, possibly unknown functional relationships between genes, and examples of exceptional relationships between orthologs suggesting differentiation in gene function between species. Some of the findings remain unexplained but could be used as starting points for more detailed studies.

VI. Experimental results: Dependencies between gene expression and transcription factor binding

The baker's yeast, *Saccharomyces cerevisiae*, is a popular eukaryotic model organism due to the representativeness of its genetic regulation and because of its easy experimental handling.

Gene expression regulation operates on several levels, of which perhaps the most crucial is transcriptional control. This is handled by a set of regulatory proteins called *transcription factors (TFs)* that bind to DNA in the gene regulatory (promoter) region and can either enhance or suppress the gene's expression. In most cases TFs interact *inter se* to make up macromolecular complexes before binding to the regulatory regions of DNA. Since TFs are manufactured by expressing the relevant genes, they are the key components of gene interaction networks. In this work, we focus on the dependencies between the TFs and gene expression, that is, on the gene regulatory network.

Regulatory interactions have been studied by measuring genome-wide expression with microarrays in knock-out mutation experiments and in time series experiments. In the knock-out experiments, a mutation is targeted to a single gene in the yeast genome to modify (usually knock out) the normal function of that gene. It is then hoped that by measuring the gene expression changes with microarrays after the mutation, the role of the mutated gene in cellular processes is revealed. Genes belonging to the same regulatory pathway as the mutated gene could be unveiled, for example. In time series experiments the goal is often to infer causality in the gene regulatory network based on the sequential changes in expression levels. However, since the interaction network between the genes is complicated, discerning the direct effects of the knock-out, or the change of expression in a time series from noise and the mass of second-order effects can be very difficult, if not impossible. At least a comprehensive, very expensive high resolution time-series experiment with numerous replications would be required. The same holds for knock-out experiments. Thus alternative approaches are worth exploring.

Gene expression is not the only source of information about gene regulation. For instance, microarray-based chromatin immunoprecipitation (ChIP) allows measuring the binding strength of the transcription factor proteins on any gene's promoter region [31]. This reveals which TFs are able to bind the specific gene's promoter and are thus potential regulators. But many TFs bind numerous gene promoter regions and are still not operational regulators. The number of false positives can be very high, and thus inferring the regulatory relationships based on the binding information alone is not in general possible.

Combining data from the several sources is a promising option, and exploratory models are perfectly suited for the first studies. We combine the functional information (gene expression) and the potential regulator information (TF binding). We make the following assumptions. First, it is assumed that the genes are co-expressed in groups that are unknown, cf. [16], [36]. Second, it is sensible to assume that a common set of transcription factors binds to the co-expressed genes. Otherwise groupwise expression would be very unlikely. This is of course an oversimplification, but it has some biological justification. To be more realistic, we do not assume that all the genes are regulated in such a manner; we relax the simplification by assuming that only *subsets* of genes behave this way, only *a subset* of transcription factors need to be the same, and co-expression needs to take place only in *a subset* of knock-out experiments or time points.

Associative clustering, when applied to expression and TF binding data, makes precisely these assumptions, and we now aim to find subsets of genes whose expression is maximally dependent on their transcription factor binding profiles. These sets then act as hypotheses for expression co-regulation.

A. Knock-out expression and TF binding

The yeast expression used in this analysis has been measured from 300 different mutation strains with cDNA microarrays [25] (`http://www.rii.com/publications/2000/cell_hughes.html`). Transcription factor binding data on genes for 113 transcription factors was obtained from [31] (`http://web.wi.mit.edu/young/regulator_network/`). After taking the logarithm of the expression ratios, imputing missing values with genewise averages, standardizing the treatmentwise variances to unity, and including only the genes appearing in both data sets, we had two full data matrices, each with 6185 genes. The number of clusters in the margin spaces was chosen to produce roughly 10 data points in each cross cluster, resulting in 30 clusters in the expression space and 20 clusters in the TF-binding space.

1) Quantitative evaluation: We first used this data to validate the performance of AC in the two tasks it addresses: maximizing the dependency and keeping the clusters homogeneous. These were measured in 10-fold crossvalidation runs with pre-validated σ for AC and pre-validated number of K-means clusters for K-IB. Pre-validation was analogous for both methods: the data was divided into two equally sized parts, and several parameter values were tried from three different random initializations. Of these the parameter value giving the best AC cost was chosen. The final cross-validation runs were also started from three different random initializations.

AC discovered dependencies in the data significantly better than the reference methods (10-fold crossvalidation, paired t-test; d.f.=9; $p < 0.001$). The dependency was measured with (natural) logarithmic Bayes factor (log-BF), the average value of which was 8.84 for AC, -46.37 for IB, and -262.29 for K-means. The value of log-BF is traditionally interpreted to signify strong evidence against the null hypothesis if it is at least 6–10 [28].

The homogeneity, or actually dispersion, of the clusters was measured simply by the sum of the componentwise variances in cross-validation. The comparison was made for both margin clusters as well as for cross clusters. Margin clusters produced by AC were statistically significantly less dispersed than those produced by IB, but for cross clusters the difference was not significant.

2) Biological results: We sought for biologically interesting findings by bootstrapping the AC (100 bootstrap data sets), and by otherwise using the same parameters as in the above cross-validation tests. A similarity matrix was generated for the genes from the bootstrap results (see Sect. IV-B), and summarized by the average-distance variant of hierarchical clustering. Clusters with average co-occurrence higher than 20 out of 100 and with the minimum size of 3 genes were chosen for the final analysis, resulting in 20 clusters.

The clusters were first screened with EASE, which found enriched gene ontology classes in 12 of the 20 clusters (Fisher's exact test, Bonferroni corrected; $p < 0.05$). It is of course likely that also clusters without significant GO enrichments are biologically meaningful, but their interpretation is more cumbersome and is therefore left for future work. In the following we present a sample of four representative AC cluster types.

The first, most notable cluster is a large set of about one hundred genes that all code for ribosomal proteins. These genes are known to be expressed often very homogeneously, and they can also often be found in conventional cluster analyses, cf. [5], [34].

The next two clusters are examples of how AC identifies and highlights modules where a subset of the genes and their main regulator(s) have been previously identified in wet-lab experiments. However, the modules also contain novel components not previously associated to the corresponding biological function.

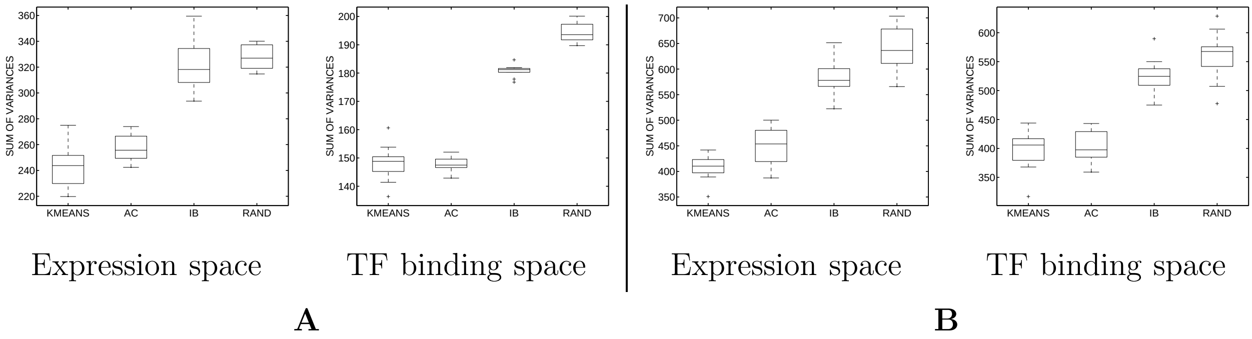
The second cluster is an example of a cluster type rarely found in conventional analyses. It contains only 4 genes, of which 3 are known to code for proteins involved in lipid metabolism and one to code for a growth factor transporter. The most reliable and strongest transcription factor bindings in this cluster are by proteins INO2/YDR123Cp and INO4/YOL108Cp that are known to form a protein complex and then regulate lipid metabolism. The fact that AC detects two interacting TFs shows that the method can be used, to a certain extent, to predict TF interactions as well. Moreover, it also unveils which potential target genes are responsible for the lipid metabolism regulation observed in wet lab experiments. In other words, the reliability of gene function annotations is enhanced through the use of AC

The third cluster of 31 genes contains 20 genes involved in amino acid and derivative metabolism. The best identified regulator for this cluster is GCN4/YEL009Cp, a transcriptional activator of amino acid biosynthetic genes known to respond to amino acid starvation. Here again, it is shown that the AC creates a partially new cluster and identifies a good candidate regulator.

About two thirds (28) of the genes in the fourth cluster, the most interesting so far, are of unknown molecular function. Even the biological process they contribute to may be unknown. The known genes map to such GO categories as “nuclear organization and biogenesis” and the most reliable transcription factor associated to genes in this cluster was YAP5p/YIR018Wp. This transcription factor is known to be activated by the main regulators (SBF and MBF [22]) of the START of the cell cycle, a time just before DNA replication. This clearly refers to cell-cycle regulation and to organization of the nucleus prior to replication.

B. Time series gene expression and TF binding

The expression data for this case study was measured during yeast cell cycle and was originally published in two different papers [10], [42] (`http://genome-www.stanford.edu/cellcycle/`



Expression space TF binding space Expression space TF binding space

Fig. 9. **A** Margin cluster and **B** cross cluster dispersion for all methods in cell-cycle experiments, demonstrating that AC produces clusters that are almost as compact as K-means clusters, whereas the IB-clusters are significantly more dispersed. RAND is a kind of an upper limit for cluster dispersion, obtained by randomly assigning samples to clusters.

`links.html`). The data consisted of 77 timepoints in total. The transcription factor binding data used here is the updated (2003) version of [31] for 106 transcription factors. In this case study the missing values were imputed with the k-nearest neighbor method ($k = 10$) [45] and logarithms were taken from both of the data sets. Including only the genes present in both data sets resulted in a total of 5618 genes. The chosen cluster numbers were 30 in the expression space and 20 in the TF-binding space.

1) Numerical results: The tests were run as described in Section VI-A. The differences in dependency modeling between all the methods were statistically significant also for this data pair (10-fold cross-validation, paired t-test; d.f.=9; $p < 0.001$). Natural logarithmic Bayes factor for AC was 32.27, for IB -13.17, and for K-means -92.30, implying that AC found a very strong dependency between the data sets.

The measure of cluster homogeneity, or actually dispersion, was the same as in the previous cases: the sum of the componentwise variances. For this data pair AC produced significantly (10-fold cross-validation, paired t-test; d.f.=9; $p < 0.001$) less dispersed cross clusters and margin clusters than IB. Figure 9 visualizes the margin cluster and cross cluster dispersion for all methods.

2) Biological results: In a similar manner as in the previous case, we sought for biological findings from the bootstrapped AC clusters. The clusters with average distance smaller than 60 (times in the same dependent cross cluster out of 100) and with more than 2 genes were chosen. This resulted in a total of 16 clusters.

Gene ontology classes were enriched statistically significantly in 13 of the 16 clusters (EASE; Fisher's exact test, Bonferroni corrected; $p < 0.05$). In the similar spirit as in the knock-out mutation case, we give a representative sample of four clusters.

Two clusters are essentially the same as in the in knock-out case study, the ribosomal proteins being the first of them.

The second cluster is the same as the most interesting (fourth) cluster in the knock-out case. This provides more evidence that the cluster represents a biologically robust motif, having a homogeneous profile in both TF-binding and expression.

The third cluster (Fig. 10) contains a significantly high number of genes involved in cell cycle regulation, and more specifically at the stage of entry into the mitotic cell cycle (9 genes out of 33). The main regulator identified in this module is SIP4p/YJL089Wp which is possibly involved in SNF1p/YDR477Wp-regulated transcriptional activation. This latter signaling factor is required for transcription in response to glucose limitation. Interestingly, SIP4p/YJL089Wp has a DNA-binding domain similar to the GAL4p/YPL248Cp transcription factor, involved in galactose response, another route in energy metabolism. Taken together, this cluster contains some clear references to cell cycle regulation on one hand and energy metabolism on the other, and proposes a set of genes that can bridge and connect these two biological processes. Thereby AC offers the hypothesis for a relation between biological functions, in addition to some clues on what genes could be involved.

The fourth cluster contains 9 genes of unknown molecular function or associated biological process. The associated transcription factor ACE2p/YLR131Cp is known to activate expression of early G1-specific genes, localizes to daughter cell nuclei after cytokinesis and there delays G1 progression in the daughters. Based on this data, the 9 genes can be predicted to act during the G1 phase of the cell-cycle, thus specifying what kind of targeted experiments are needed to establish their function.

VII. Conclusion and future work

We have introduced a new approach for a relatively little-studied machine learning or data mining problem: From data sets of co-occurring samples, find what is in common. We have formulated the problem probabilistically, extending earlier mutual information-based approaches. The new solution is better-justified for finite (relatively small) data sets.

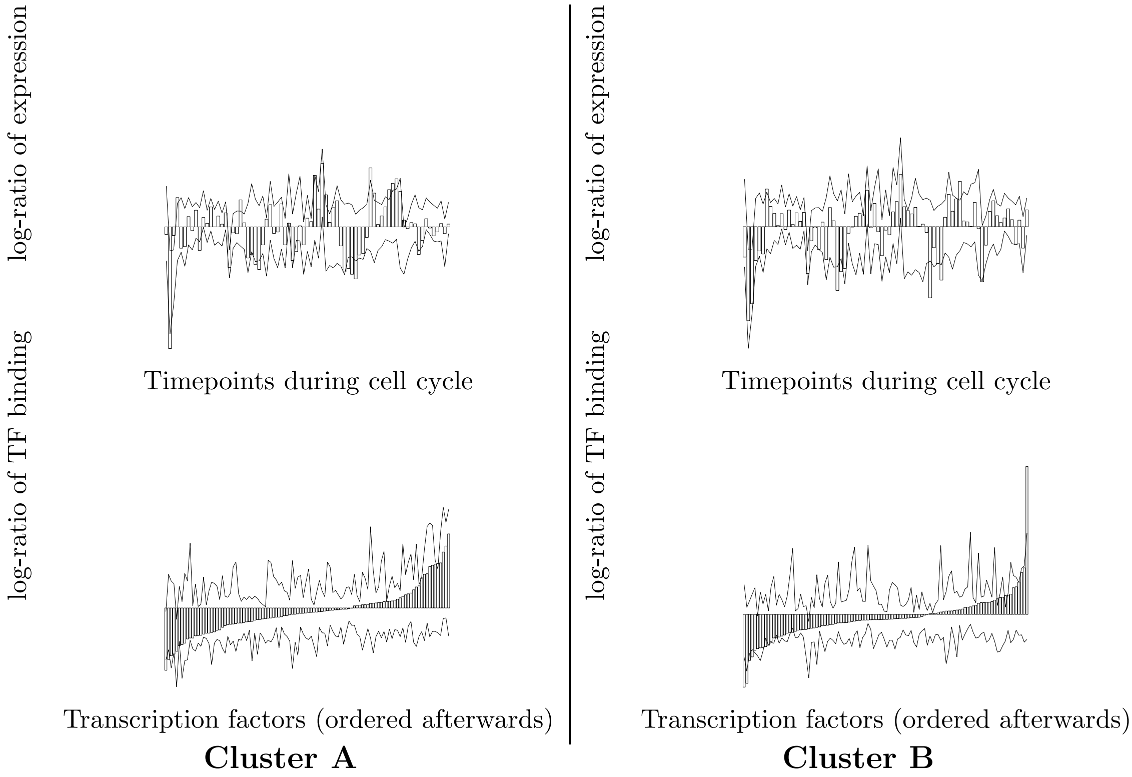


Fig. 10. Two examples of bootstrapped cross clusters, associated to cell cycle, that reveal both known and novel dependencies between gene expression and TF binding. The upper figures show the average expression profiles (bars) of the clusters and confidence intervals (curves). The periodicity of the cell cycle in the expression is clearly visible. The lower figures show the average TF-binding profile of the clusters with confidence intervals. The average TF-bindings rising above the confidence interval are considered reliable. Note that the confidence intervals are very conservative; they have been estimated based on random clusters. In the **Cluster A** there was only one reliable TF binding, SIP4. It could be verified from the literature (see text for details). SIP4 binds also the genes in **Cluster B**, but additionally there is one extremely strongly binding TF, SFL1 (the rightmost bar). Its putative regulatory interaction with the gene cluster during cell cycle is a new finding.

The introduced method, coined associative clustering (AC), summarizes dependencies between data sets as clusters of similar samples having similar dependencies. Such a method is particularly needed for mining functional genomics data where measurements are available about different aspects of the same set of functioning genes. Then a key challenge is to find commonalities between the measurements. The answer should reveal characteristics of the genes, not only characteristics of the measurement setups.

The work is pure machine learning in the sense that the model is a general-purpose semiparametric model which learns to fit a new data set instead of being manually tailored.

As a result, it is probably not as accurate as more specific models, but it can be expected to be faster and easier to apply to new problems. Its main intended application area is in exploratory data analysis, "looking at the dependencies in the data" in the first stages of a research project.

The method was validated and applied in two functional genomics studies. The first found regularities and differences between functioning of orthologous genes in different organisms, suggesting evolutionary conservation and divergence. The second explored regulatory interactions between gene expression and transcription factor binding. Both trivial and unexpected findings were made: known regularities, outliers, and hints about unexpected regularities.

While the proposed method was shown to be viable already as such, it can be further improved. We did not address the problem of choosing an optimal number of clusters. If clustering is interpreted as a partitioning or quantization of data to compress its presentation, then the exact number of clusters is not a crucial parameter, but nevertheless the results could be improved by optimizing it. Since the task is formulated in Bayesian terms, Bayesian complexity control methods are applicable in principle. The setting is not standard, however, because of the non-standard (new) use of the Bayes factors, and because of discontinuities in the objective function.

Another direction of improvement is regularization of the solution. Dependency-searching methods may potentially overfit the data, which is well-known from canonical correlation analysis and can be avoided by regularization. We have developed two regularization methods for AC with one fixed margin. "Entropy regularization" was used here because it is easier in practice and has not been shown to be worse than the alternative [27]. In the present case bootstrap also helped. Another related question is which kinds of priors to use for the distributional parameters. The simple constant Dirichlet priors used in this work may be too informative. Hierarchical modeling should be more appropriate but it is computationally more complex.

A third area worth investigating is the parameterization of the clusters. It should be investigated whether the hard Voronoi regions, used up to now because they are easily interpretable and make the theory manageable, could be replaced by smooth and more regular-sized clusters. Alternatively, the degrees of freedom of the clusterings could be directly reduced to regularize the solution.

Finally, a comprehensive comparison of the relative merits of dependency maximization and more traditional Bayes networks and graphical models of the whole joint distribution should be carried out. It is clear that the two approaches focus on different properties of data, and that our semiparametric models need less prior knowledge than specialized models of gene regulation, for instance, and are hence more general-purpose. We expect that exploratory models of the type introduced here are viable as complementary methods for gathering the necessary prior knowledge for the more specific models.

Appendix

Tissues in mouse-human data

The first 21 tissues are considered to be common for both species. (Listed in the following order: tissue number: human tissue: mouse tissue. Tissues are separated with commas.)

Common tissues: 1: cerebellum: cerebellum, 2: cortex: cortex, 3: amygdala: amygdala, 4: testis: testis, 5: placenta: placenta, 6: thyroid: thyroid, 7: prostate: prostate, 8: ovary: ovary, 9: uterus: uterus, 10: 0DRG: 0DRG, 11: salivary gland: salivary gland, 12: trachea: trachea, 13: lung: lung, 14: thymus: thymus, 15: spleen: spleen, 16: adrenal gland: adrenal gland, 17: kidney: kidney, 18: liver: liver, 19: heart: heart, 20: caudate nucleus: striatum, 21: spinal cord: spinal cord lower,

Non-common tissues: 22: fetal brain: digits, 23: whole brain: gall bladder, 24: thalamus: hippocampus, 25: corpus callosum: large intestine, 26: pancreas: adipose tissue, 27: pituitary gland: lymph node, 28: prostate Cancer: eye, 29: OVR278E: skeletal muscle, 30: OVR278S: snout epidermis, 31: fetal liver: tongue, 32: HUVEC: trigeminal, 33: THY+: bladder, 34: THY-: small intestine, 35: myelogenous k-562: stomach, 36: lymphoblastic molt-4: hypothalamus, 37: burkitts Daudi: epidermis, 38: bukitts Raji: spinal cord upper, 39: hep3b: bone, 40: A2058: brown fat, 41: DOHH2: olfactory bulb, 42: GA10: mammary gland, 43: HL60: umbilical cord, 44: K422: bone marrow, 45: ramos: frontal cortex, 46: WSU: -.

Acknowledgment

The authors would like to thank Jaakko Peltonen for the code for IB. This work has been supported by the Academy of Finland, decisions #79017 and #207467.

References

[1] M. Ashburner et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[2] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms.* Wiley, New York, 1993.

[3] S. Becker. Mutual information maximization: models of cortical self-organization. *Network: Computation in Neural Systems*, 7:7–31, 1996.

[4] S. Becker and G. E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355:161–163, 1992.

[5] M. Beer and S. Tavazoie. Predicting gene expression from sequence. *Cell*, 117:185–198, 2004.

[6] S. Bergmann, J. Ihmels, and N. Barkai. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biology*, 2:85–93, 2004.

[7] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, New York, 1995.

[8] H. Bono and Y. Okazaki. Functional transcriptomes: comparative analysis of biological pathways and processes in eukaryotes to infer genetic networks among transcripts. *Current Opinion in Structural Biology*, 12:355–361, 2002.

[9] S. B. Carroll. Genetics and the making of Homo sapiens. *Nature*, 422:849–857, 2003.

[10] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodickaa, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.

[11] A. G. Clark et al. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, 302:1960–1963, 2003.

[12] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman&Hall, New York, 1993.

[13] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences, USA*, 95:14863–14868, 1998.

[14] W. Enard et al. Intra- and inter-specific variation of primate gene expression patterns. *Science*, 296:340–343, 2002.

[15] R. M. Ewing and J.-M. Claverie. EST databases as multi-conditional gene expression datasets. In *Proceedings of Pacific Symposium on Biocomputing*, volume 5, pages 427–439, 2000.

[16] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:559–584, 2000.

[17] N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby. Multivariate information bottleneck. In *Proc. Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 152–161. Morgan Kaufmann Publishers, San Francisco, CA, 2001.

[18] V. Ganti, J. Gehrke, R. Ramakrishnan, and W.-Y. Loh. A framework for measuring changes in data characteristics. In *Proceedings of ACM PODS 1999, 18th Symposium on Principles of Database Systems*, pages 126–137. 1999.

[19] H. Ge1, Z. Liu, G. M. Church, and M. Vidal. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nature Genetics*, 29:482–486, 2001.

[20] I. J. Good. On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Annals of Statistics*, 4:1159–1189, 1976.

[21] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[22] C. E. Horak, N. M. Luscombe, J. Qian, P. Bertone, S. Piccirrillo, M. Gerstein, and M. Snyder. Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae. *Genes and Development*, 16:3017–3033, 2002.

[23] D. Hosack, G. Dennis Jr., B. Sherman, H. Lane, and R. Lempicki. Identifying biological themes within lists of genes with EASE. *Genome Biology*, 4(R70), 2003.

[24] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

[25] T. R. Hughes et al. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.

[26] J. L. Jiménez, M. P. Mitchell, and J. G. Sgouros. Microarray analysis of orthologous genes: conservation of the translational machinery across species at the sequence and expression level. *Genome Biology*, 4:R4, 2002.

[27] S. Kaski, J. Sinkkonen, and A. Klami. Discriminative clustering. *Neurocomputing*, 2005. Accepted for publication.

[28] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[29] M. K. Kerr and G. A. Churchill. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences*, 98:8961–8965, 2001.

[30] P. Khaitovich, G. Weiss, M. Lachmann, I. Hellmann, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge, and S. Pääbo. A neutral model of transcriptome evolution. *PLoS Biology*, 2:0682–0689, 2004.

[31] T. I. Lee et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298:799–804, 2002.

[32] G. J. McLachlan, K.-A. Do, and C. Ambroise. *Analyzing microarray gene expression data*. Wiley, New York, 2004.

[33] S. R. Neves, P. T. Ram, and R. Iyengar. G Protein Pathways. *Science*, 296:1636–1639, 2002.

[34] J. Nikkilä, P. Törönen, S. Kaski, J. Venna, E. Castrén, and G. Wong. Analysis and visualization of gene expression data using self-organizing maps. *Neural Networks*, 15:953–966, 2002. Special issue on New Developments on Self-Organizing Maps.

[35] J. Peltonen, J. Sinkkonen, and S. Kaski. Sequential information bottleneck for finite data. In R. Greiner and D. Schuurmans, editors, *Proceedings of the Twenty-First International Conference on Machine Learning (ICML 2004)*, pages 647–654. Omnipress, Madison, WI, 2004.

[36] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*, 34:166–176, 2003.

[37] J. Sinkkonen and S. Kaski. Clustering based on conditional distributions in an auxiliary space. *Neural Computation*, 14:217–239, 2002.

[38] J. Sinkkonen, S. Kaski, J. Nikkilä, and L. Lahti. Associative Clustering (AC): Technical Details. Technical Report A84, Publications in Computer and Information Science, Laboratory of Computer and Information Science, Helsinki University of Technology, Espoo, Finland, 2005.

[39] J. Sinkkonen, J. Nikkilä, L. Lahti, and S. Kaski. Associative clustering. In Boulicaut, Esposito, Giannotti, and Pedreschi, editors, *Machine Learning: ECML2004 (Proceedings of the ECML'04, 15th European Conference on Machine Learning)*, pages 396–406. Springer, Berlin, 2004.

[40] N. Slonim. *The information bottleneck: theory and applications*. PhD thesis, Hebrew University, Jerusalem, 2002.

[41] N. Slonim, N. Friedman, and N. Tishby. Unsupervised document classification using sequential information maximization. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 129–136. ACM Press, 2002.

[42] P. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.

[43] A. I. Su et al. Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences, USA*, 99:4465–4470, 2002.

[44] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377. Urbana, Illinois, 1999.

[45] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17:520–5, 2001.

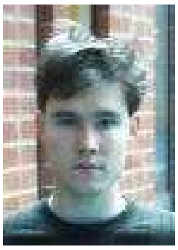
[46] D. L. Wheeler et al. Database resources of the national center for biotechnology. *Nucleic Acids Research*, 31:28–33, 2003.

Samuel Kaski (M'96–SM'02) received the D.Sc. (Ph.D.) degree in computer science from Helsinki University of Technology, Espoo, Finland, in 1997.

He is currently Professor of Computer Science at University of Helsinki, Finland. His main research areas are statistical machine learning and data mining, bioinformatics, and information retrieval.

Janne Nikkilä received the M.Sc.(Tech.) degree from Helsinki University of Technology, Espoo, Finland, in 1999. He is currently finalizing his Ph.D. thesis about exploratory clustering analysis methods applied to genomic data sets at Laboratory of Computer and Information Science (Neural Networks Research Centre), Helsinki University of Technology, and is also partly affiliated with University of Helsinki.

Janne Sinkkonen received an M.A. degree in psychology from University of Helsinki in 1996, and the Ph.D. degree on machine learning from Helsinki University of Technology (HUT) in 2004. He has worked as a researcher in a Helsinki University brain research group during 1990's, and as a researcher at the HUT Neural Networks Research Centre during 2000's. He is currently at Xtract Ltd.



Leo Lahti received the M.Sc.(Tech.) degree from Helsinki University of Technology, Espoo, Finland, in 2003. He is currently postgraduate researcher at the Laboratory of Computer and Information Science (Neural Networks Research Centre), Helsinki University of Technology, focusing on research and development of data analysis methods for bioinformatics.

Juha E.A. Knuuttila is a graduate student at the University of Jyväskylä, Finland where he is pursuing for the masters degree in molecular biology. For the doctoral thesis he has started studying the plasticity of the adult brain in different research paradigms by measuring the alterations of gene expression and the activity status of some plasticity related proteins at the Neuroscience Center, University of Helsinki, Finland.

Christophe Roos graduated in genetics and mathematics (1982) at the University of Helsinki, Finland, whereafter he performed a Ph.D thesis in molecular biology (1986) at the University of Strasbourg, France. After having directed a Drosophila developmental biology research group at the University of Helsinki, he joined (2000) Medicel Ltd., a company developing a systems biology software platform. His principal scientific interests proceed from the use of bioinformatics in developmental biology.