Supplementary Material for: Information based clustering

Noam Slonim, Gurinder Singh Atwal, Gašper Tkačik, and William Bialek

Joseph Henry Laboratories of Physics, and
Lewis-Sigler Institute for Integrative Genomics,
Princeton University, Princeton, New Jersey 08544
{nslonim, gatwal, gtkacik, wbialek}@princeton.edu

Contents

1 Introduction

This technical report provides the supplementary material for the paper entitled "Information based clustering". It is organized as follows. In Section 2 we present in detail the iterative clustering algorithm used in our experiments. In Section 3 we describe the validation scheme used to determine the statistical significance of our results. In Section 4, Section 5, and Section 6 we provide all the experimental results in all three applications. In particular, we highlight some of the results that seem to deserve special attention. Parts of the text and figures of the main paper are repeated here for convenience.

This report does *not* deal with the technical details of estimating mutual (and multi) information from empirical data. The reader is referred to (*1*) for a complete description of the estimation procedure used in our experiments.[1] Nonetheless, some of the results regarding information estimation are repeated here for completeness.

As mentioned in the corresponding sections, all the experimental results and relevant code, including a freely available web application, are available at *http://www.genomics.princeton.edu/biophysics-theory*.

[1] See also *http://www.genomics.princeton.edu/biophysics-theory* for the relevant software and a freely available web implementation.

2 The Iclust algorithm

As implied by the main paper, a typical clustering task involves many choices at different levels of the analysis. The choice of the specific algorithm, or optimization routine being used, is only another choice among many, and not necessarily the most important one. In our work we concentrated on offering a principled approach to all the decisions that *precede* the choice of the algorithm. Thus, the specific algorithm we used is only briefly mentioned in the main text and here we provide a more detailed description. We emphasize that we used this algorithm mainly because it emerges directly out of the theoretical analysis. Other procedures that aim to optimize the same target functional are certainly plausible and we expect future research to elucidate the potential (dis)advantages of such alternatives.

For brevity, let us further concentrate on the conventional case $r = 2$ where only pairwise interactions are considered. The derivation for the general case is similar. For $r = 2$, any stationary point of our target functional, $\mathcal{F}$, must obey:

$$P(C|\mathrm{i}) = \frac{P(C)}{Z(\mathrm{i};T)} \exp\left\{\frac{1}{T}[2s(C;\mathrm{i}) - s(C)]\right\}, \tag{1}$$

where $Z(\mathrm{i};T)$ is a normalization function, $s(C;\mathrm{i})$ is the expected similarity between i and a member of cluster C,

$$s(C;\mathrm{i}) = \sum_{\mathrm{i}_1=1}^{N} P(\mathrm{i}_1|C)s(\mathrm{i}_1,\mathrm{i}), \tag{2}$$

and $s(C)$ is the average similarity among pairs chosen independently out of the cluster C,

$$s(C) = \sum_{\mathrm{i}_1=1}^{N}\sum_{\mathrm{i}_2=1}^{N} P(\mathrm{i}_1|C)P(\mathrm{i}_2|C)s(\mathrm{i}_1,\mathrm{i}_2). \tag{3}$$

Eq. (1) defines an implicit set of equations since the right hand side depends on $P(i|C)$ and $P(C)$, which in particular depend on $P(C|i)$ through Bayes' rule ($P(i|C) = \frac{P(C|i)P(i)}{P(C)}$) and through $P(C) = \sum_{i=1}^{N} P(C|i)P(i)$. This is a common situation in variational methods, also present, for example, in conventional rate-distortion clustering (*4*), in maximum likelihood estimation with hidden variables (*5*) and in the Information Bottleneck framework (*6*). The standard strategy is to turn the self-consistency condition into an iterative algorithm by simply applying the self-consistent equations iteratively. Specifically, let us denote the intermediate solution of the algorithm at the m'th

iteration by $P^{(m)}(C|i)$. Then, at the $m+1$'th iteration, the algorithm applies the following update rule:

$$P^{(m+1)}(C|\mathrm{i}) \leftarrow P^{(m)}(C) \exp\left\{\frac{1}{T}[2s^{(m)}(C;\mathrm{i}) - s^{(m)}(C)]\right\},\ \forall\, C = 1 : N_c\ , \tag{4}$$

followed by a normalization step. Notice, that $\{P^{(m)}(C),\ s^{(m)}(C;\mathrm{i}),\ s^{(m)}(C)\}$ are all calculated using the previous $P^{(m)}(C|i)$. A Pseudo-code of this algorithm is given in Figure 1. It is easy to verify that with a straightforward implementation, the complexity of this algorithm is $O(N^3 \cdot N_c)$ for a single pass over the entire data. We will refer to this algorithm as the Iclust algorithm.

To gain some intuition let us consider a typical situation where i is relatively similar to elements in C, but very different from elements in C'. Thus, the exponent will be positive for i and C, but might be negative for i and C'. Consequently, while applying the update step the assignment of i to C will be boosted while its assignment to C' will cut down. This is clearly a desirable outcome, which in particular should increase $\mathcal{F}$. Thus, since $\mathcal{F}$ is upper bounded (as a sum of information terms), after a finite number of such updates the algorithm is expected to converge to a fixed-point which corresponds to a (possibly local) stationary point of $\mathcal{F}$.

This example also illustrates one of the differences between our algorithm and previous approaches. While in the Blahut-Arimoto algorithm in rate-distortion (*4*), in the iterative Information Bottleneck algorithm (*6*), and typically also in EM for maximum likelihood (*5*), the sign of the exponent is constant (for a given i), this is not true in our case. In principle, such a non-constant exponent sign should imply faster convergence to a local stationary point, but might also imply higher sensitivity to the random initialization of $P(C|i)$. Thus, as done in other works, we typically perform several runs with different random initializations of $P(C|i)$ from which we choose the best solution, i.e., the one that maximizes $\mathcal{F}$.

The Iclust algorithm presented here uses a "sequential", or "incremental" iterative procedure in which the updates for some i incorporate the implications of the updates for "preceding" elements, $i' \neq i$. As a simple example, consider the case where we have three elements ($N = 3$) and two clusters ($N_c = 2$). We start from some random conditional distribution matrix, $P^{(0)}(C|i)$, which in particular defines $s^{(0)}(C),\ \forall\, C = 1 : 2$. At the first iteration we find a new distribution for the first element ($i = 1$) over the two clusters. Thus, we now have a new conditional distribution matrix, $P^{(1)}(C|i)$ which differs from the previous $P^{(0)}(C|i)$ only by its first row. This distribution is used to define

Input:

Pairwise similarity matrix, $s(i_1, i_2),\ \forall\ i_1 = 1, ..., N,\ i_2 = 1, ..., N$.
Trade-off parameter, T .
Requested number of clusters, N_c .
Convergence parameter, ϵ .

Output:

A (typically "soft") partition of the N elements into N_c clusters.

Initialization:

$m = 0$.
$P^{(m)}(C|i) \leftarrow$ A random (normalized) distribution $\forall\ i = 1, ..., N$.

While True

For every $i = 1, ..., N$:

- $P^{(m+1)}(C|\mathrm{i}) \leftarrow P^{(m)}(C) \exp\left\{\frac{1}{T}[2s^{(m)}(C;\mathrm{i}) - s^{(m)}(C)]\right\},\ \forall\ C = 1, ..., N_c$.
- $P^{(m+1)}(C|\mathrm{i}) \leftarrow \frac{P^{(m+1)}(C|\mathrm{i})}{\sum_{C'=1}^{N_c} P^{(m+1)}(C'|\mathrm{i})},\ \forall\ C = 1, ..., N_c$.
- $m \leftarrow m + 1$.

If $\forall\ i = 1, ..., N,\ \forall\ C = 1, ..., N_c$ we have $|P^{(m+1)}(C|i) - P^{(m)}(C|i)| \leq \epsilon$,
Break.

Figure 1: Pseudo-code of the Iclust algorithm. Extending the algorithm for the general case (of more than pairwise relations, $r > 2$) is straightforward. In principle we repeat this procedure for different initializations and choose the solution which maximizes $\mathcal{F} = \langle s \rangle - TI(C;\mathrm{i})$.

$s^{(1)}(C),\ \forall\ C = 1:2$. Now, in the next iteration, we find a new distribution for the second element ($i = 2$) over the two clusters. This yields another new conditional distribution matrix, $P^{(2)}(C|i)$ which differs from the previous $P^{(1)}(C|i)$ only by its middle row. And so on. This process is somewhat in the spirit of the "incremental EM" (*7*). An alternative optimization routine, which seems somewhat less natural in our case, would be "parallel" optimization, used, e.g., in standard EM. In this case, if we continue our example, at the first iteration we will update *all* the rows in the conditional distribution matrix, $P^{(0)}(C|i)$ using $s^{(0)}(C)$, to find the new $P^{(1)}(C|i)$.

In some extreme cases the above algorithm might produce a non-monotonic behavior in $\mathcal{F}$. That is, some of the updates might reduce $\mathcal{F}$, suggesting that obtaining a general proof of convergence is a challenging goal. Nonetheless, even in these extreme cases, and more generally in all our experiments (which included more than 1000 runs over real world problems with different T and N_c values), the algorithm always converged to a stationary point. Moreover, for the regime $T \geq max_{i_1,i_2} s(i_1, i_2)$ it is possible to prove this convergence analytically (the details will be presented elsewhere).

3 Evaluating clusters' coherence

In this section we describe in detail the validation scheme we used to determine the statistical significance of our results. In general, the core information for such a validation process is a set of *annotations* (or labels) provided for every data item we clustered. Importantly, these annotations are not used during the (unsupervised) clustering process but rather are exposed only for the post-clustering validation. Every data item might be assigned with more than one annotation via different sources of information. These annotations reflect to some extent the "real" structure of the data that one wishes to reveal through the clustering process.

To be more concrete, let us assume that we clustered N elements where each one of these elements is assigned with some set of annotations. Formally, this could be represented through an *annotation matrix*, denoted as $\mathbf{A}$, with N rows and R columns, where R is the number of distinct annotations in our data. Thus, $\mathbf{A}(i,j) = 1$ if and only if the i-th element is assigned with the j-th annotation, and zero otherwise. A simple example is given in Table 1.

While we examine a single cluster, consisting of $n < N$ elements, the first question we might ask is whether some annotations occur in this cluster with a "suspiciously" high frequency. Let us consider a specific annotation a_j that is assigned to $K \leq N$ elements in the entire population and to $x \leq n$ elements in the cluster. The probability of this event, under the null hypothesis that the cluster was chosen at random, is given by the *hyper–geometric* distribution:

$$P_{hyper}(x \,|\, n, K, N) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \,. \tag{5}$$

Table 1: A simple example for an annotation matrix. Here, the total number of elements is $N = 5$ and the total number of distinct annotations is $R = 4$. The first element is assigned with the second and third annotations, and so on.

Element index	a_1	a_2	a_3	a_4
$element_1$	0	1	1	0
$element_2$	1	0	1	1
$element_3$	1	0	0	0
$element_4$	0	1	1	1
$element_5$	1	0	1	1

Table 2: Examples of P-values. When the annotation is over-abundant in the cluster (with respect to its frequency in the entire population) the P-value is reduced accordingly.

N (Population size)	K (Annot. freq.)	n (Cluster size)	x (Annot. freq. in cluster)	$Pval$
1000	100	50	5	0.57
1000	100	50	20	10^{-8}
1000	20	100	2	0.61
1000	20	100	20	10^{-21}

The corresponding P-value is defined as the tail of this distribution:

$$Pval(x \mid n, K, N) = \sum_{x'=x}^{\min(\mathrm{K,n})} P_{hyper}(x' \mid n, K, N) \; . \tag{6}$$

In words, it is the probability of observing x or more elements in the cluster with annotation a_j where the cluster is chosen independently of this annotation. Alternatively, it is the probability of *wrongly* rejecting the hypothesis that the cluster has nothing to do with the annotation a_j. The smaller the P-value the more unlikely this null hypothesis becomes. To gain some intuition, several examples are presented in Table 2.

Having defined the statistical significance of a single event we need to bear in mind that in a single cluster one typically observes several (perhaps many) different annotations. Naturally, the more hypotheses one tests the less surprising it is to find one with a small P-value, even in a randomly chosen cluster. The simplest and most conservative approach to correct for this multiple hypotheses testing effect is to apply the Bonferroni correction (see, e.g., (*8*)). Specifically, if the statistical significance level is q (e.g., $q = 0.05$), an event is considered significant if and only if its P-value satisfies:

$$Pval < \frac{q}{H} \; , \tag{7}$$

where H is the number of hypotheses being tested. We will say that a cluster is *"enriched"* with the annotation if the corresponding P-value satisfies Eq. (7).

Finally, while the above procedure determines the significance of every annotation that occurs in the cluster, it is also useful to have a single "score" that roughly summarizes how homogeneous the cluster is with respect to all the given annotations. Different alternatives have been proposed to this end and here we use the "coherence" score,

recently suggested by Segal *et. al* (*9*). Specifically, the *coherence* of a cluster is defined as the percentage of elements in this cluster covered by some annotation that was found to be enriched in this cluster. In particular, this means that the coherence of a cluster is above zero if and only if it is enriched with at least one annotation. That is, there is at least a single “hint” regarding the reason of forming this cluster.

4 First application: The yeast ESR data

4.1 Description of the data

We considered experiments on the response of gene expression levels in yeast to various forms of environmental stress (*2*). Previous analysis of expression patterns from all ~ 6000 genes identified a group of 283 stress–induced and 585 stress–repressed genes that had apparently "nearly identical but opposite" expression profiles (*2*). This collection of 868 genes was thus termed the yeast environmental stress response (ESR) module. As seen in Figure 2, differences in expression profiles within the ESR module are indeed relatively subtle. More recent manual analysis with attention to background biological data suggests that some of these differences are biologically significant (*3*). Thus, it seems a good challenge for our approach to ask if we can discover automatically any meaningful substructure in these data.

Each of the 868 ESR genes was represented by its log-ratio expression profile in the 173 microarray "stress" experiments (*2*).[2] These data are available at *http://genome-www.stanford.edu/yeast_stress/data.shtml*. The list of the 868 ESR genes was taken from Figure 3 at the same website.

4.2 Mutual information estimation results

From these data we estimated all the $\sim 376,000$ mutual information relations, as described in (*1*), ending up with a matrix of 868×868 information relations which defined the input to our clustering procedure. For convenience, we provide here some statistics of the estimated mutual information values. For a complete description, including different verification schemes that support the reliability of our estimates, the reader is referred to (*1*).

The average estimated mutual information was $0.48\ bits$ with a variance of $0.0425\ bits$. This relatively high average value corresponds to the strong positive/negative linear correlations known to be present in these data. Almost 7000 pairs had a mutual information greater than $1\ bit$ where the maximal estimated mutual information was $1.58\ bits$. All the pairwise mutual information relations are presented in Figure 3, where the genes are sorted according to the clustering partition into $N_c = 20$ clusters that we analyze in detail (see below). The self-information relations were set to $I(i;i) = \log_2(5)$ which corresponds to the maximal possible information under a quantization into 5 bins (*1*).

[2] Notice, that the log transformation has no effect on our analysis since the mutual information is invariant to such transformations (*1*).

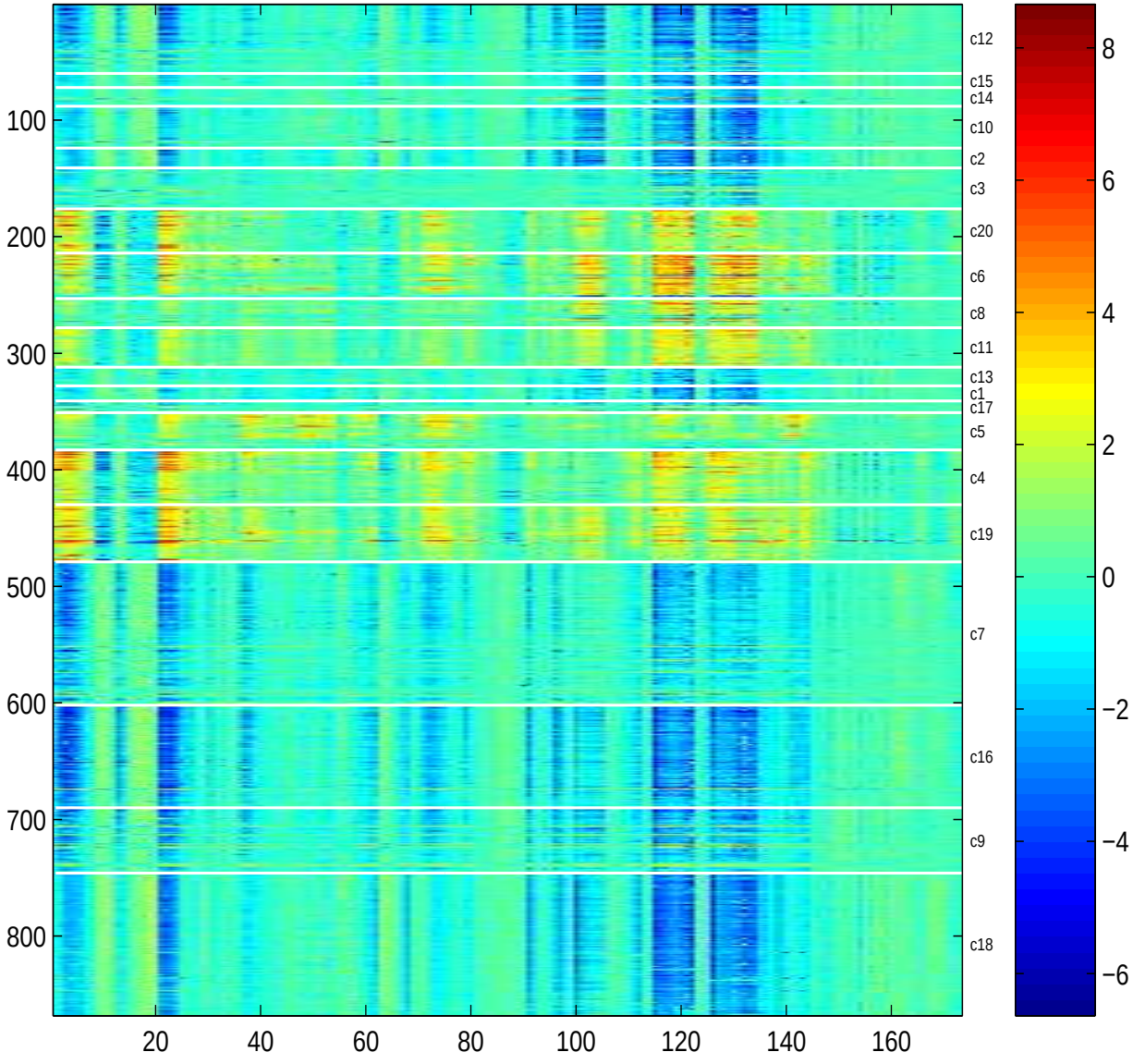

Figure 2: Expression profiles of the 868 genes in the ESR data across the 173 microarray "stress" experiments. Data taken from Gasch *et. al* (*2*). Missing values are set to zero. The genes are sorted according to the clustering partition into 20 clusters that we analyze in detail later on. Inside each cluster, genes are sorted according to the average mutual information relation with other cluster members.

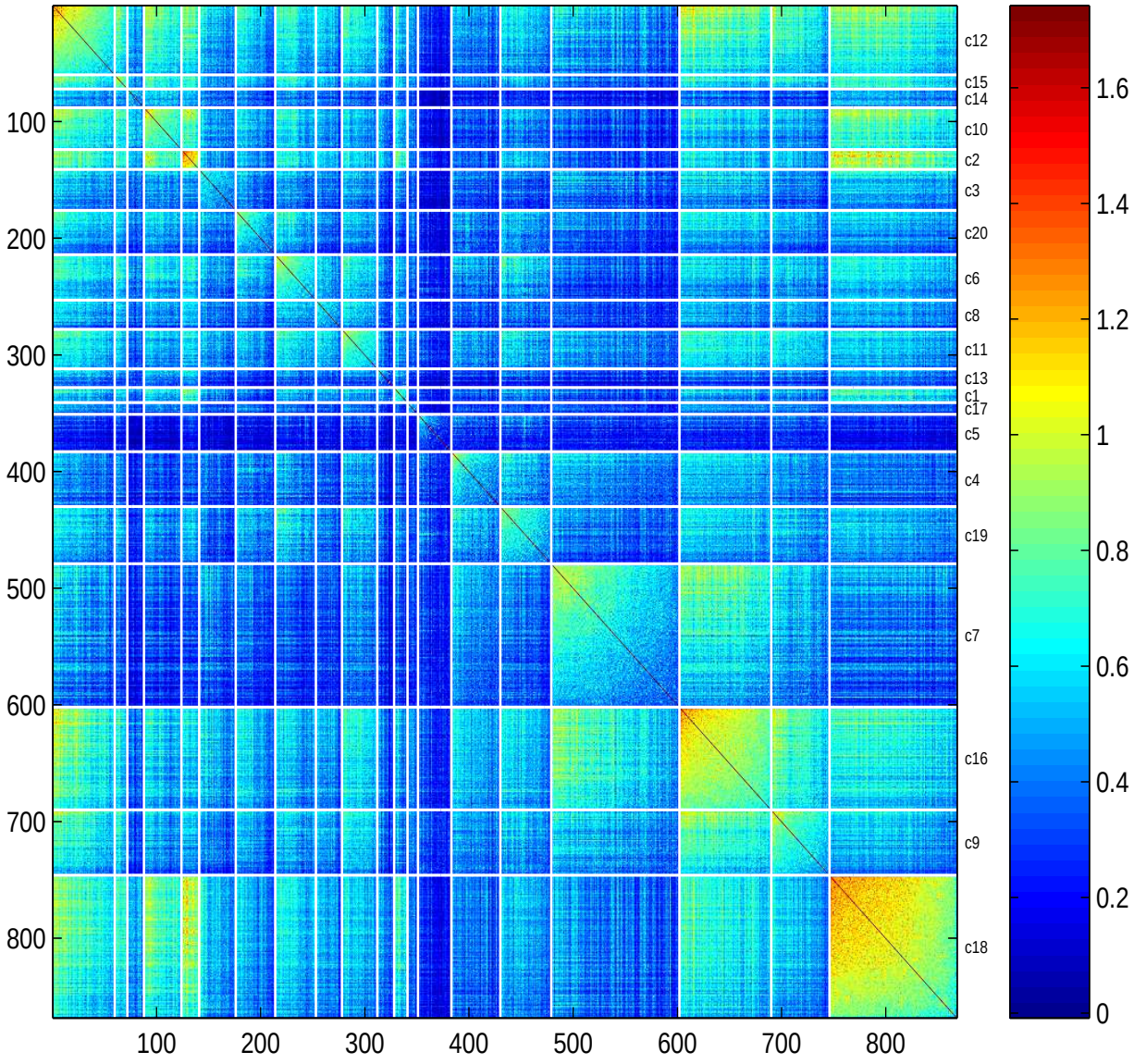

Figure 3: Pairwise mutual information relations for the 868 genes in the ESR data. The genes are sorted according to the clustering partition into 20 clusters that we analyze in detail later on. Inside each cluster, genes are sorted according to the average mutual information relation with other cluster members.

4.3 Implementation details and quality-complexity trade-off curves

Given the pairwise mutual information matrix we applied the Iclust algorithm described in Section 2. Recall that our target functional, $\mathcal{F}$, is given by:

$$\mathcal{F} = \langle s \rangle - TI(C;\mathrm{i}), \tag{8}$$

where T is a (temperature) trade-off parameter, $\langle s \rangle = \sum_{C=1}^{N_c} P(C)s(C)$ measures the quality of the clusters: the average mutual information among intra–cluster pairs, and $I(C;\mathrm{i})$ measures the complexity of the clusters: the cost of coding cluster identity.

For a fixed number of clusters, N_c, the term $\langle s \rangle$ gradually saturates as the temperature T is lowered, while $I(C;\mathrm{i})$ increases accordingly. We explored this trade-off for different numbers of clusters: $N_c = 5, 10, 15, 20$. For each of these values we tried several values of T. Specifically, we found that $\frac{1}{T} = \{5, 10, 15, 20, 25\}$ typically suffices to obtain a relatively clear saturation of $\langle s \rangle$, hence we present the results for these T values.

For each $\{N_c\ , T\}$ pair we performed 10 different random initializations ending up with 10 (possibly) different local maxima of $\mathcal{F}$, from which we chose the best one. The resulting trade-off curves are presented in the left panel of Figure 4. For a given N_c , as T is lowered, $\langle s \rangle$ increases but so does $I(C;\mathrm{i})$. In addition, the solutions become more deterministic. For example, for $N_c = 20$ and $\frac{1}{T} = 15$, only $\sim\ 44\%$ of the genes have nearly deterministic assignment (i.e., $P(C|i) > 0.9$ for a particular C). For $\frac{1}{T} = 25$ this percentage boosts up to $\sim\ 85\%$.

It is important to realize that the entire continuum of solutions, represented by these curves, may encompass a lot of insights about the data. Nonetheless, for brevity, we will further focus our analysis on solutions for which the saturation of $\langle s \rangle$ is relatively clear, i.e., on the four solutions with $N_c = \{5, 10, 15, 20\}$ and $\frac{1}{T} = 25$. In all these partitions most of the genes (between 75% to 85%) had a nearly deterministic assignment ($P(C|i) > 0.9$ for a particular C). Hence, for further simplicity, in the rest of the analysis we treat these solutions as "hard" partitions where every gene is assigned solely to its most probable cluster. In the next section we explore the possible hierarchical relations between these four solutions. In later sections we analyze in detail the specific solution with $\{N_c = 20,\ \frac{1}{T} = 25\}$ that obtained the highest $\langle s \rangle$ value.

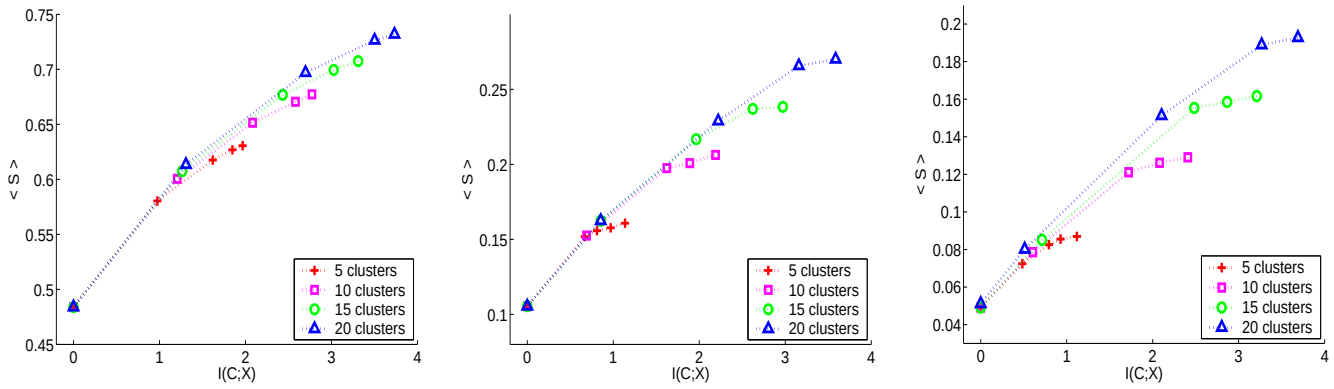

Figure 4: **(Left)** Tradeoff curves obtained for the ESR data. Each curve describes the solutions obtained for a particular N_c value, i.e., for a fi xed number of clusters. Different points along each curve correspond to different local maxima of $\mathcal{F}$ at different T values. The results are presented for $\frac{1}{T} = \{5, 10, 15, 20, 25\}$ which suffi ces to obtain a relatively clear saturation of the average mutual information per cluster, $\langle s \rangle$. In Section 4.4 we explore the possible hierarchical relations between the four "saturated" solutions at $\frac{1}{T} = 25$ and $N_c = \{5, 10, 15, 20\}$. Further detailed analysis refers to the solution with $N_c = 20$ and $\frac{1}{T} = 25$ that obtained the highest $\langle s \rangle$ value. **(Middle)** Similar tradeoff curves that were obtained for the SP500 data. The results are presented for $\frac{1}{T} = \{15, 20, 25, 30, 35\}$ which suffi ces to obtain a relatively clear saturation of $\langle s \rangle$. Notice, that due to the lower average mutual information relations in these data (with respect to the ESR example), one must apply lower T values to obtain a clear saturation. In Section 5.4 we explore the possible hierarchical relations between the four "saturated" solutions at $\frac{1}{T} = 35$ and $N_c = \{5, 10, 15, 20\}$. Further detailed analysis refers to the solution with $N_c = 20$ and $\frac{1}{T} = 35$. **(Right)** Similar tradeoff curves that were obtained for the EachMovie data. The results are presented for $\frac{1}{T} = \{20, 25, 30, 35, 40\}$ which suffi ces to obtain a relatively clear saturation of $\langle s \rangle$. In Section 6.4 we explore the possible hierarchical relations between the four "saturated" solutions at $\frac{1}{T} = 40$ and $N_c = \{5, 10, 15, 20\}$. Further detailed analysis refers to the solution with $N_c = 20$ and $\frac{1}{T} = 40$.

4.4 Comparing solutions at different numbers of clusters

A common dichotomy in the cluster analysis literature is between hierarchical versus non-hierarchical, or "partitional" clustering algorithms (see, e.g., (*10*)). What is often missed, though, is the fact that applying a hierarchical clustering algorithm over a given data typically enforces the output to be of hierarchical nature, regardless of whether the data structure indeed calls for this view. For example, applying an agglomerative clustering algorithm over the ESR data will produce, by definition, a nested "tree-like" hierarchy of partitions, although *a priori* it is not obvious whether a functional classification of these genes should be of a hierarchical nature or not.

In principle, this issue calls for a theoretical analysis which is out of the scope of this work. Nonetheless, since our main theme is to make as little assumptions as possible, we clearly do not want to enforce a hierarchical output structure but rather to examine directly this issue, from an unbiased perspective.

To this end we apply the following simple scheme. Given several solutions *that were found independently* at different numbers of clusters, we ask to what extent these solutions form a hierarchy. This is done in two steps. First, for every cluster we identify its "best parent" in the next (less detailed) level. Specifically, if C is some cluster at a partition with N_c clusters, then its "best parent" in a less detailed partition with $N'_c < N_c$ clusters will be the one that includes the maximal number of C members. Second, we visualize how well this "parent" includes its "son" through the type of the edge that we draw between the two clusters.

The hierarchical graph produced by this scheme is different from the standard output of hierarchical clustering algorithms in several aspects. To start, a cluster might have more than one parent if its members are equally distributed among several clusters in the less detailed partition. Next, a cluster might have no sons if it is not the "best parent" of any cluster at the more detailed partition. Last, the characteristics of the edges convey further information regarding how well the independent solutions form a hierarchy. In particular, a graph with many high quality inclusion edges is a good indication that the given data is hierarchical in nature. On the other hand, a graph in which many of the inclusions from one level to the other are only partial, suggests that a hierarchical view of the data is somewhat misleading.

We applied this scheme to the four solutions we obtained independently for $N_c = \{5, 10, 15, 20\}$ with $\frac{1}{T} = 25$. The results are presented in Figure 5. Here, we distinguish between 4 discrete levels of inclusion that correspond to 4 edge types. As can be seen in the figure, the independent solutions form an "approximated" hierarchical structure. Interestingly, some functional modules are better preserved than others across the different levels. For example, the "ribosome cluster", $c18$, is clearly identified at all the independent solutions.

There are clearly different ways to implement the above idea. For example, the width of the edge can reflect the percentage of the the "son" members that are absorbed in the "best parent" cluster in a continuous manner. In addition, one can draw an edge from a cluster to all its parents in the less detailed partition, not just to the one that absorbs its most significant portion, and so on. Nonetheless, we found the above, relatively simple, presentation satisfactory for our needs.

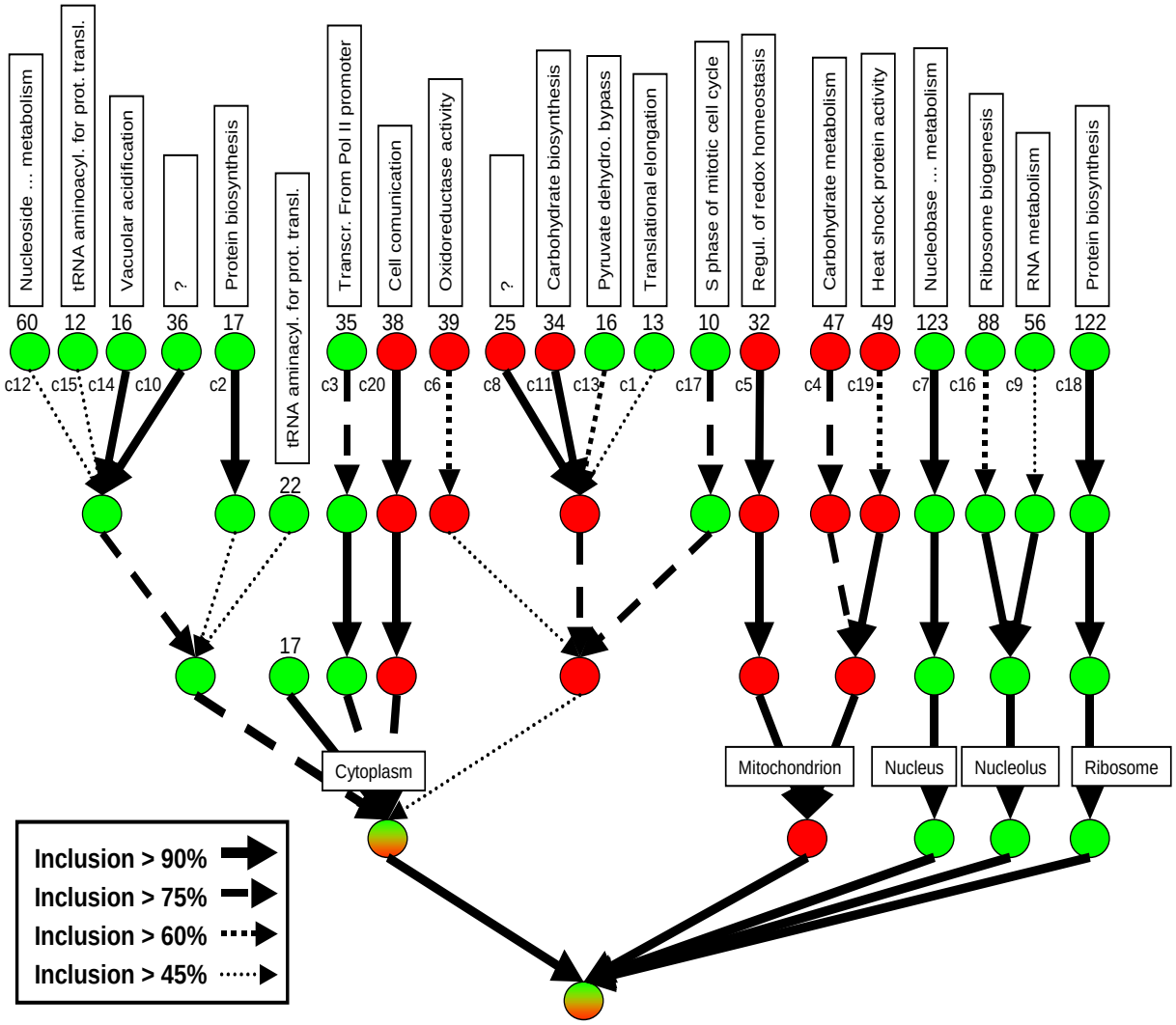

Figure 5: Relations between the optimal solutions with $N_c = \{5, 10, 15, 20\}$ at $\frac{1}{T} = 25$ for the ESR data. At the upper level, $N_c = 20$ clusters, and the clusters are sorted as in Figure 2 and Figure 3. The numbers above every cluster indicate the number of genes in this cluster. The title of each cluster correspond to the most enriched GO_{BP} (biological process) annotation in the cluster, i.e., to the GO_{BP} annotation with the smallest P-value in the cluster (see Section 4.5.1). The only exceptions are $c6$, not enriched in GO_{BP}, and $c19$, enriched with a non-informative annotation ("response to stress"). For these two clusters we use their most enriched GO_{MF} (molecular function) annotation as a title. The titles of the five clusters at the lower level ($N_c = 5$) are by their most enriched GO_{CC} (cellular component) annotation. Notice, that most clusters were enriched with more than one annotation, hence the short titles might be too concise in some cases (see Section 4.6 for a detailed description of every cluster at the top level). Red and green clusters represent clusters with a clear majority of stress–induced or stress–repressed genes, respectively. In the "cytoplasm" cluster we had a relatively balanced mixture of stress-repressed (58%) and stress-induced (42%) genes.

Table 3: A small subset of the $\mathbf{A}_{BP}$ annotation matrix, constructed for the ESR data out of the GO_{BP} ontology.

ORF	**Metabolism**	**Transcription**	**RNA processing**	**Ribosome biogenesis**	**...**
YKL144C	1	1	0	0	...
YML060W	1	0	0	0	...
YGR251W	1	1	1	1	...
YLL036C	1	0	1	0	...
YNL163C	0	0	0	1	...
...	...	...	...	...	...

4.5 Coherence results

4.5.1 Constructing the annotation matrices

As already mentioned, clusters' coherence is estimated with respect to a given annotation matrix. For yeast genes, different sources of information may provide these data. One such important resource is the Gene Ontology (GO) database (*11*) which is the one that we use in this work (specifically, we used the December 2003 version).

The GO database consists of three structured ontologies (controlled vocabularies) that describe gene products in terms of their associated Biological Processes (the GO_{BP} ontology), Molecular Functions (the GO_{MF} ontology), and Cellular Components (the GO_{CC} ontology). For each of these three ontologies we constructed a corresponding annotation matrix. Thus, for example, if $\mathbf{A}_{BP}$ is the matrix constructed for the GO_{BP} ontology then $\mathbf{A}_{BP}(i,j) = 1$ if and only if the i-th gene in our data is assigned with the j-th biological process in this ontology. A small subset of this annotation matrix is represented in Table 3.

Each of the GO ontologies is organized in a hierarchical manner where more specific annotations correspond to nodes which are more distant from the ontology "root". This might yield evaluation difficulties if one considers only the particular GO terms with which a gene is annotated (*14*). An example is given in Table 4. Here, several genes that were all assigned in the same cluster are annotated with different specific GO_{BP} terms and their functional relationship becomes evident only if one notices that all these annotations have a common (more general) ancestor in the ontology. We applied a standard routine to overcome this difficulty where every gene was assigned not only with its direct GO annotations but also with all the ancestors of these annotations in the GO hierarchy. This is consistent

Table 4: An example for a subset of genes from a single cluster that are assigned with different specific GO_{BP} terms. The functional relationship between these genes becomes statistically significant only if one considers the fact that all these annotations have a common ancestor in the GO_{BP} database, the "tRNA aminoacylation for protein translation" term.

ORF	**Direct GO_{BP} annotation**
YDR037W	lysyl-tRNA aminoacylation
YGR094W	valyl-tRNA aminoacylation
YLR060W	phenylalanyl-tRNA aminoacylation
YNEuclidean47W	cysteinyl-tRNA aminoacylation
YPL160W	leucyl-tRNA aminoacylation

Table 5: Details of the different annotation matrices used for evaluating the statistical significance of the obtained clusters for the yeast ESR genes. [a]Data source for constructing the annotation matrix. [b]Number of distinct annotations in the annotation matrix, assigned with at least two genes and thus participate in the analysis. [c]Number of genes assigned with at least one annotation and thus participate in the analysis. Notice that this number determines the "population size" (N) for the P-value estimation. [d]Average number of distinct annotations per gene. [e]Maximal number of distinct annotations for a single gene.

Data source[a]	**# Annotations**[b]	**# Genes**[c]	**Avg. # Annot. per gene**[d]	**Maximal # Annot. per gene**[e]
GO_{BP} (*11*)	472	614	11.4	63
GO_{MF} (*11*)	215	561	4.6	18
GO_{CC} (*11*)	94	747	5.4	14

with the GO terminology in which if a GO term describes some gene product then all its parent terms in the ontology also apply to that gene product.

Last, while estimating clusters' coherence we removed annotations that were assigned with less than two genes in our data (since these annotations obviously can not be enriched in any cluster). We also removed from the analysis genes that were not assigned with any annotation (or assigned with the "unknown" annotation). The details of the resulting annotation matrices are given in Table 5.

4.5.2 Coherence results and comparison to other clustering algorithms

We estimated the statistical coherence of the clusters obtained at the low-temperature end of the trade-off curves where $\frac{1}{T} = 25$. This coherence was estimated with respect to each of the three Gene Ontologies. To gain some perspective, we applied similar analysis with the most recent release of the "Cluster" software, called "Cluster 3.0" (*12*). This software is considered to be a state-of-the-art (and quite popular) tool for cluster analysis of gene expression data. We

experimented extensively with all the basic algorithms available by this software. These include two different variants of iterative K–means clustering ("K–means" and "K–medians") and four different variants of hierarchical clustering ("Complete linkage", "Average linkage", "Centroid linkage", and "Single linkage"). With each of these algorithms we tried three standard similarity measures: the Pearson correlation ("centered correlation") (*13*), the absolute value of the Pearson correlation, and the Euclidean distance. Thus, altogether we compared our performance to 18 different configurations of this software which are probably the most commonly used configurations. For the 6 "K–means" variants we tried 100 different random initializations in every run, from which the best solution (with the smallest sum of within-cluster distances) is chosen. The comparison was undertaken to all the different numbers of clusters, $N_c = 5, 10, 15, 20$. The results are summarized in Table 6 to Table 9. The average results are given Figure 6.

In all cases the Iclust algorithm was clearly superior to *all* the 12 hierarchical algorithms we tried. It should be stressed that these algorithms are considered a powerful tool for analyzing genomic datasets, and many published applications are based on this type of hierarchical analysis. Nonetheless, standard hierarchical clustering typically failed to see a significant sub-structure in the ESR module. In most cases Iclust was also superior to the average performance of the 6 K–means variants, and in some cases (e.g., $N_c = 5$) it was in fact superior to all the K–means variants. Averaging over all three Gene Ontologies and over all four N_c values, Iclust obtains a coherence of $\sim 56\%$ while the average K–means coherence is $\sim 42\%$ and the average Hierarchical coherence is $\sim 12\%$.

We further repeated this comparison with the 18 competing algorithms while considering the log_2 ratios of expression profiles as input, instead of the raw ratios. Even under this pre–process (to which our approach is invariant), the Iclust average performance is superior to almost all the 18 alternatives, typically by a significant margin. Specifically, when averaging over all three Gene Ontologies and over all four N_c values, the average K–means coherence is $\sim 52\%$ and the average Hierarchical coherence is $\sim 19\%$. While there exists some intuitive motivation for the log_2 pre–process it is certainly not clear what is the formal justification of this step. Clearly, from a principled point of view, a clustering approach which is invariant to such transformations is preferable.

Table 6: Coherence results for the ESR data with respect to the three Gene Ontologies with $N_c = 20$ clusters. [a]Clustering algorithm. In the "$\langle$ $K-$means $\rangle$" row we present the average results of all the 6 $K-$means variants. For each of these variants we performed 100 runs from which the best solution is chosen. In the "$\langle$ Hier. $\rangle$" row we present the average results of all the 12 Hierarchical clustering variants. In parenthesis we present the results where the input are the log_2 of the expression ratio profiles. [b]Correlation measure used by the algorithm. "PC" stands for the (centered) Pearson Correlation. "|PC|" is the absolute value of this correlation. "Euclidean" stands for the Euclidean distance. [c]Number of clusters with a positive coherence with respect to the GO_{BP} ontology. [d]Average coherence of all 20 clusters with respect to the GO_{BP} ontology. [e]Number of clusters with a positive coherence with respect to the GO_{MF} ontology. [f]Average coherence of all 20 clusters with respect to the GO_{MF} ontology. [g]Number of clusters with a positive coherence with respect to the GO_{CC} ontology. [h]Average coherence of all 20 clusters with respect to the GO_{CC} ontology.

$N_c = 20$		**BP**	**BP**	**MF**	**MF**	**CC**	**CC**
Algorithm [a]	**Similarity** [b]	N_c^{pos} [c]	$\langle Coh \rangle$ [d]	N_c^{pos} [e]	$\langle Coh \rangle$ [f]	N_c^{pos} [g]	$\langle Coh \rangle$ [h]
Iclust	mutual information	17	51	16	41	14	33
$K-$means	PC	11 (13)	30 (43)	11 (11)	31 (31)	10 (12)	19 (30)
$K-$means	|PC|	9 (15)	27 (50)	8 (14)	24 (40)	8 (16)	26 (42)
$K-$means	Euclidean	7 (15)	23 (52)	9 (15)	26 (39)	5 (16)	13 (51)
$K-$medians	PC	11 (15)	35 (51)	13 (16)	34 (48)	10 (15)	35 (46)
$K-$medians	|PC|	12 (15)	38 (41)	16 (16)	43 (39)	13 (11)	37 (35)
$K-$medians	Euclidean	16 (18)	49 (52)	15 (14)	39 (44)	13 (16)	43 (51)
$\langle$ $K-$means $\rangle$		11.0 (15.2)	33.7 (48.2)	12.0 (14.3)	32.8 (40.2)	9.8 (14.3)	28.8 (42.5)
Hier - Comp. linkage	PC	9 (13)	29 (41)	10 (10)	25 (30)	7 (12)	19 (34)
Hier - Comp. linkage	|PC|	9 (10)	25 (26)	12 (9)	31 (27)	7 (10)	17 (26)
Hier - Comp. linkage	Euclidean	1 (13)	2 (43)	3 (11)	8 (32)	1 (8)	2 (27)
Hier - Avg. linkage	PC	5 (7)	17 (20)	5 (5)	18 (17)	4 (4)	11 (12)
Hier - Avg. linkage	|PC|	5 (4)	17 (10)	5 (2)	18 (8)	4 (2)	10 (4)
Hier - Avg. linkage	Euclidean	1 (9)	2 (29)	1 (4)	1 (17)	2 (6)	6 (16)
Hier - Centr. linkage	PC	4 (3)	12 (10)	4 (3)	12 (10)	4 (2)	11 (8)
Hier - Centr. linkage	|PC|	4 (4)	12 (12)	3 (4)	7 (11)	4 (2)	9 (4)
Hier - Centr. linkage	Euclidean	0 (4)	0 (13)	0 (4)	0 (12)	1 (2)	1 (8)
Hier - Sing. linkage	PC	2 (2)	8 (8)	2 (2)	7 (7)	2 (2)	8 (8)
Hier - Sing. linkage	|PC|	2 (0)	6 (0)	1 (0)	5 (0)	0 (0)	0 (0)
Hier - Sing. linkage	Euclidean	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
$\langle$ Hier. $\rangle$		3.5 (5.8)	10.8 (17.7)	3.8 (4.5)	11.0 (14.2)	3.0 (4.2)	7.8 (12.2)

Table 7: Coherence results for the ESR data with respect to the three Gene Ontologies with $N_c = 15$ clusters. The column and row definitions are as in Table 6. Again, in parenthesis we present the results where the input are the log_2 of the expression ratio profiles.

$N_c = 15$		**BP**	**BP**	**MF**	**MF**	**CC**	**CC**
Algorithm [a]	**Similarity** [b]	N_c^{pos} [c]	$\langle Coh \rangle$ [d]	N_c^{pos} [e]	$\langle Coh \rangle$ [f]	N_c^{pos} [g]	$\langle Coh \rangle$ [h]
Iclust	mutual information	12	51	14	54	14	52
$K-$means	PC	7 (14)	29 (55)	8 (13)	32 (49)	7 (10)	18 (38)
$K-$means	$\|$PC$\|$	10 (14)	40 (47)	9 (11)	32 (37)	8 (12)	27 (38)
$K-$means	Euclidean	10 (12)	33 (50)	8 (13)	36 (46)	3 (11)	14 (44)
$K-$medians	PC	11 (13)	40 (46)	11 (13)	41 (49)	10 (14)	41 (47)
$K-$medians	$\|$PC$\|$	11 (14)	42 (50)	11 (13)	31 (44)	10 (11)	35 (38)
$K-$medians	Euclidean	11 (14)	50 (58)	12 (13)	42 (43)	11 (13)	46 (61)
$\langle$ $K-$means $\rangle$		10.0 (13.5)	39.0 (51.0)	9.8 (12.7)	35.7 (44.7)	8.2 (11.8)	30.2 (44.3)
Hier - Comp. linkage	PC	8 (11)	32 (43)	9 (8)	31 (31)	6 (9)	20 (44)
Hier - Comp. linkage	$\|$PC$\|$	4 (8)	17 (29)	7 (7)	21 (29)	5 (8)	15 (32)
Hier - Comp. linkage	Euclidean	0 (8)	0 (33)	1 (8)	2 (29)	1 (6)	2 (27)
Hier - Avg. linkage	PC	5 (5)	21 (22)	4 (5)	18 (21)	4 (3)	13 (12)
Hier - Avg. linkage	$\|$PC$\|$	4 (4)	15 (13)	3 (3)	11 (12)	3 (2)	5 (5)
Hier - Avg. linkage	Euclidean	2 (7)	8 (36)	1 (4)	1 (22)	2 (4)	8 (14)
Hier - Centr. linkage	PC	4 (3)	16 (13)	4 (3)	16 (15)	4 (3)	14 (12)
Hier - Centr. linkage	$\|$PC$\|$	4 (3)	16 (11)	3 (3)	7 (11)	4 (3)	11 (6)
Hier - Centr. linkage	Euclidean	0 (3)	0 (15)	0 (3)	0 (11)	0 (2)	0 (11)
Hier - Sing. linkage	PC	2 (2)	11 (11)	2 (2)	9 (9)	2 (2)	11 (11)
Hier - Sing. linkage	$\|$PC$\|$	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Hier - Sing. linkage	Euclidean	0 (0)	0 (0)	0 (0)	0 (0)	1 (0)	6 (0)
$\langle$ Hier. $\rangle$		2.8 (4.5)	11.3 (18.8)	2.8 (3.8)	9.7 (15.8)	2.7 (3.5)	8.8 (14.5)

Table 8: Coherence results for the ESR data with respect to the three Gene Ontologies with $N_c = 10$ clusters. The column and row definitions are as in Table 6. Again, in parenthesis we present the results where the input are the log_2 of the expression ratio profiles.

$N_c = 10$		**BP**	**BP**	**MF**	**MF**	**CC**	**CC**
Algorithm [a]	**Similarity** [b]	N_c^{pos} [c]	$\langle Coh \rangle$ [d]	N_c^{pos} [e]	$\langle Coh \rangle$ [f]	N_c^{pos} [g]	$\langle Coh \rangle$ [h]
Iclust	mutual information	7	50	7	43	9	53
K–means	PC	8 (9)	45 (52)	8 (8)	44 (47)	7 (9)	37 (56)
K–means	|PC|	7 (9)	41 (48)	6 (8)	42 (41)	8 (8)	39 (48)
K–means	Euclidean	5 (10)	27 (62)	6 (10)	30 (57)	3 (8)	22 (55)
K–medians	PC	9 (10)	51 (57)	8 (9)	45 (53)	9 (10)	49 (54)
K–medians	|PC|	7 (9)	45 (52)	8 (9)	50 (47)	9 (8)	41 (56)
K–medians	Euclidean	7 (9)	48 (62)	8 (9)	46 (60)	7 (9)	49 (58)
$\langle$ K–means $\rangle$		7.2 (9.3)	42.8 (55.5)	7.3 (8.8)	42.8 (50.8)	7.2 (8.7)	39.5 (54.5)
Hier - Comp. linkage	PC	6 (8)	33 (44)	7 (5)	43 (32)	5 (7)	26 (43)
Hier - Comp. linkage	|PC|	4 (6)	24 (33)	6 (5)	32 (37)	5 (6)	22 (30)
Hier - Comp. linkage	Euclidean	2 (7)	12 (41)	2 (7)	8 (39)	2 (5)	7 (32)
Hier - Avg. linkage	PC	3 (4)	19 (30)	3 (4)	20 (30)	3 (3)	18 (19)
Hier - Avg. linkage	|PC|	2 (4)	8 (20)	1 (3)	7 (19)	1 (2)	1 (7)
Hier - Avg. linkage	Euclidean	0 (4)	0 (33)	0 (5)	0 (28)	0 (4)	0 (20)
Hier - Centr. linkage	PC	3 (3)	19 (19)	3 (3)	21 (20)	3 (3)	18 (18)
Hier - Centr. linkage	|PC|	4 (3)	19 (21)	3 (3)	11 (17)	4 (3)	16 (9)
Hier - Centr. linkage	Euclidean	0 (3)	0 (21)	1 (3)	2 (17)	0 (3)	0 (9)
Hier - Sing. linkage	PC	2 (2)	16 (16)	2 (2)	13 (14)	2 (2)	17 (17)
Hier - Sing. linkage	|PC|	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Hier - Sing. linkage	Euclidean	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
$\langle$ Hier. $\rangle$		2.2 (3.7)	12.5 (23.2)	2.3 (3.3)	13.1 (21.1)	2.1 (3.2)	10.4 (17.0)

Table 9: Coherence results for the ESR data with respect to the three Gene Ontologies with $N_c = 5$ clusters. The column and row definitions are as in Table 6. Again, in parenthesis we present the results where the input are the log_2 of the expression ratio profiles.

$N_c = 5$		**BP**	**BP**	**MF**	**MF**	**CC**	**CC**
Algorithm [a]	**Similarity** [b]	N_c^{pos} [c]	$\langle Coh \rangle$ [d]	N_c^{pos} [e]	$\langle Coh \rangle$ [f]	N_c^{pos} [g]	$\langle Coh \rangle$ [h]
Iclust	mutual information	5	75	5	77	5	86
K–means	PC	5 (5)	62 (65)	5 (5)	63 (65)	5 (5)	75 (73)
K–means	|PC|	5 (5)	61 (70)	5 (5)	62 (67)	5 (5)	75 (70)
K–means	Euclidean	3 (5)	43 (71)	3 (5)	35 (56)	3 (4)	39 (65)
K–medians	PC	5 (5)	64 (62)	5 (5)	65 (63)	5 (5)	72 (72)
K–medians	|PC|	5 (5)	57 (59)	5 (5)	52 (58)	5 (5)	75 (69)
K–medians	Euclidean	4 (5)	52 (71)	4 (5)	59 (60)	4 (4)	68 (57)
$\langle$ K–means $\rangle$		4.5 (5.0)	56.5 (66.3)	4.5 (5.0)	56.0 (61.5)	4.5 (4.7)	67.3 (67.7)
Hier - Comp. linkage	PC	4 (4)	42 (44)	5 (4)	52 (46)	4 (4)	37 (57)
Hier - Comp. linkage	|PC|	4 (4)	47 (51)	5 (3)	34 (44)	3 (4)	30 (45)
Hier - Comp. linkage	Euclidean	1 (3)	11 (37)	2 (4)	13 (49)	0 (4)	0 (36)
Hier - Avg. linkage	PC	3 (3)	38 (39)	3 (3)	40 (47)	3 (3)	36 (37)
Hier - Avg. linkage	|PC|	0 (1)	0 (6)	0 (1)	0 (13)	0 (2)	0 (8)
Hier - Avg. linkage	Euclidean	0 (2)	0 (31)	0 (3)	0 (30)	0 (2)	0 (33)
Hier - Centr. linkage	PC	3 (2)	39 (32)	3 (2)	41 (27)	3 (2)	36 (33)
Hier - Centr. linkage	|PC|	3 (1)	21 (8)	2 (0)	19 (0)	3 (1)	13 (6)
Hier - Centr. linkage	Euclidean	0 (1)	0 (8)	0 (0)	0 (0)	0 (1)	0 (6)
Hier - Sing. linkage	PC	2 (2)	32 (32)	2 (2)	27 (27)	2 (2)	33 (33)
Hier - Sing. linkage	|PC|	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Hier - Sing. linkage	Euclidean	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
$\langle$ Hier. $\rangle$		1.7 (1.9)	19.2 (24.0)	1.8 (1.8)	18.8 (23.6)	1.5 (2.1)	15.4 (24.5)

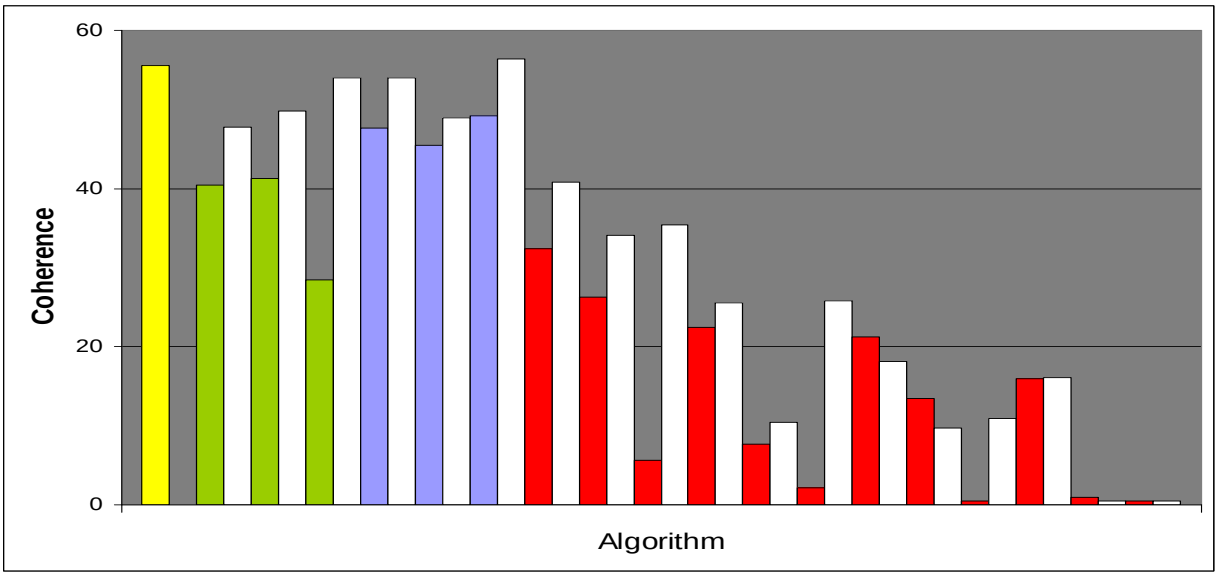

Figure 6: **ESR data**: Comparison of average coherence results of the Iclust algorithm (yellow) with conventional clustering algorithms (*12*): K–means (green); K–medians (blue); Hierarchical (red). For the hierarchical algorithms, four different variants are tried: complete, average, centroid, and single linkage, respectively from left to right. For every algorithm, three different similarity measures are applied: Pearson correlation (left); absolute value of Pearson correlation (middle); Euclidean distance (right). The white bars correspond to applying the algorithm over the $\log_2$ transformation of the expression ratios. In all cases, the results are averaged over all the different numbers of clusters that we tried: $N_c = 5, 10, 15, 20$, and over the three Gene Ontologies.
.

4.6 Detailed results for the $N_c = 20$ clusters partition

In Table 10 we present all the enriched annotations for the Iclust partition with $N_c = 20$ clusters and $\frac{1}{T} = 25$. Further examination of these clusters yields several immediate observations.

First, in several cases the extracted clusters consist of both induced and repressed genes which are negatively correlated. For example, $c5$ consists of 26 induced genes (enriched with "oxidoreductase activity") and 6 repressed ones. In Figure 7A we see that the genes in this cluster have a relatively augmented response under Menadione exposure as opposed to a reduced response in a stationary phase.

In Figure 7B we display the average behavior of the 22 induced genes in $c8$ versus the 49 induced genes in $c19$ in two opposing temperature shifts. Clearly, the genes in $c19$ are more sensitive to this treatment which is consistent

Table 10: Enriched GO annotations in the Iclust solution with $N_c = 20$ clusters. Clusters are ordered as in Figure 2, Figure 3, and Figure 5. Only annotations with a P-value below 0.05 (Bonferroni corrected) are presented. [a]Cluster index, number of "repressed genes" in the cluster, and number of "induced" genes in the cluster. [b]Cluster coherence (in percentage) in each of the three Gene Ontologies. [c]Enriched annotations for the GO_{BP} ontology. In parentheses: $(x/K, p)$ stands for the number of genes in the cluster to which this annotations is assigned, the number of genes in the ESR module to which this annotations is assigned, and the Bonferroni corrected P-value, respectively. [d]Enriched annotations for the GO_{MF} ontology. Parenthesis are as in the previous column. [e]Enriched annotations for the GO_{CC} ontology. Parenthesis are as in the previous column.

C index[a]	**Coh.**[b]	**BP Enriched annot.**[c]	**MF Enriched annot.**[d]	**CC Enriched annot.**[e]
c12 Rep : 57 Ind : 3	BP : 18 MF : 9 CC : 0	nucleoside monophosphate metabolism (5/6,0.0032) purine nucleoside monophosphate meta (5/6,0.0032) ribonucleoside monophosphate metabol (5/6,0.0032) purine ribonucleoside monophosphate (5/6,0.0032) histidine biosynthesis (4/4,0.0076) histidine metabolism (4/4,0.0076) histidine family amino acid metaboli (4/4,0.0076) histidine family amino acid biosynth (4/4,0.0076) nucleoside monophosphate biosynthesi (4/4,0.0076) purine nucleoside monophosphate bios (4/4,0.0076) ribonucleoside monophosphate biosynt (4/4,0.0076) purine ribonucleoside monophosphate (4/4,0.0076) biogenic amine biosynthesis (4/4,0.0076) amine biosynthesis (7/16,0.0187) amino acid derivative biosynthesis (4/5,0.0359)	diphosphotransferase activity (4/4,0.0047) – – – – – – – – – – – – – – – –	– – – – – – – – – – – – – – – –
c15 Rep : 12 Ind : 0	BP : 67 MF : 75 CC : 0	tRNA aminoacylation for protein tran (8/14,0.0000) – – – – –	tRNA ligase activity (9/17,0.0000) RNA ligase activity (9/17,0.0000) ligase activity, forming carbon-oxyg (9/17,0.0000) ligase activity, forming aminoacyl-t (9/17,0.0000) ligase activity, forming phosphoric (9/17,0.0000) ligase activity (9/23,0.0000)	– – – – – –
c14 Rep : 15 Ind : 1	BP : 57 MF : 43 CC : 53	vacuolar acidifi cation (3/3,0.0010) regulation of pH (3/4,0.0038) monovalent inorganic cation homeosta (3/4,0.0038) hydrogen ion homeostasis (3/4,0.0038) organelle organization and biogenesi (6/38,0.0080) vacuolar transport (3/5,0.0093) cation homeostasis (3/7,0.0317) asparagine biosynthesis (2/2,0.0488) aspartate family amino acid biosynth (2/2,0.0488) – –	hydrogen-transporting ATPase activit (3/3,0.0005) ion transporter activity (4/8,0.0006) monovalent inorganic cation transpor (3/4,0.0020) hydrogen ion transporter activity (3/4,0.0020) ATPase activity, coupled to transmem (3/5,0.0048) cation transporter activity (3/6,0.0095) asparagine synthase (glutamine-hydro) (2/2,0.0232) primary active transporter activity (3/9,0.0382) P-P-bond-hydrolysis-driven transport (3/9,0.0382) hydrolase activity, acting on acid a (3/9,0.0382) ATPase activity, coupled to transmem (3/9,0.0382)	chaperonin-containing T-complex (3/3,0.0001) hydrogen-translocating V-type ATPase (3/3,0.0001) vacuolar membrane (3/8,0.0073) cytoskeleton (3/8,0.0073) hydrogen-transporting ATPase V0 doma (2/2,0.0079) membrane (5/47,0.0312) – – – – –
c10 Rep : 35 Ind : 1	BP : 0 MF : 0 CC : 0	– – –	– – –	– – –
c2 Rep : 17 Ind : 0	BP : 100 MF : 80 CC : 100	protein biosynthesis (15/189,0.0000) macromolecule biosynthesis (15/189,0.0000) biosynthesis (15/236,0.0000) translational elongation (5/10,0.0000) protein metabolism (15/252,0.0000) – –	structural constituent of ribosome (12/127,0.0001) structural molecule activity (12/128,0.0001) – – – – –	ribosome (14/153,0.0000) ribonucleoprotein complex (14/186,0.0000) cytosolic large ribosomal subunit (10/75,0.0000) large ribosomal subunit (10/75,0.0000) cytosolic ribosome (sensu Eukarya) (12/132,0.0000) cytosol (12/165,0.0001) cytoplasm (16/525,0.0430)
c3 Rep : 29 Ind : 6	BP : 17 MF : 15 CC : 23	transcription from Pol II promoter (5/11,0.0077) – –	ribonuclease activity (4/10,0.0387) – –	DNA-directed RNA polymerase II, core (4/7,0.0036) DNA-directed RNA polymerase II, holo (4/9,0.0122) cytoplasmic exosome (RNase complex) (3/5,0.0282)
c20 Rep : 0 Ind : 38	BP : 25 MF : 0 CC : 22	cell communication (5/18,0.0147) signal transduction (4/12,0.0359) – –	– – – –	vacuole (7/26,0.0014) storage vacuole (5/20,0.0288) lytic vacuole (5/20,0.0288) vacuole (sensu Fungi) (5/20,0.0288)

C index[a]	Coh.[b]	BP Enriched annot.[c]	MF Enriched annot.[d]	CC Enriched annot.[e]
c6 Rep : 3 Ind : 36	BP : 0 MF : 33 CC : 7	– – –	oxidoreductase activity (6/37,0.0255) – –	peroxisomal matrix (2/2,0.0353) – –
c8 Rep : 3 Ind : 22	BP : 0 MF : 0 CC : 0	– – –	– – –	– – –
c11 Rep : 0 Ind : 34	BP : 63 MF : 58 CC : 100	carbohydrate biosynthesis (5/8,0.0000) gluconeogenesis (4/5,0.0002) hexose biosynthesis (4/5,0.0002) alcohol biosynthesis (4/5,0.0002) monosaccharide biosynthesis (4/5,0.0002) regulation of carbohydrate metabolis (3/3,0.0015) regulation of gluconeogenesis (3/3,0.0015) regulation of biosynthesis (3/3,0.0015) negative regulation of biosynthesis (3/3,0.0015) negative regulation of metabolism (3/3,0.0015) negative regulation of gluconeogenes (3/3,0.0015) negative regulation of carbohydrate (3/3,0.0015) energy pathways (6/25,0.0015) energy derivation by oxidation of or (6/25,0.0015) protein amino acid phosphorylation (4/8,0.0021) phosphorylation (4/9,0.0037) main pathways of carbohydrate metabo (4/11,0.0093) phosphorus metabolism (4/11,0.0093) phosphate metabolism (4/11,0.0093) carbohydrate metabolism (6/35,0.0120) glucose metabolism (4/12,0.0138) regulation of metabolism (3/6,0.0284) hexose metabolism (4/15,0.0363)	protein kinase activity (5/12,0.0000) phosphotransferase activity, alcohol (5/19,0.0004) kinase activity (5/27,0.0027) protein serine/threonine kinase acti (3/6,0.0035) cyclic-nucleotide dependent protein (2/2,0.0101) cAMP-dependent protein kinase activi (2/2,0.0101) transferase activity (7/103,0.0496) –	lipid particle (3/5,0.0029) cAMP-dependent protein kinase comple (2/2,0.0123) cytoplasm (20/525,0.0139) –
c13 Rep : 16 Ind : 0	BP : 40 MF : 20 CC : 0	pyruvate dehydrogenase bypass (3/3,0.0002) pyruvate metabolism (3/3,0.0002) fermentation (2/2,0.0189) ethanol fermentation (2/2,0.0189) glycolytic fermentation (2/2,0.0189) alcohol metabolism (4/25,0.0303)	pyruvate decarboxylase activity (2/3,0.0264) – – – – –	– – – – – –
c1 Rep : 13 Ind : 0	BP : 73 MF : 36 CC : 20	translational elongation (4/10,0.0004) nascent polypeptide association (2/2,0.0096) methionine metabolism (2/2,0.0096) serine family amino acid metabolism (2/3,0.0286)	translation elongation factor activi (3/7,0.0032) translation factor activity, nucleic (4/29,0.0253) translation regulator activity (4/31,0.0328) –	nascent polypeptide-associated compl (2/2,0.0021) – – –
c17 Rep : 7 Ind : 3	BP : 80 MF : 33 CC : 25	S phase of mitotic cell cycle (4/10,0.0000) DNA replication (4/10,0.0000) DNA replication and chromosome cycle (4/14,0.0000) mitotic cell cycle (4/14,0.0000) cell cycle (4/22,0.0002) DNA metabolism (4/24,0.0003) DNA dependent DNA replication (3/8,0.0004) cell proliferation (4/27,0.0004) DNA replication, priming (2/2,0.0016) DNA replication initiation (2/3,0.0048) lagging strand elongation (2/3,0.0048) DNA strand elongation (2/4,0.0095) DNA repair (2/8,0.0438)	alpha DNA polymerase activity (2/2,0.0025) DNA-directed DNA polymerase activity (2/4,0.0148) – – – – – – – – – – – – –	alpha DNA polymerase:primase complex (2/2,0.0010) replication fork (2/4,0.0060) – – – – – – – – – – – – –
c5 Rep : 6 Ind : 26	BP : 24 MF : 45 CC : 11	regulation of redox homeostasis (3/5,0.0358) cell redox homeostasis (3/5,0.0358) oxygen and reactive oxygen species m (4/12,0.0455)	oxidoreductase activity (9/37,0.0000) – –	mitochondrial intermembrane space (3/6,0.0167) – –
c4 Rep : 1 Ind : 46	BP : 62 MF : 89 CC : 7	carbohydrate metabolism (11/35,0.0000) response to stress (10/50,0.0050) –	catalytic activity (24/280,0.0008) hydrolase activity, acting on carbon (3/4,0.0201) –	alpha,alpha-trehalose-phosphate synt (3/3,0.0053) – –
c19 Rep : 0 Ind : 49	BP : 45 MF : 48 CC : 0	response to stress (14/50,0.0000) – –	heat shock protein activity (6/8,0.0000) oxidoreductase activity (9/37,0.0031) chaperone activity (6/19,0.0139)	– – –

C index[a]	Coh.[b]	BP Enriched annot.[c]	MF Enriched annot.[d]	CC Enriched annot.[e]
c7 Rep : 114 Ind : 9	BP : 89 MF : 0 CC : 78	nucleobase, nucleoside, nucleotide a (46/188,0.0000) transcription, DNA-dependent (35/117,0.0000) transcription (35/119,0.0000) RNA metabolism (33/119,0.0000) RNA processing (31/110,0.0000) ribosome biogenesis (30/110,0.0001) transcription from Pol I promoter (28/102,0.0002) rRNA processing (25/90,0.0008) ribosome biogenesis and assembly (32/136,0.0013) cellular process (48/263,0.0037) cell growth and/or maintenance (47/256,0.0044) cytoplasm organization and biogenesi (35/172,0.0142) transcription from Pol III promoter (7/13,0.0392) cell organization and biogenesis (37/196,0.0479)	– – – – – – – – – – – – – – –	nucleus (83/288,0.0000) nucleolus (40/118,0.0000) nucleoplasm (13/34,0.0125) DNA-directed RNA polymerase III comp (7/13,0.0295) – – – – – – – – – – –
c16 Rep : 87 Ind : 1	BP : 96 MF : 77 CC : 88	ribosome biogenesis (48/110,0.0000) ribosome biogenesis and assembly (51/136,0.0000) RNA processing (43/110,0.0000) RNA metabolism (44/119,0.0000) cytoplasm organization and biogenesi (52/172,0.0000) transcription from Pol I promoter (40/102,0.0000) rRNA processing (37/90,0.0000) transcription, DNA-dependent (41/117,0.0000) transcription (41/119,0.0000) nucleobase, nucleoside, nucleotide a (51/188,0.0000) cell organization and biogenesis (52/196,0.0000) cellular process (55/263,0.0000) cell growth and/or maintenance (54/256,0.0000) processing of 20S pre-rRNA (16/37,0.0001) ribosomal large subunit biogenesis (9/13,0.0002) ribosomal large subunit assembly and (10/26,0.0380)	snoRNA binding (14/25,0.0000) RNA binding (21/70,0.0000) methyltransferase activity (10/20,0.0002) transferase activity, transferring o (10/20,0.0002) RNA helicase activity (9/17,0.0004) ATP dependent RNA helicase activity (8/14,0.0007) ATP dependent helicase activity (8/14,0.0007) RNA dependent ATPase activity (8/14,0.0007) helicase activity (9/18,0.0008) RNA methyltransferase activity (7/11,0.0011) nucleic acid binding (23/109,0.0041) binding (25/129,0.0080) S-adenosylmethionine-dependent methy (6/13,0.0482) – – –	nucleolus (49/118,0.0000) nucleus (68/288,0.0000) small nucleolar ribonucleoprotein co (11/27,0.0007) – – – – – – – – – – – – – –
c9 Rep : 48 Ind : 8	BP : 62 MF : 58 CC : 33	RNA metabolism (17/119,0.0025) RNA processing (16/110,0.0039) nucleobase, nucleoside, nucleotide a (21/188,0.0083) transcription, DNA-dependent (15/117,0.0369) transcription (15/119,0.0451)	binding (19/129,0.0006) nucleic acid binding (15/109,0.0251) – – –	nucleolus (18/118,0.0336) – – – –
c18 Rep : 122 Ind : 0	BP : 100 MF : 98 CC : 99	protein biosynthesis (112/189,0.0000) macromolecule biosynthesis (112/189,0.0000) biosynthesis (112/236,0.0000) protein metabolism (112/252,0.0000) metabolism (112/523,0.0000) ribosomal small subunit assembly and (8/10,0.0013) regulation of translational fi delity (6/7,0.0082) regulation of translation (8/12,0.0107) ribosomal subunit assembly (15/36,0.0263) – –	structural constituent of ribosome (110/127,0.0000) structural molecule activity (110/128,0.0000) – – – – – – – – –	cytosolic ribosome (sensu Eukarya) (110/132,0.0000) ribosome (110/153,0.0000) cytosol (110/165,0.0000) ribonucleoprotein complex (110/186,0.0000) cytosolic large ribosomal subunit (s (62/75,0.0000) large ribosomal subunit (62/75,0.0000) cytosolic small ribosomal subunit (s (48/56,0.0000) small ribosomal subunit (48/56,0.0000) eukaryotic 48S initiation complex (48/56,0.0000) eukaryotic 43S preinitiation complex (49/61,0.0000) cytoplasm (113/525,0.0000)

with the enrichment of "heat shock protein activity" in this cluster.

Clusters $c18$ consists of 122 repressed genes which were mainly ribosomal proteins. In Figure 7C we see that the genes in this clusters exhibit a distinguished expression "dynamics" under, e.g., Diamide treatment, a fact that was already mentioned in (*3*). On the other hand, cluster $c16$ consists of 87 repressed genes and is enriched for "ribosome biogenesis" and other related annotations. In the same figure we see that this cluster exhibit another distinctive behavior with respect to the rest of the repressed genes.

In Figure 7D we consider again two clusters, $c2$ and $c7$, which seem to involve ribosomal proteins and ribosome biogenesis, respectively. As seen in the figure, when the cells converge to a quiescent state under Nitrogen depletion, these two clusters exhibit quite different behaviors.

In Figure 7E we see another intriguing behavior of two clusters, $c15$ and $c17$, under steady-state growth in different temperatures. From the GO annotations we find that $c15$, which includes 12 repressed genes, is enriched for "tRNA aminoacylation", while $c17$ which includes 7 repressed genes is enriched with cell cycle related annotations. Figure 7F demonstrates that the distinction between these two clusters is not redundant as they display different behaviors under, e.g., hyper–osmotic shock.

As two complementary validation schemes we used the regulator-promoter region interactions reported in (*15*)[3] and the DNA-binding sequence motifs provided in (*16*).[4] In most of our clusters we found enrichment of regulatory interactions and/or known sequence motif in the corresponding upstream sequences ($Pval < 0.05$, Bonferroni corrected). For example, $c5,\ c19,$ and $c17$ were enriched for YAP1, HSF1, and MBP1, respectively. As YAP1 is known to be involved with response to oxidative stress, HSF1 with response to heat, and MBP1 with cell cycle regulation, these enrichments are clearly in consistent with the GO enrichments for the same clusters. $c18$ and $c2$ (Figs. 4C,D) were enriched with FHL1 which is required for rRNA processing, and $c18$ was further enriched with RAP1 - known to be involved in regulating ribosomal proteins, and with four other regulators (GAT3, YAP5, PDR1, and RGM1),

[3]In these data, every gene is "annotated" with 106 "P-value" scores that determine the probability of this gene being regulated by each of 106 yeast transcriptional regulators. By considering only interactions with a "P-value" lower than 0.005 we constructed out of these data an annotation matrix with 868 (gene) rows, 106 (regulator) columns and a total 1307 predicted interactions.

[4]Here, again, one can construct an annotation matrix where $A(i,j) = 1$ if and only if the $1,000$ base-pair promoter sequence of the i-th gene includes the j-th motif. After considering only the 100 most frequent motifs we ended up with an annotation matrix, with 868 (gene) rows, 100 (motif) columns and $19,517$ predicted interactions.

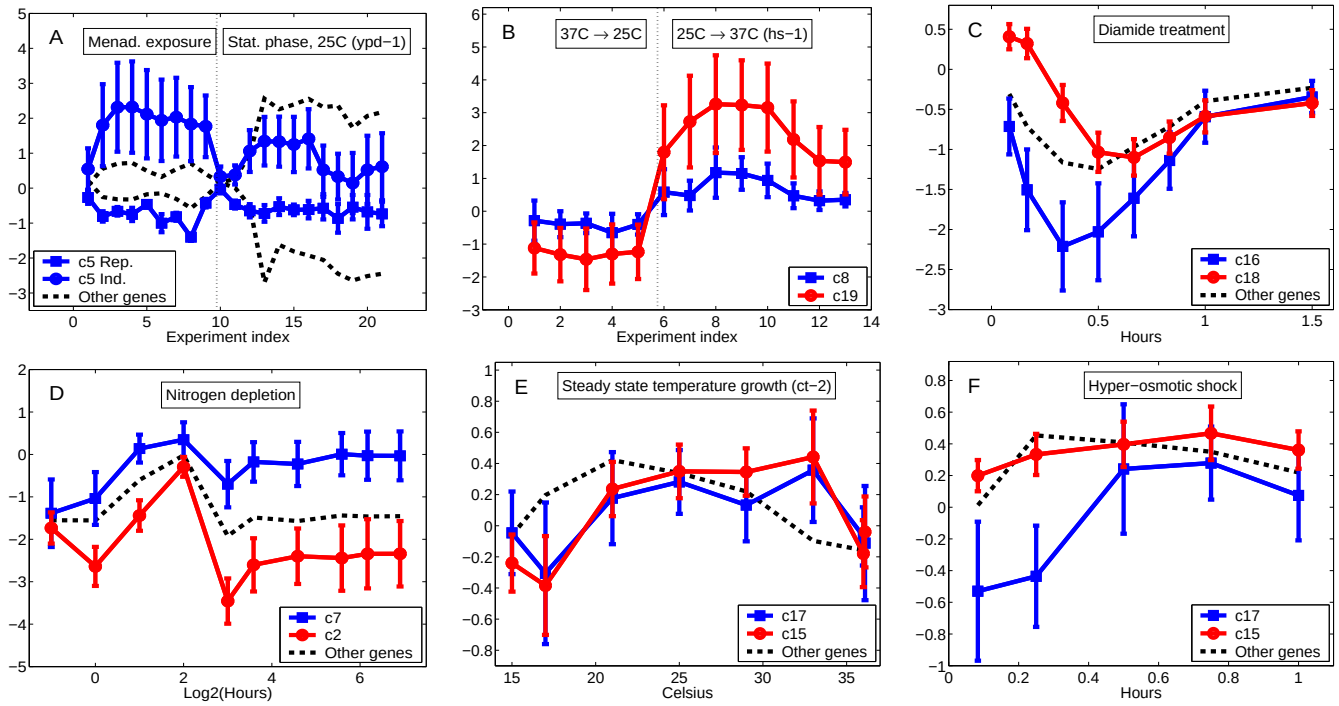

Figure 7: Examples of the average behavior of some of the clusters obtained with $N_c = 20$. Error-bars indicate standard deviation. The vertical axis measures the $\log_2$ of expression ratio. The dashed ("Other genes") curve displays the average behavior of the Repressed genes, excluding those in the clusters that are mentioned in the figure. In panel A the upper dashed curve corresponds to the average behavior of the Induced genes, excluding those in $c5$. **(A)** $c5$ in Menadione exposure and stationary phase. **(B)** $c8$ and $c19$ in different temperatures shifts. **(C)** $c16$ and $c18$ in Diamide treatment. **(D)** $c7$ and $c2$ in Nitrogen depletion. **(E)** $c17$ and $c15$ in steady-state growth. **(F)** $c17$ and $c15$ in hyper–osmotic shock.

suggesting similar, yet not identical regulatory programs for these two functionally related clusters. $c16$ was enriched for ABF1 and both $c7$ and $c16$ were enriched with several motifs which are known to be related to rRNA processing and synthesis, consistently with the GO annotations enriched for these clusters.

In a separate text file we provide complete details of the specific partition obtained by Iclust for $N_c = 20$ clusters. Similar detailed results for $N_c = 15, 10, 5$ (including the analogous tables of Table 10) are available at request and will be posted online in the corresponding web site.

4.7 A cluster enriched with uncharacterized genes

In the statistical validation of our clusters (Section 4.5) we removed from the analysis uncharacterized genes since we were mainly interested to gain some specific insights regarding the function of our clusters. Nonetheless, the distribution of the uncharacterized genes among our clusters yields an intriguing result. One might have suspected that almost every process in the cell has a few components that have not been identified, and hence that as these processes are regulated there would be a handful of unknown genes that are regulated in concert with many genes of known function. For at least one of our clusters, our results reveal a different picture.

Notice, that given the fraction of uncharacterized genes in a cluster and the corresponding fraction at the entire population, one can use the hyper–geometric distribution to calculate a P-value for this event (see Section 3). Applying this to our partition into $N_c = 20$ clusters we find that $c7$ is significantly enriched with genes that are uncharacterized in the GO_{BP} and GO_{MF} ontologies.

Specifically, out of the 123 genes in $c7$, 72 has an unknown molecular function. This level of concentration has a (P-value) probability of $\sim 10^{-8}$ to have arisen by chance. Moreover, if we consider only the repressed genes in the ESR module (since $c7$ consists mainly of such genes), we see that 69 out of the 114 repressed genes in $c7$ are uncharacterized in the GO_{MF} ontology, which has a P-value of $\sim 10^{-15}$.

Closer examination of the GO_{BP} *characterized* genes in the same cluster reveals several enriched annotations (see Table 10) related to ribosome biogenesis and ribosomal RNA processing, suggesting that most of the previously unannotated genes in this cluster are involved in these processes as well. Nonetheless, the extremely high concentration of uncharacterized genes in this cluster suggests that these genes are involved with biological processes which are harder to detect and characterize with the current technologies.

Finally, it is also worthwhile to point out that the cluster $c7$ is extremely preserved when one tries to find partitions with a smaller number of clusters, as demonstrated in Figure 5. In fact, all the "parent" clusters of this $c7$ cluster (for $N_c = 5, 10, 15$) were similarly enriched for GO_{BP} and GO_{MF} uncharacterized genes.

5 Second application: The SP500 data

5.1 Description of the data

In our second application we consider a very different data set, the companies in the Standard and Poor's 500 list. We used the May 2004 listing of the 500 companies, available at *http://www.standardandpoors.com*.

For these companies we downloaded the day–to–day fractional changes in stock price during the trading days between December 2, 2002, and December 31, 2003, (a total of 273 trading days). These data are available at *http://wrds.wharton.upenn.edu* and are presented in Figure 8.[5]

5.2 Mutual information estimation results

From these data we estimated all the $\sim 125,000$ mutual information relations, as described in (*1*), ending up with a matrix of 500×500 information relations which defined the input to our clustering procedure. For convenience, we provide here some statistics of the estimated mutual information values. For a complete description, including different verification schemes that support the reliability of our estimates, the reader is referred to (*1*).

The average estimated mutual information was $0.10\ bits$ with a variance of $0.0054\ bits$. The maximal estimated mutual information was $0.97\ bits$. All the pairwise mutual information relations are presented in Figure 9, where the companies are sorted according to the clustering partition into $N_c = 20$ clusters that we analyze in detail (see below). The self-information relations were set to $I(i;i) = \log_2(5)$ which corresponds to the maximal possible information under a quantization into 5 bins (*1*).

5.3 Implementation details and quality-complexity trade-off curves

Given the pairwise mutual information matrix we applied the Iclust algorithm described in Section 2. As in the first application, we explored the trade-off between $\langle s \rangle$ and $I(C;\mathrm{i})$ for different numbers of clusters: $N_c = 5, 10, 15, 20$ and for different values of the trade-off parameter, T. Specifically, we found that $\frac{1}{T} = \{15, 20, 25, 30, 35\}$ were typically sufficient to obtain a relatively clear saturation of $\langle s \rangle$, hence we present the results for these T values. For

[5] Notice, that we identifi ed the different companies by their ticker symbols as reported in *http://www.standardandpoors.com*. However, these symbols are not unique, and as a result the database at *http://wrds.wharton.upenn.edu* returned slightly more than 500 companies, for which only for 501 the data was available for the entire 2003 year, hence these 501 are the companies we consider in our analysis.

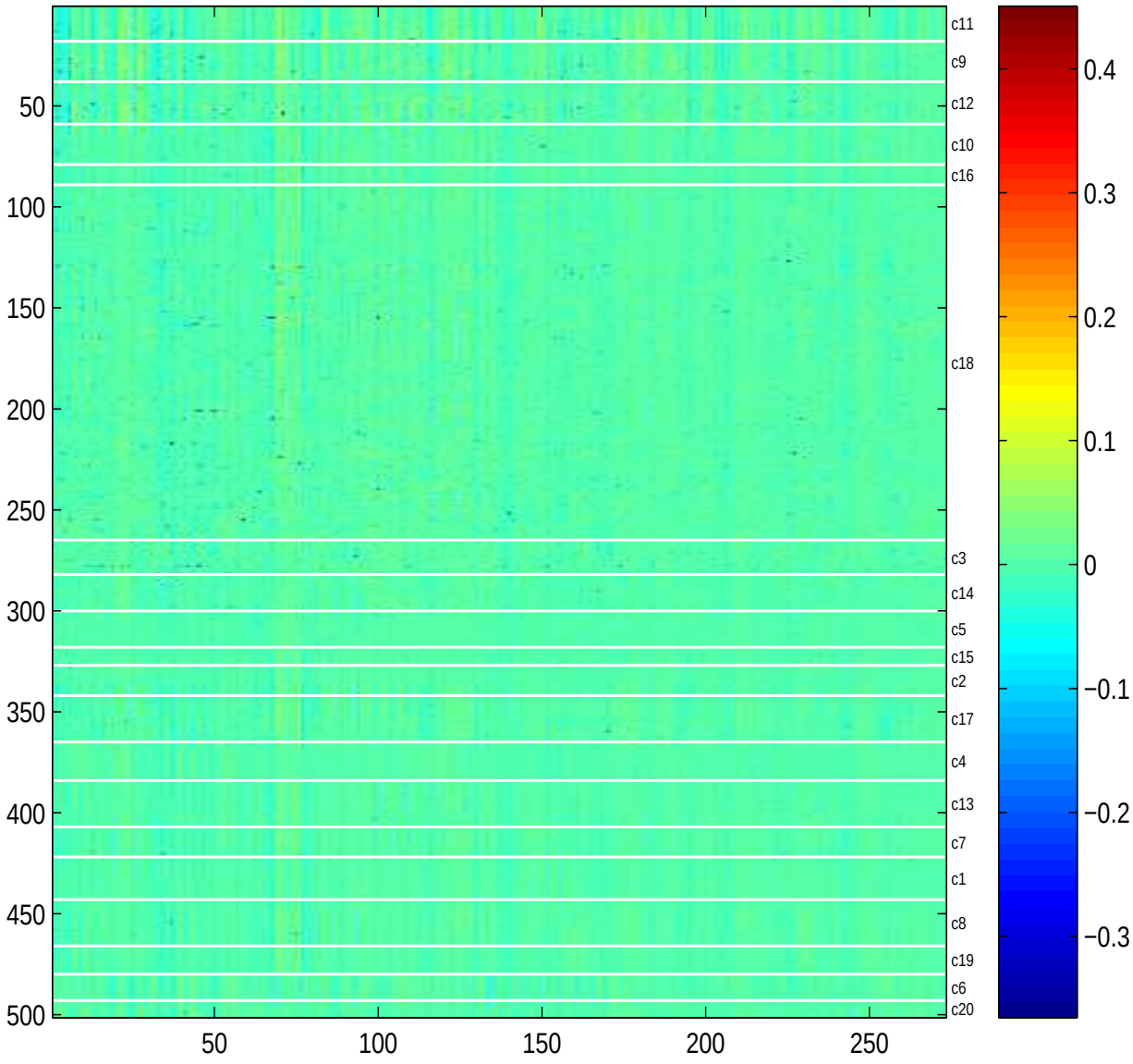

Figure 8: Fractional changes in stock price of the Standard and Poor's companies we considered during the 273 trading days of December 2002 – December 2003. The companies are sorted according to the clustering partition into 20 clusters that we analyze in detail later on. Inside each cluster, companies are sorted according to the average mutual information relation with other cluster members.

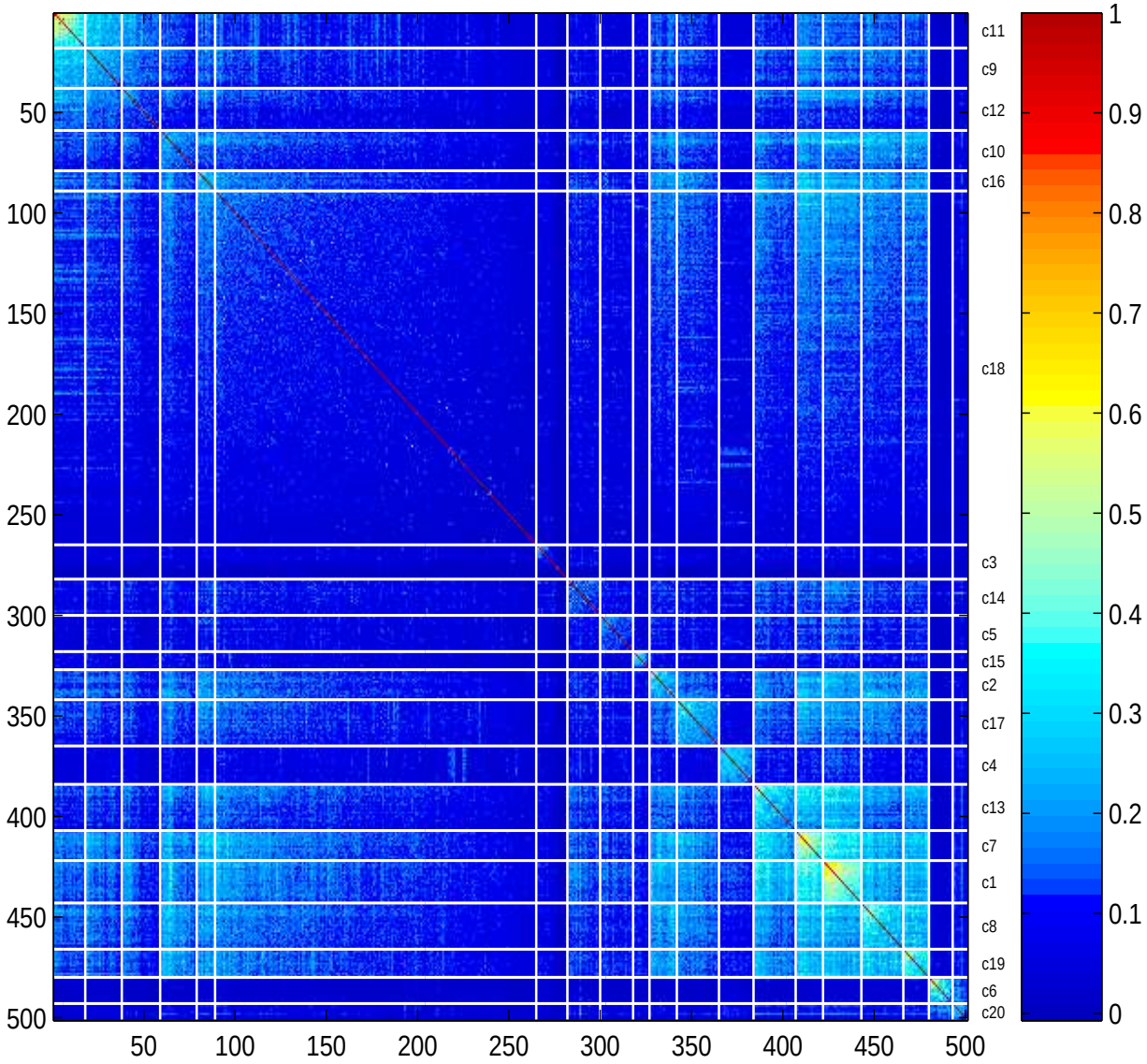

Figure 9: Pairwise mutual information relations for the SP500 data. The companies are sorted according to the clustering partition into 20 clusters that we analyze in detail later on. Inside each cluster, companies are sorted according to the average mutual information relation with other cluster members.

each $\{N_c\ , T\}$ pair we performed 10 different random initializations ending up with 10 (possibly) different local maxima of $\mathcal{F}$, from which we chose the best one. The resulting trade-off curves are presented in the middle panel of Figure 4.

As before, as T is lowered, $\langle s \rangle$ increases but so does $I(C;\mathrm{i})$. In addition, the solutions become more deterministic. For example, for $N_c = 20$ and $\frac{1}{T} = 25$, only $\sim \ 36\%$ of the companies have nearly deterministic assignment ($P(C|i) > 0.9$ for a particular C). On the other hand, for $\frac{1}{T} = 35$, all the assignments are nearly deterministic ($P(C|i) > 0.9$).

For brevity, we will further focus our analysis on solutions for which the saturation of $\langle s \rangle$ is relatively clear, i.e., on the four solutions with $N_c = \{5, 10, 15, 20\}$ and $\frac{1}{T} = 35$. In all these partitions almost all of the companies had a nearly deterministic assignment ($P(C|i) > 0.9$ for a particular C). Hence, for further simplicity, in the rest of the analysis we treat these solutions as "hard" partitions where every company is assigned solely to its most probable cluster. In the next section we explore the possible hierarchical relations between these four solutions. In Section 5.6 we analyze in detail the specific solution with $\{N_c = 20,\ \frac{1}{T} = 35\}$ that obtained the highest $\langle s \rangle$ value.

5.4 Comparing solutions at different numbers of clusters

We examine directly how well our independent solutions form a hierarchical structure. Accordingly, we apply exactly the same scheme as described in Section 4.4 to the four solutions we obtained independently for $N_c = \{5, 10, 15, 20\}$ with $\frac{1}{T} = 35$. The results are presented in Figure 10. Again, the independent solutions form only an "approximated" hierarchy. Nonetheless, this approximation seems more suitable in this case, as demonstrated, e.g., by the larger percentage of nearly perfect inclusion relations (solid red bold edges in the figure). It should be noted that indeed the standard classification of these companies is hierarchical in nature (see Section 5.5.1).

Again, it is worthwhile to point out that some of the clusters are better preserved than others across the different levels. For example, the "Semiconductors Equipment" cluster, $c11$, is clearly identified at all the independent solutions.

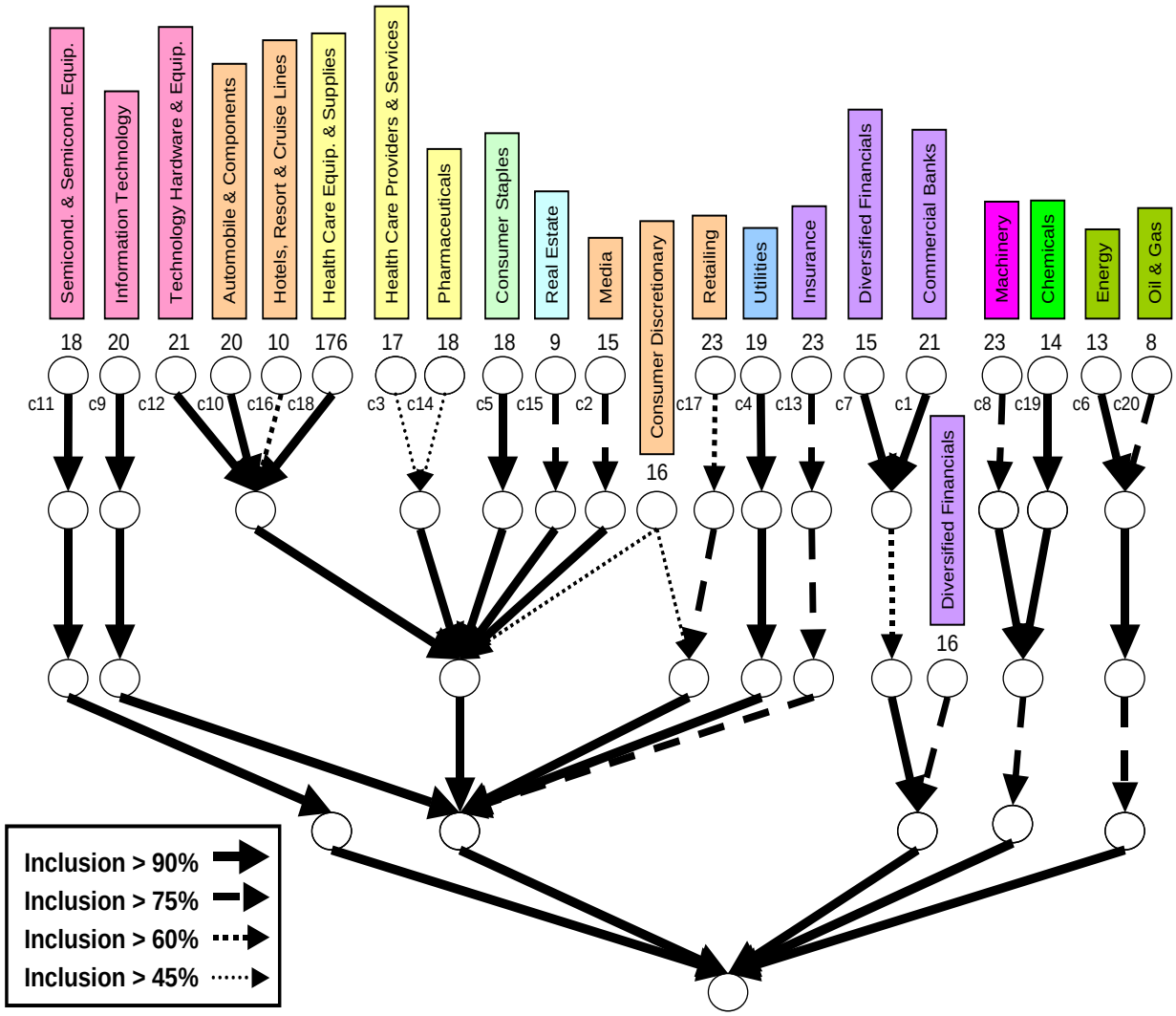

Figure 10: Relations between the optimal solutions with $N_c = \{5, 10, 15, 20\}$ at $\frac{1}{T} = 35$ for the SP500 data. At the upper level, $N_c = 20$ clusters, and the clusters are sorted as in Figure 8 and Figure 9. The numbers above every cluster indicate the number of companies in this cluster. The title of each cluster correspond to the most enriched annotation in the cluster, i.e., to the annotation with the smallest P-value in the cluster. Similar color of text boxes indicate that the corresponding annotations belong to the same major sector of economy (see Section 5.5.1). Notice, that most clusters were enriched with more than one annotation, hence the short titles might be too concise in some cases (see Section 5.6 for a detailed description of every cluster).

5.5 Coherence results

5.5.1 Constructing the annotation matrices

We used the Global Industry Classification Standard (GICS) methodology which classifies companies at four different levels: sector, industry group, industry, and sub-industry (see *http://www.standardandpoors.com*).

These four levels are arranged in a well defined "tree-like" hierarchy. The bottom ("sector") level consists of 10 different annotations: "Consumer Discretionary", "Consumer Staples", "Energy", "Financials", "Health Care", "Industrials", "Information Technology", "Materials", "Telecommunication Services", and " Utilities". The next ("industry group") level consists of 24 distinct annotations. The next ("industry") level consists of 59 distinct annotations. The last ("sub-industry") level consists of 114 distinct annotations. Thus, altogether there are 207 different annotations where every company is assigned with exactly 4 annotations, one at every level of the hierarchy.

As in the first application, while estimating clusters' coherence we removed annotations that were assigned with less than two companies in our data, ending up with a total of 178 distinct annotations.

5.5.2 Coherence results and comparison to other clustering algorithms

We estimated the statistical coherence of the clusters obtained at the low-temperature end of the trade-off curves where $\frac{1}{T} = 35$. To gain some perspective, we applied similar analysis with the "Cluster 3.0" software (*12*). We experimented with the same 18 basic configurations as in the previous application (K-means variants, again with 100 different initializations), and applied the comparison to all the different numbers of clusters we examined, $N_c = 5, 10, 15, 20$. The results are summarized in Table 11 to Table 14 and in Figure 11.

In all cases, Iclust was superior to the average performance of the K–means and the hierarchical "Cluster 3.0" variants. In fact, except for the "K–medians" configurations, none of the other algorithms came even close to the Iclust performance. Averaging over all four N_c values, Iclust obtains an average coherence of $\sim 90\%$ while the average K–means coherence is $\sim 79\%$ and the average Hierarchical coherence is only $\sim 19\%$.

It is interesting to point out that although the annotations for these data are arranged in a relatively simple and clear hierarchical structure, the performance of the hierarchical algorithms are still relatively poor, perhaps due to the

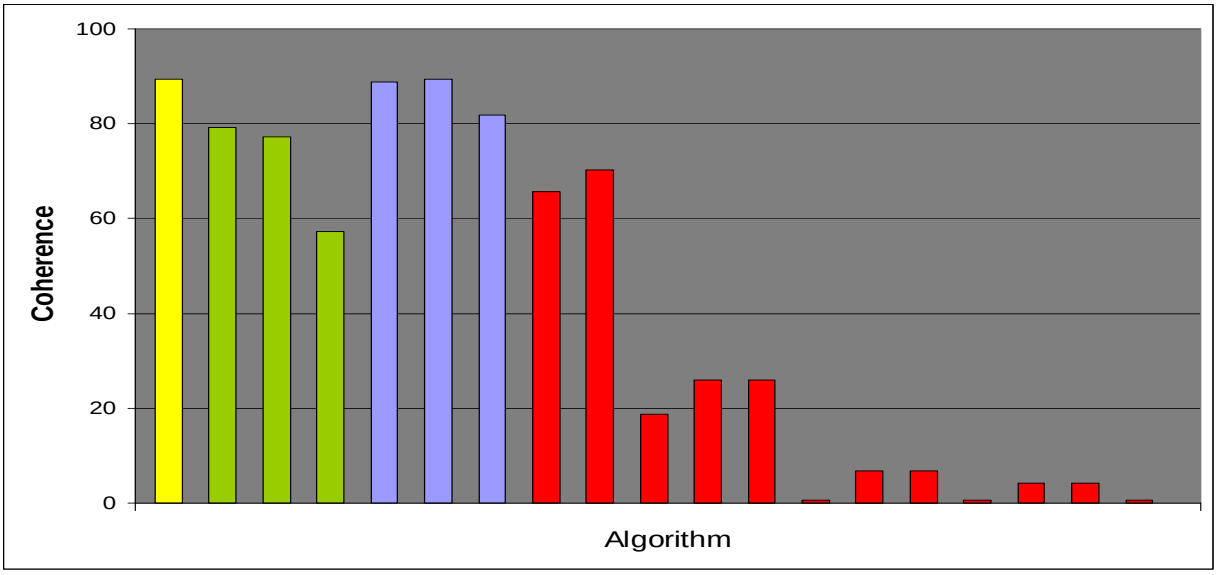

Figure 11: **SP500 data**: Comparison of average coherence results of the Iclust algorithm (yellow) with conventional clustering algorithms (*12*): K–means (green); K–medians (blue); Hierarchical (red). For the hierarchical algorithms, four different variants are tried: complete, average, centroid, and single linkage, respectively from left to right. For every algorithm, three different similarity measures are applied: Pearson correlation (left); absolute value of Pearson correlation (middle); Euclidean distance (right). In all cases, the results are averaged over all the different numbers of clusters that we tried: $N_c = 5, 10, 15, 20$.

greedy nature of these optimization routines which typically yield sub-optimal solutions.

5.6 Detailed results for the $N_c = 20$ clusters partition

In Table 15 we present all the enriched annotations for the Iclust partition with $N_c = 20$ clusters and $\frac{1}{T} = 35$. Several specific results are noted in the following.

First, 8 out of the 20 clusters are perfectly (100%) coherent. For example, $c11$ consists of 18 companies which are all "Information Technology" companies, mainly sub–classified as "Semiconductors & Semiconductor Equipment" companies like Intel and Texas Instruments; $c17$ consists of mainly different types of stores: "Department Stores" like Sears, "General Merchandise Stores" like Target, "Speciality Stores" like Staples, and so on;

Perhaps more interesting is the relatively subtle distinction between $c7$ and $c1$ (both perfectly coherent). The former includes mainly companies which are classified as "Investment Banking & Brokerage" (e.g., Merril Lynch) or

Table 11: Coherence results for the SP500 data with respect to the GICS companies' annotations with $N_c = 20$ clusters. [a]Clustering algorithm. In the "$\langle$ $K-$means $\rangle$" row we present the average results of all the 6 $K-$means variants. For each of these variants we performed 100 runs from which the best solution is chosen. In the "$\langle$ Hier. $\rangle$" row we present the average results of all the 12 Hierarchical clustering variants. [b]Correlation measure used by the algorithm. "PC" stands for the (centered) Pearson Correlation. "|PC|" is the absolute value of this correlation. "Euclidean" stands for the Euclidean distance. [c]Number of clusters with a positive coherence. [d]Average coherence of all 20 clusters.

$N_c = 20$ **Algorithm** [a]	**Similarity** [b]	N_c^{pos} [c]	$\langle Coh \rangle$ [d]
Iclust	mutual information	20	86
$K-$means	PC	19	74
$K-$means	|PC|	17	69
$K-$means	Euclidean	15	58
$K-$medians	PC	20	85
$K-$medians	|PC|	20	88
$K-$medians	Euclidean	20	81
$\langle$ $K-$means $\rangle$		18.5	75.8
Hier - Comp. linkage	PC	16	70
Hier - Comp. linkage	|PC|	16	70
Hier - Comp. linkage	Euclidean	4	12
Hier - Avg. linkage	PC	7	32
Hier - Avg. linkage	|PC|	7	32
Hier - Avg. linkage	Euclidean	0	0
Hier - Centr. linkage	PC	2	10
Hier - Centr. linkage	|PC|	2	10
Hier - Centr. linkage	Euclidean	0	0
Hier - Sing. linkage	PC	2	10
Hier - Sing. linkage	|PC|	2	10
Hier - Sing. linkage	Euclidean	0	0
$\langle$ Hierarchical $\rangle$		4.8	21.3

Table 12: Coherence results for the SP500 data with respect to the GICS companies' annotations with $N_c = 15$ clusters. The column and row definitions are as in Table 11.

$N_c = 15$ **Algorithm** [a]	**Similarity** [b]	N_c^{pos} [c]	$\langle Coh \rangle$ [d]
Iclust	mutual information	15	93
K–means	PC	12	69
K–means	|PC|	13	68
K–means	Euclidean	11	54
K–medians	PC	15	90
K–medians	|PC|	15	88
K–medians	Euclidean	15	85
$\langle$ K–means $\rangle$		13.5	75.7
Hier - Comp. linkage	PC	11	63
Hier - Comp. linkage	|PC|	11	63
Hier - Comp. linkage	Euclidean	2	5
Hier - Avg. linkage	PC	6	32
Hier - Avg. linkage	|PC|	6	32
Hier - Avg. linkage	Euclidean	0	0
Hier - Centr. linkage	PC	1	7
Hier - Centr. linkage	|PC|	1	7
Hier - Centr. linkage	Euclidean	0	0
Hier - Sing. linkage	PC	1	7
Hier - Sing. linkage	|PC|	1	7
Hier - Sing. linkage	Euclidean	0	0
$\langle$ Hierarchical $\rangle$		3.3	18.6

Table 13: Coherence results for the SP500 data with respect to the GICS companies' annotations with $N_c = 10$ clusters. The column and row definitions are as in Table 11.

$N_c = 10$ **Algorithm** [a]	**Similarity** [b]	N_c^{pos} [c]	$\langle Coh \rangle$ [d]
Iclust	mutual information	10	91
K–means	PC	10	84
K–means	|PC|	10	85
K–means	Euclidean	8	63
K–medians	PC	10	90
K–medians	|PC|	10	90
K–medians	Euclidean	10	77
$\langle$ K–means $\rangle$		9.7	81.5
Hier - Comp. linkage	PC	8	64
Hier - Comp. linkage	|PC|	8	64
Hier - Comp. linkage	Euclidean	4	22
Hier - Avg. linkage	PC	2	20
Hier - Avg. linkage	|PC|	2	20
Hier - Avg. linkage	Euclidean	0	0
Hier - Centr. linkage	PC	1	10
Hier - Centr. linkage	|PC|	1	10
Hier - Centr. linkage	Euclidean	0	0
Hier - Sing. linkage	PC	0	0
Hier - Sing. linkage	|PC|	0	0
Hier - Sing. linkage	Euclidean	0	0
$\langle$ Hierarchical $\rangle$		2.2	17.5

Table 14: Coherence results for the SP500 data with respect to the GICS companies' annotations with $N_c = 5$ clusters. The column and row definitions are as in Table 11.

$N_c = 5$ **Algorithm** [a]	**Similarity** [b]	N_c^{pos} [c]	$\langle Coh \rangle$ [d]
Iclust	mutual information	5	88
K–means	PC	5	90
K–means	|PC|	5	87
K–means	Euclidean	4	54
K–medians	PC	5	90
K–medians	|PC|	5	92
K–medians	Euclidean	5	84
$\langle$ K–means $\rangle$		4.8	82.8
Hier - Comp. linkage	PC	4	66
Hier - Comp. linkage	|PC|	5	84
Hier - Comp. linkage	Euclidean	3	36
Hier - Avg. linkage	PC	1	20
Hier - Avg. linkage	|PC|	1	20
Hier - Avg. linkage	Euclidean	0	0
Hier - Centr. linkage	PC	0	0
Hier - Centr. linkage	|PC|	0	0
Hier - Centr. linkage	Euclidean	0	0
Hier - Sing. linkage	PC	0	0
Hier - Sing. linkage	|PC|	0	0
Hier - Sing. linkage	Euclidean	0	0
$\langle$ Hierarchical $\rangle$		1.2	18.8

"Asset Management & Custody Banks", while the latter corresponds to "Commercial (Regional) Banks" like PNC. Indeed, in Figure 10 we see that these two clusters nicely merge with each other at the independent solution found for $N_c = 15$ clusters. A similar relativly subtle distinction is also captured between $c6$ and $c20$ (again, both are perfectly coherent), where both clusters correspond to different sub–classifications of the "Oil & Gas" category. As for the previous pair, these two clusters also merge for $N_c = 15$.

Even in clusters with non perfect coherence we typically see a clear reasoning behind the automatically recovered structure. For example, $c16$ is basically enriched only for 3 "Hotels Resorts & Cruise Lines" companies with a coherence level of only 30%. Nonetheless, it further contains two banks (MBNA and Capital One Financial) which specialise in credit card issuing and therefore consumer spending, a company (CINTAS) which is a builder of corporate identity, and another company (Paychex) which handles payroll and human resource services for employees. In addition, the Walt Disney Co. is also in this cluster, presumably due to its parks and resorts division.

In a separate text file we provide complete details of this partition. Similar detailed results for $N_c = 15, 10, 5$ (including the analogous tables of Table 15) are available upon request and will be posted online in the corresponding web site.

Table 15: Enriched GICS annotations in the Iclust solution with $N_c = 20$ clusters for the SP500 data. Clusters are ordered as in Figure 8, Figure 9, and Figure 10. Only annotations with a P-value below 0.05 (Bonferroni corrected) are presented. [a]Cluster index. [c]Cluster size. [c]Cluster coherence (in percentage). [d]Enriched annotations. In parentheses: $(x/K, p)$ stands for the number of companies in the cluster to which this annotations is assigned, the number of companies in the entire data to which this annotations is assigned, and the Bonferroni corrected P-value, respectively.

C index[a]	**C size**[b]	**Coh.**[c]	**Enriched annot.**[d]
c11	18	100	4530 Semiconductors& Semiconductor Equipment (16/19,0.000000) 453010 Semiconductor& Semiconductor Equipment (16/19,0.000000) 45301020 Semiconductors (12/15,0.000000) 45 Information Technology (18/81,0.000000) 45301010 Semiconductor Equipment (4/4,0.000013)
c9	20	95	45 Information Technology (19/81,0.000000) 4520 Technology Hardware& Equipment (10/35,0.000002) 451030 Software (6/15,0.000155) 452030 Electronic Equipment& Instruments (5/10,0.000276) 4510 Software& Services (7/27,0.000617) 452010 Communications Equipment (5/14,0.001970) 45201020 Communications Equipment (5/14,0.001970) 45103010 Application Software (4/8,0.002368) 45203020 Electronic Manufacturing Services (3/4,0.004177)
c12	21	95	4520 Technology Hardware& Equipment (13/35,0.000000) 45 Information Technology (17/81,0.000000) 45202010 Computer Hardware (5/7,0.000035) 452010 Communications Equipment (6/14,0.000132) 45201020 Communications Equipment (6/14,0.000132) 452020 Computers& Peripherals (5/10,0.000383) 501020 Wireless Telecommunication Services (2/2,0.040138) 50102010 Wireless Telecommunication Services (2/2,0.040138)
c10	20	65	2510 Automobiles& Components (6/9,0.000005) 251010 Auto Components (4/6,0.000863) 201010 Aerospace& Defense (4/9,0.006685) 20101010 Aerospace& Defense (4/9,0.006685) 25101010 Auto Parts& Equipment (3/4,0.006730) 2010 Capital Goods (7/37,0.008990) 25102010 Automobile Manufacturers (2/2,0.046560)
c16	10	30	25301020 Hotels Resorts& Cruise Lines (3/4,0.000546) 2530 Hotels Restaurants& Leisure (3/11,0.020860) 253010 Hotels Restaurants& Leisure (3/11,0.020860)
c18	176	19	351010 Health Care Equipment& Supplies (13/13,0.000133) 35101010 Health Care Equipment (11/11,0.001186) 2020 Commercial Services& Supplies (11/12,0.009850) 202010 Commercial Services& Supplies (11/12,0.009850) 2030 Transportation (9/9,0.010371)
c3	17	83	351020 Health Care Providers& Services (9/16,0.000000) 3510 Health Care Equipment& Services (9/29,0.000000) 35 Health Care (10/47,0.000000) 35102030 Managed Health Care (4/5,0.000015) 35102015 Health Care Services (2/4,0.045569) 35102020 Health Care Facilities (2/4,0.045569)

C index[a]	C size[b]	Coh.[c]	Enriched annot.[d]
c14	18	94	352020 Pharmaceuticals (10/13,0.000000) 35202010 Pharmaceuticals (10/13,0.000000) 3520 Pharmaceuticals& Biotechnology (10/18,0.000000) 35 Health Care (12/47,0.000000) 501010 Diversifi ed Telecommunication Services (5/9,0.000066) 50101020 Integrated Telecommunication Services (5/9,0.000066) 50 Telecommunication Services (5/11,0.000232) 5010 Telecommunication Services (5/11,0.000232)
c5	18	94	30 Consumer Staples (17/35,0.000000) 3020 Food Beverage& Tobacco (12/19,0.000000) 302020 Food Products (9/10,0.000000) 30202030 Packaged Foods& Meats (8/9,0.000000) 3030 Household& Personal Products (4/6,0.000341) 303010 Household Products (3/4,0.003000) 30301010 Household Products (3/4,0.003000) 302010 Beverages (3/6,0.014308)
c15	9	83	4040 Real Estate (5/6,0.000000) 404010 Real Estate (5/6,0.000000) 40401010 Real Estate Investment Trusts (5/6,0.000000) 40 Financials (5/80,0.004491)
c2	15	93	2540 Media (10/14,0.000000) 254010 Media (10/14,0.000000) 25401040 Publishing (7/7,0.000000) 25 Consumer Discretionary (14/83,0.000000) 25401020 Broadcasting& Cable TV (2/3,0.031371) 252010 Household Durables (3/11,0.040516)
c17	23	100	2550 Retailing (19/30,0.000000) 25 Consumer Discretionary (21/83,0.000000) 255030 Multiline Retail (9/11,0.000000) 255040 Specialty Retail (10/17,0.000000) 25503010 Department Stores (5/7,0.000050) 25503020 General Merchandise Stores (4/4,0.000065) 25504010 Apparel Retail (3/3,0.001574) 25504040 Specialty Stores (4/8,0.004005) 30101040 HyperMarkets& Super Centers (2/2,0.036344)
c4	19	100	55 Utilities (19/36,0.000000) 5510 Utilities (19/36,0.000000) 551010 Electric Utilities (14/22,0.000000) 55101010 Electric Utilities (14/22,0.000000) 551020 Gas Utilities (3/6,0.007516) 55102010 Gas Utilities (3/6,0.007516)
c13	23	100	4030 Insurance (19/21,0.000000) 403010 Insurance (19/21,0.000000) 40 Financials (23/80,0.000000) 40301040 Property& Casualty Insurance (9/9,0.000000) 40301020 Life& Health Insurance (6/7,0.000000) 40301030 Multi-line Insurance (3/3,0.001203) 401020 Thrifts& Mortgage Finance (3/6,0.021900) 40102010 Thrifts& Mortgage Finance (3/6,0.021900)
c7	15	100	4020 Diversifi ed Financials (15/24,0.000000) 402030 Capital Markets (13/16,0.000000) 40 Financials (15/80,0.000000) 40203020 Investment Banking& Brokerage (6/7,0.000000) 40203010 Asset Management& Custody Banks (6/8,0.000000)
c1	21	100	401010 Commercial Banks (21/23,0.000000) 4010 Banks (21/29,0.000000) 40101015 Regional Banks (16/17,0.000000) 40 Financials (21/80,0.000000) 40101010 Diversifi ed Banks (5/6,0.000003)

C index[a]	C size[b]	Coh.[c]	Enriched annot.[d]
c8	23	83	201060 Machinery (12/14,0.000000) 2010 Capital Goods (16/37,0.000000) 20 Industrials (16/58,0.000000) 20106020 Industrial Machinery (7/9,0.000000) 20106010 Construction& Farm Machinery& Heavy Trucks (5/5,0.000004) 151050 Paper& Forest Products (3/5,0.023469)
c19	14	93	151010 Chemicals (11/14,0.000000) 15 Materials (13/33,0.000000) 1510 Materials (13/33,0.000000) 15101020 Diversifi ed Chemicals (5/6,0.000001) 15101050 Specialty Chemicals (4/5,0.000026) 15101040 Industrial Gases (2/2,0.009228) 15103020 Paper Packaging (2/3,0.027226)
c6	13	100	10 Energy (13/23,0.000000) 1010 Energy (13/23,0.000000) 101010 Energy Equipment& Services (7/7,0.000000) 10102020 Oil& Gas Exploration& Production (6/7,0.000000) 10101020 Oil& Gas Equipment& Services (4/4,0.000002) 101020 Oil& Gas (6/16,0.000005) 10101010 Oil& Gas Drilling (3/3,0.000105)
c20	8	100	101020 Oil& Gas (8/16,0.000000) 10 Energy (8/23,0.000000) 1010 Energy (8/23,0.000000) 10102010 Integrated Oil& Gas (5/6,0.000000) 10102030 Oil& Gas Refi ning& Marketing& Transportation (2/3,0.004224)

6 Third application: The EachMovie data

6.1 Description of the data

In our third test case we consider the $EachMovie$ dataset, movie ratings provided by more than $70,000$ viewers (see *http://www.research.digital.com/SRC/eachmovie/*.) These data are inherently quantized as only six discrete possible ratings were used. Indeed, many real life clustering problems involve such "categorical data". In these cases the issue of what similarity measure to use seems even more obscure, especially if the descriptive attributes are not naturally ordered. Thus, our approach perfectly suits these cases and in fact can benefit from this data format as no quantization scheme need be applied while estimating the information relations. We represented each movie by its ratings from different viewers and focused on the 500 movies that got the maximal number of votes. These data are presented in Figure 12 and are available at *http://www.research.digital.com/SRC/eachmovie/*.

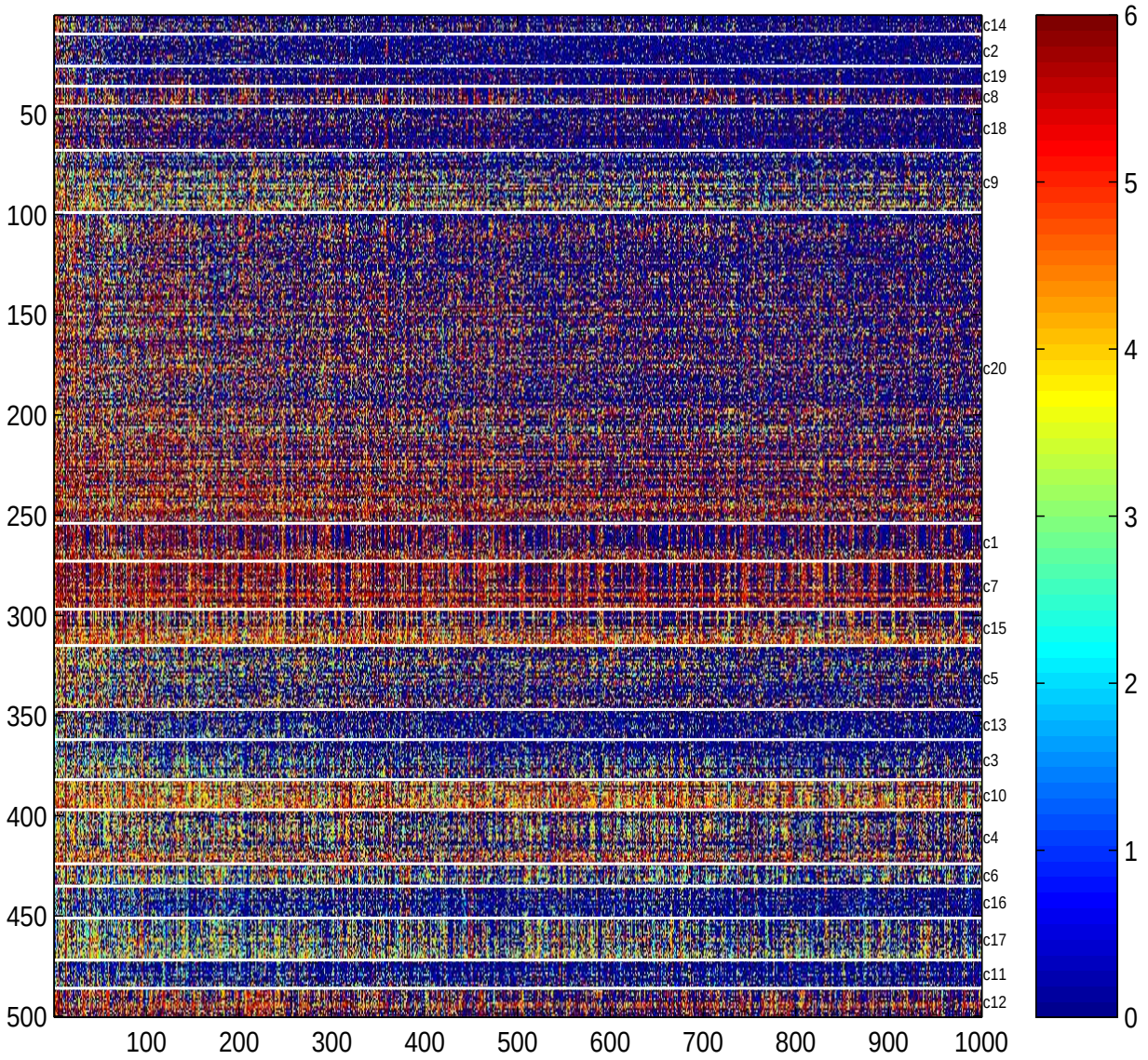

Figure 12: Discrete movie ratings for the 500 movies with the maximal number of votes in the EachMovie data. The ratings are presented only for the 1000 viewers who rated the maximal number of movies. Zeros represent "missing values" (i.e., no vote). The movies are sorted according to the clustering partition into 20 clusters that we analyze in detail later on. Inside each cluster, movies are sorted according to the average mutual information relation with other cluster members.

6.2 Mutual information estimation results

From these data we estimated all the $\sim 125,000$ mutual information relations, as described in (*1*), ending up with a matrix of 500×500 information relations which defined the input to our clustering procedure. Notice, that while estimating the mutual information for a pair of movies, only viewers who voted for both movies were considered.

The average estimated mutual information was $0.052\ bits$ with a variance of $0.0026\ bits$. The maximal estimated mutual information was $0.89\ bits$. All the pairwise mutual information relations are presented in Figure 13, where the movies are sorted according to the clustering partition into $N_c = 20$ clusters that we analyze in detail (see below). The self-information relations were set to $I(i;i) = \log_2(6)$ which corresponds to the maximal possible information under a quantization into 6 bins.

For a complete description of the mutual information estimation procedure, including different verification schemes that support the reliability of our estimates, the reader is referred to (*1*).

6.3 Implementation details and quality-complexity trade-off curves

Given the pairwise mutual information matrix we applied the Iclust algorithm described in Section 2. As in the previous applications, we explored the trade-off between $\langle s \rangle$ and $I(C;\mathrm{i})$ for different numbers of clusters: $N_c = 5, 10, 15, 20$ and for different values of the trade-off parameter, T. Specifically, we found that $\frac{1}{T} = \{20, 25, 30, 35, 40\}$ were typically sufficient to obtain a relatively clear saturation of $\langle s \rangle$, hence we present the results for these T values. For each $\{N_c\ , T\}$ pair we performed 10 different random initializations ending up with 10 (possibly) different local maxima of $\mathcal{F}$, from which we chose the best one. The resulting trade-off curves are presented in the right panel of Figure 4.

As before, as T is lowered, $\langle s \rangle$ and $I(C;\mathrm{i})$ increase and the solutions become more deterministic. For example, for $N_c = 20$ and $\frac{1}{T} = 30$, only $\sim\ 32\%$ of the movies have nearly deterministic assignment, while for $\frac{1}{T} = 40$ almost all the movie assignments are nearly deterministic ($P(C|i) > 0.9$ for a particular C).

For brevity, we will further focus our analysis on solutions for which the saturation of $\langle s \rangle$ is relatively clear, i.e., on the four solutions with $N_c = \{5, 10, 15, 20\}$ and $\frac{1}{T} = 40$. In all these partitions almost all of the movies had a nearly

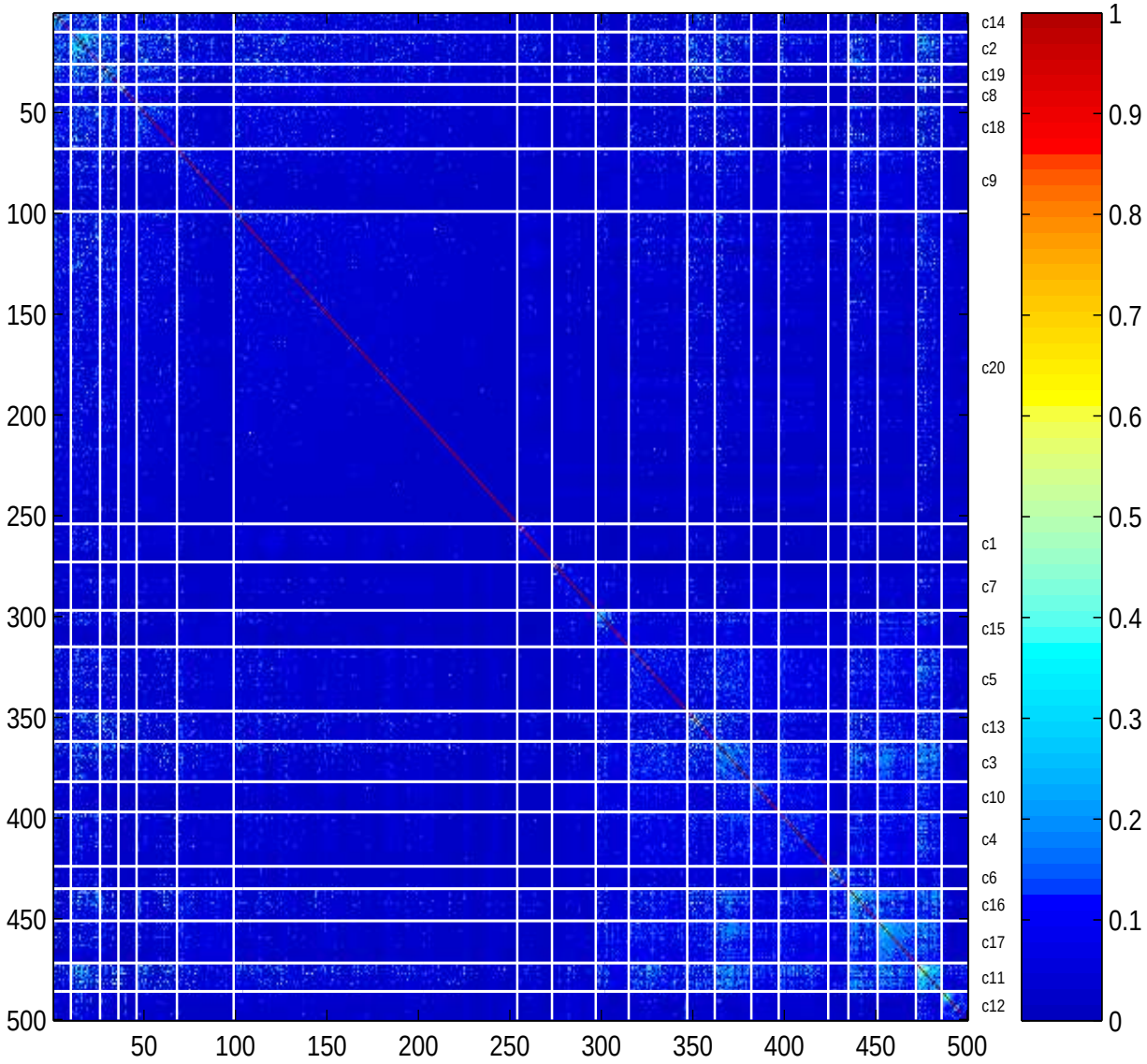

Figure 13: Pairwise mutual information relations for the EachMovie data. The movies are sorted according to the clustering partition into 20 clusters that we analyze in detail later on. Inside each cluster, movies are sorted according to the average mutual information relation with other cluster members.

deterministic assignment ($P(C|i) > 0.9$ for a particular C). Hence, for further simplicity, in the rest of the analysis we treat these solutions as "hard" partitions where every movie is assigned solely to its most probable cluster. In the next section we explore the possible hierarchical relations between these four solutions. In Section 6.6 we analyze in detail the specific solution with $\{N_c = 20,\ \frac{1}{T} = 40\}$ that obtained the highest $\langle s \rangle$ value.

6.4 Comparing solutions at different numbers of clusters

We examine directly how well our independent solutions form a hierarchical structure by applying the same scheme as in Section 4.4. The results are presented in Figure 14. Clearly, the relations between solutions at different numbers of clusters are relatively weak, suggesting that the obtained solutions are not that robust. Only a few clusters are somewhat preserved, like the "Family-Animation-Classic" cluster, $c12$, or the "Action" cluster, $c9$.

6.5 Coherence results

6.5.1 Constructing the annotation matrices

We used the genre labels provided for these data to construct the annotation matrix. Specifically, these labels are: "Action" (110 movies), "Animation" (25 movies), "Art-Foreign" (45 movies), "Classic" (44 movies), "Comedy" (149 movies), "Drama" (160 movies), "Family" (67 movies), "Horror" (33 movies), "Romance" (61 movies), and "Thriller" (90 movies). Almost half of the movies were annotated with more than one genre and the average number of genre annotations per movie was 1.6, with a maximal number of 4 different genres for a single movie.

It is important to notice that these annotations are probably too broad, providing a somewhat simplistic view of the structure in these data. For example, it is quite reasonable that more subtle distinctions like the movie director and/or main actors are reflected in the viewer preferences that were used to cluster the movies. Nonetheless, for practical reasons we used these broad labels as a "first order approximation" for our evaluation.

6.5.2 Coherence results and comparison to other clustering algorithms

We estimated the statistical coherence of the clusters obtained at the low-temperature end of the trade-off curves where $\frac{1}{T} = 40$. As before, to gain some perspective, we applied similar analysis with the "Cluster 3.0" software (*12*). We experimented with the same 18 basic configurations as in the previous applications (K-means variants, again

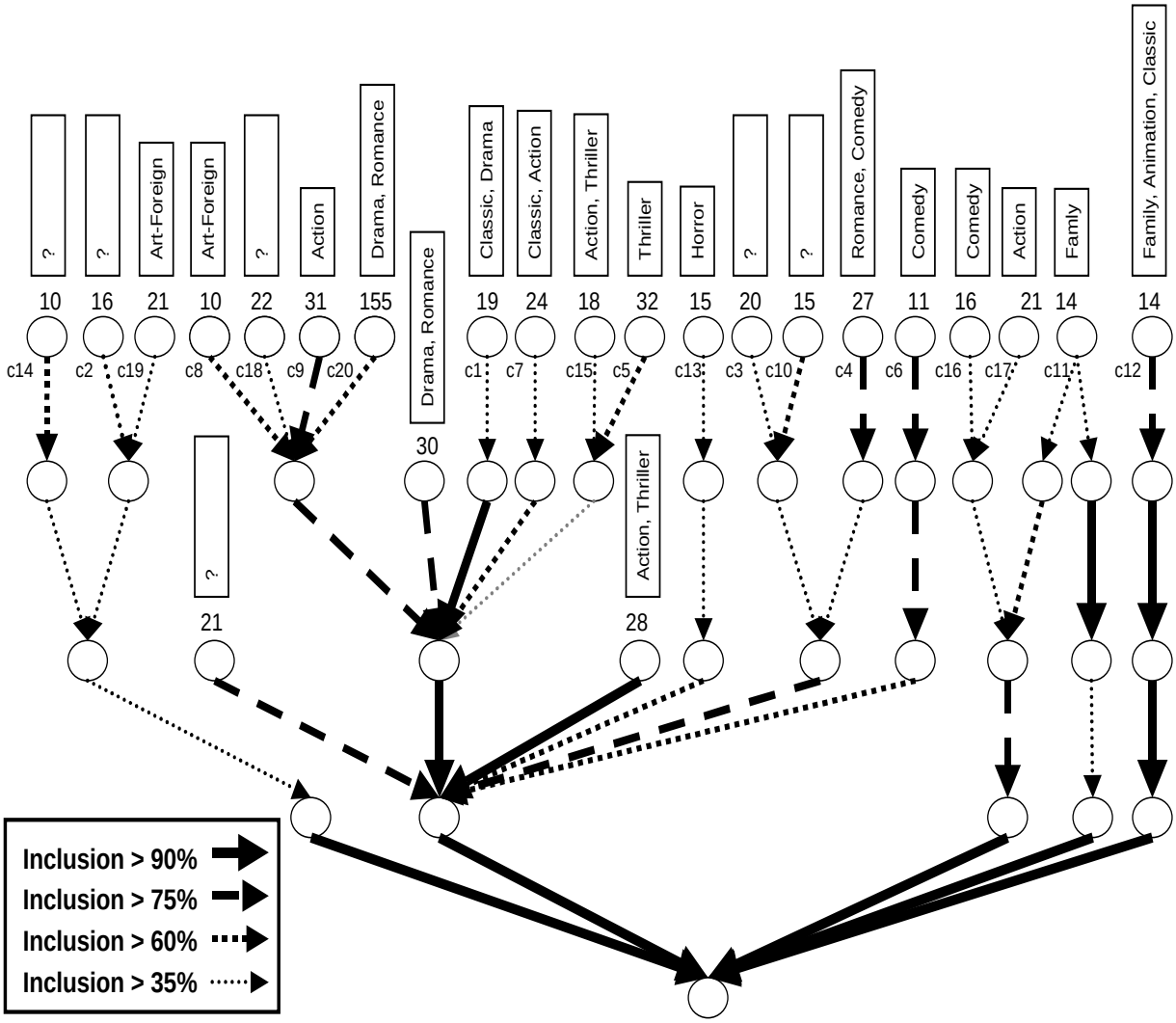

Figure 14: Relations between the optimal solutions with $N_c = \{5, 10, 15, 20\}$ at $\frac{1}{T} = 40$ for the EachMovie data. At the upper level, $N_c = 20$ clusters, and the clusters are sorted as in Figure 12 and Figure 13. The numbers above every cluster indicate the number of movies in this cluster. The title of each cluster correspond to (all) the enriched "genre" annotation in the cluster, i.e., to all the annotation with a (Bonferroni corrected) P-value below 0.05. See Section 6.6 for a detailed description of every cluster.

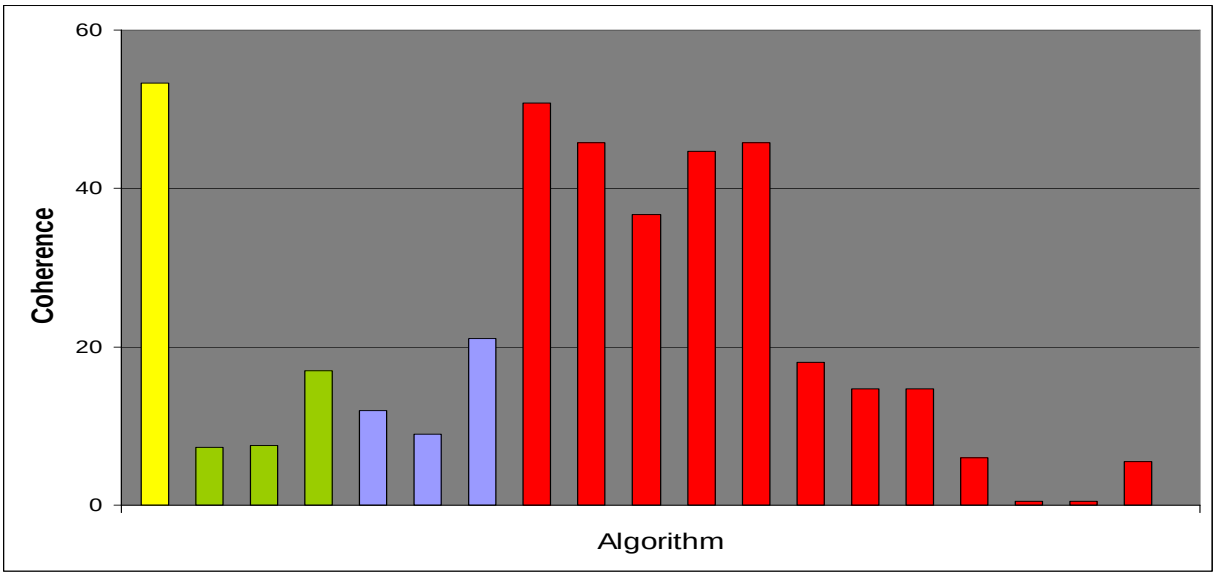

Figure 15: **EachMovie data**: Comparison of average coherence results of the Iclust algorithm (yellow) with conventional clustering algorithms (*12*): K–means (green); K–medians (blue); Hierarchical (red). For the hierarchical algorithms, four different variants are tried: complete, average, centroid, and single linkage, respectively from left to right. For every algorithm, three different similarity measures are applied: Pearson correlation (left); absolute value of Pearson correlation (middle); Euclidean distance (right). In all cases, the results are averaged over all the different numbers of clusters that we tried: $N_c = 5, 10, 15, 20$.

with 100 different initializations), and applied the comparison to all the different numbers of clusters we examined, $N_c = 5, 10, 15, 20$. The results are summarized in Table 16 to Table 19 and in Figure 15.

In all cases, Iclust was clearly superior to the average performance of the K–means and the Hierarchical "Cluster 3.0" variants. In fact, except for the hierarchical complete-linkage configuration with the Pearson correlation as the similarity measure, none of the other algorithms came even close to the Iclust performance. Averaging over all four N_c values, Iclust obtains an average coherence of $\sim 53\%$ while the average K–means coherence is only $\sim 12\%$ and the average Hierarchical coherence is $\sim 24\%$.

Notice, that in contrast to the previous applications, here the K–means algorithms are inferior to some of the Hierarchical algorithms (and both are inferior to Iclust). These results demonstrate that while standard clustering algorithms might work well in certain circumstances and fail completely in others, our principled and model–independent approach maintains a high and robust performance across a wide variety of applications.

Table 16: Coherence results for the EachMovie data with respect to the movie genre annotations with $N_c = 20$ clusters. [a]Clustering algorithm. In the "$\langle$ $K-$means $\rangle$" row we present the average results of all the 6 $K-$means variants. For each of these variants we performed 100 runs from which the best solution is chosen. In the "$\langle$ Hier. $\rangle$" row we present the average results of all the 12 Hierarchical clustering variants. [b]Correlation measure used by the algorithm. "PC" stands for the (centered) Pearson Correlation. "|PC|" is the absolute value of this correlation. "Euclidean" stands for the Euclidean distance. [c]Number of clusters with a positive coherence. [d]Average coherence of all 20 clusters.

$N_c = 20$ **Algorithm** [a]	**Similarity** [b]	N_c^{pos} [c]	$\langle Coh \rangle$ [d]
Iclust	mutual information	15	54
$K-$means	PC	1	3
$K-$means	|PC|	2	4
$K-$means	Euclidean	5	12
$K-$medians	PC	2	5
$K-$medians	|PC|	4	8
$K-$medians	Euclidean	2	6
$\langle$ $K-$means $\rangle$		2.7	6.3
Hier - Comp. linkage	PC	17	55
Hier - Comp. linkage	|PC|	16	51
Hier - Comp. linkage	Euclidean	10	34
Hier - Avg. linkage	PC	12	43
Hier - Avg. linkage	|PC|	12	43
Hier - Avg. linkage	Euclidean	5	19
Hier - Centr. linkage	PC	4	16
Hier - Centr. linkage	|PC|	4	16
Hier - Centr. linkage	Euclidean	2	8
Hier - Sing. linkage	PC	0	0
Hier - Sing. linkage	|PC|	0	0
Hier - Sing. linkage	Euclidean	1	5
$\langle$ Hierarchical $\rangle$		6.9	24.2

Table 17: Coherence results for the EachMovie data with respect to the movie genre annotations with $N_c = 15$ clusters. The column and row definitions are as in Table 16.

$N_c = 15$ **Algorithm** [a]	**Similarity** [b]	N_c^{pos} [c]	$\langle Coh \rangle$ [d]
Iclust	mutual information	11	54
K–means	PC	1	2
K–means	|PC|	1	1
K–means	Euclidean	2	6
K–medians	PC	2	5
K–medians	|PC|	1	3
K–medians	Euclidean	4	14
$\langle$ K–means $\rangle$		21.8	5.2
Hier - Comp. linkage	PC	13	54
Hier - Comp. linkage	|PC|	11	47
Hier - Comp. linkage	Euclidean	6	29
Hier - Avg. linkage	PC	10	47
Hier - Avg. linkage	|PC|	10	46
Hier - Avg. linkage	Euclidean	3	16
Hier - Centr. linkage	PC	2	8
Hier - Centr. linkage	|PC|	2	8
Hier - Centr. linkage	Euclidean	1	3
Hier - Sing. linkage	PC	0	0
Hier - Sing. linkage	|PC|	0	0
Hier - Sing. linkage	Euclidean	1	7
$\langle$ Hierarchical $\rangle$		4.9	22.1

Table 18: Coherence results for the EachMovie data with respect to the movie genre annotations with $N_c = 10$ clusters. The column and row definitions are as in Table 16.

$N_c = 10$ **Algorithm** [a]	**Similarity** [b]	N_c^{pos} [c]	$\langle Coh \rangle$ [d]
Iclust	mutual information	9	57
K–means	PC	1	3
K–means	|PC|	1	6
K–means	Euclidean	4	20
K–medians	PC	2	6
K–medians	|PC|	2	6
K–medians	Euclidean	6	27
$\langle$ K–means $\rangle$		2.7	11.3
Hier - Comp. linkage	PC	8	43
Hier - Comp. linkage	|PC|	8	44
Hier - Comp. linkage	Euclidean	5	36
Hier - Avg. linkage	PC	7	43
Hier - Avg. linkage	|PC|	8	47
Hier - Avg. linkage	Euclidean	2	16
Hier - Centr. linkage	PC	2	12
Hier - Centr. linkage	|PC|	2	12
Hier - Centr. linkage	Euclidean	1	4
Hier - Sing. linkage	PC	0	0
Hier - Sing. linkage	|PC|	0	0
Hier - Sing. linkage	Euclidean	1	10
$\langle$ Hierarchical $\rangle$		3.7	22.3

Table 19: Coherence results for the EachMovie data with respect to the movie genre annotations with $N_c = 5$ clusters. The column and row definitions are as in Table 16.

$N_c = 5$ **Algorithm** [a]	**Similarity** [b]	N_c^{pos} [c]	$\langle Coh \rangle$ [d]
Iclust	mutual information	5	48
K–means	PC	2	21
K–means	|PC|	2	19
K–means	Euclidean	3	30
K–medians	PC	4	32
K–medians	|PC|	2	19
K–medians	Euclidean	4	37
$\langle$ K–means $\rangle$		2.8	26.3
Hier - Comp. linkage	PC	5	51
Hier - Comp. linkage	|PC|	5	41
Hier - Comp. linkage	Euclidean	4	48
Hier - Avg. linkage	PC	4	46
Hier - Avg. linkage	|PC|	4	47
Hier - Avg. linkage	Euclidean	2	21
Hier - Centr. linkage	PC	2	23
Hier - Centr. linkage	|PC|	2	23
Hier - Centr. linkage	Euclidean	1	9
Hier - Sing. linkage	PC	0	0
Hier - Sing. linkage	|PC|	0	0
Hier - Sing. linkage	Euclidean	0	0
$\langle$ Hierarchical $\rangle$		2.4	25.8

6.6 Detailed results for the $N_c = 20$ clusters partition

In Table 20 we present all the enriched annotations for the Iclust partition with $N_c = 20$ clusters and $\frac{1}{T} = 40$. Several results should be noted specifically.

For example, $c12$ consisted solely of 14 classic family movies like “The Wizard of Oz”, “Snow White” and so on. $c8$ consisted mainly of “Art-Foreign” movies, including all the “Three Colors” trilogy by Kieslowski. $c15$ included all seven “Star Trek” movies. Moreover, some of the obtained clusters reflect more subtle distinctions than the broad genre definitions. For example, both $c4$ and $c6$ were enriched for “Comedy”, but while $c4$ was further enriched for “Romance” $c6$ consisted mainly of Jim Carrey and Adam Sandler movies. Both $c7$ and $c17$ were enriched for “Action”, but $c7$ was further enriched for “Classic” with some emphasis over Science Fiction movies like the “Star Wars” trilogy, the “Terminator” movies, “Alien”, “Back to the Future” and so on. In contrast $c17$ consisted mainly of movies starring Sylvester Stallone, Jean-Claude Van Damme etc.

In a separate text file we provide complete details of the specific partition obtained by Iclust for $N_c = 20$ clusters. Similar detailed results for $N_c = 15, 10, 5$ (including the analogous tables of Table 20) are available at request and will be posted online in the corresponding web site.

Table 20: Enriched genre annotations in the Iclust solution with $N_c = 20$ clusters for the EachMovie data. Clusters are ordered as in Figure 12, Figure 13, and Figure 14. Only annotations with a P-value below 0.05 (Bonferroni corrected) are presented. [a]Cluster index. [b]Cluster size. [c]Cluster coherence (in percentage). [d]Enriched annotations. In parentheses: $(x/K, p)$ stands for the number of movies in the cluster to which this annotations is assigned, the number of movies in the entire data to which this annotations is assigned, and the Bonferroni corrected P-value, respectively.

C index[a]	**C size**[b]	**Coh.**[c]	**Enriched annot.**[d]
c14	10	0	–
c2	16	0	–
c19	10	50	Art-Foreign (5/45,0.005254)
c8	10	70	Art-Foreign (7/45,0.000019)
c18	22	0	–
c9	31	55	Action (17/110,0.000281)
c20	155	55	Drama (68/160,0.001170) Romance (30/61,0.011591)
c1	19	95	Classic (10/44,0.000004) Drama (15/160,0.000214)
c7	24	71	Classic (10/44,0.000067) Action (13/110,0.003526)
c15	18	94	Action (16/110,0.000000) Thriller (10/90,0.001778)
c5	32	39	Thriller (12/90,0.034492)
c13	15	40	Horror (6/33,0.001412)
c3	20	0	–
c10	15	0	–
c4	27	74	Romance (12/61,0.000100) Comedy (17/149,0.001613)
c6	11	100	Comedy (11/149,0.000001)
c16	16	87	Comedy (13/149,0.000012)
c17	21	76	Action (16/110,0.000000)
c11	14	71	Family (10/67,0.000003)
c12	14	100	Family (13/67,0.000000) Animation (8/25,0.000000) Classic (5/44,0.019004)

References and Notes

1. Slonim, N., Atwal, G. S., Tkačik, G. & Bialek W. (2005) *http://arxiv.org/abs/cs.IT/0502017*.

2. Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel–Harel, O., Eisen, M. B., Storz, G., Botstein, D. & Brown, P. O. (2000) *Mol Biol Cell* **11,** 4241–4257.

3. Gasch, A. P. (2002) in *Topics in Current Genetics*, eds. Hohmann, S. & Mager, P. (Springer-Verlag, Heidelberg), Vol. 1, pp. 11–70.

4. Cover, T. M. & Thomas, J. A. (1991) *Elements of Information Theory* (John Wiley and Sons, New York).

5. Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) *Journal of the Royal Statistical Society B* **39**, 1–38.

6. Tishby, N., Pereira, F. C. & Bialek, W. (1999) in *Proc. of the 37th Annual Allerton Conference on Communication, Control and Computing*, eds. Hajek, B. & Sreenivas, R. S. (Urbana, Illinois), pp. 369-477.

7. Neal, R. M. & Hinton, G. E. (1998) in *Learning in Graphical Models*, ed. Jordan, M. I. (Kluwer Academic Publishers, Dordrecht), pp. 355–368.

8. Durrett, R. (1991) *Probability Theory and Examples* (Wadsworth and Brookes, Cole, California).

9. Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. & Friedman, N. (2003) *Nature Gen* **34**, 166–176.

10. Jain, A. K., Murty, M. N. & Flynn, P. J. (1999) *ACM Computing Surveys* **31**, 264–323.

11. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig J. T. *et al.* (2000) *Nature Gen* **25** 25–29.

12. de Hoon, M. J. L., Imoto, S., Nolan, J. & Miyano, S. (2004) *Bioinformatics* **20**, 1453–1454.

13. Woan, G. (2000) *The Cambridge Handbook of Physics Formulas* (Cambridge University Press).

14. Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B. & D Botstein (2003) *Proc Nat Acad Sci (USA)* **100**, 8348–8353.

15. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. R., Thompson, C. M., Simon, I. *et al.* (2002) *Science* **298**, 799–804.

16. Pilpel, Y., Sudarsanam, P. & Church, G. M. (2001) *Nature Gen* **29**, 153–159.