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, NJ 08544

Communicated by David Mumford, Brown University, Providence, RI, September 9, 2005 (received for review April 8, 2005)

In an age of increasingly large data sets, investigators in many different disciplines have turned to clustering as a tool for data analysis and exploration. Existing clustering methods, however, typically depend on several nontrivial assumptions about the structure of data. Here, we reformulate the clustering problem from an information theoretic perspective that avoids many of these assumptions. In particular, our formulation obviates the need for defining a cluster "prototype," does not require an *a priori* similarity metric, is invariant to changes in the representation of the data, and naturally captures nonlinear relations. We apply this approach to different domains and find that it consistently produces clusters that are more coherent than those extracted by existing algorithms. Finally, our approach provides a way of clustering based on collective notions of similarity rather than the traditional pairwise measures.

information theory | rate distortion | cluster analysis | gene expression

The idea that complex data can be grouped into clusters or categories is central to our understanding of the world, and this structure arises in many diverse contexts (e.g., Table 1). In popular culture we group films or books into genres; in business we group companies into sectors of the economy; in biology we group the molecular components of cells into functional units or pathways, and so on. Typically, these groupings are first constructed by hand using specific but qualitative knowledge; e.g., Dell and Apple belong in the same group because they both make computers. The challenge of clustering is to ask whether these qualitative groupings can be derived automatically from objective, quantitative data. Is our intuition about sectors of the economy derivable, for example, from the dynamics of stock prices? Are the functional units of the cell derivable from patterns of gene expression under different conditions (1, 2)? The literature on clustering, even in the context of gene expression, is vast (3). Our goal here is not to suggest yet another clustering algorithm, but rather to focus on questions about the formulation of the clustering problem. We are led to an approach, grounded in information theory, that should have wide applicability.

Our intuition about clustering starts with the obvious notion that similar elements should fall within the same cluster, whereas dissimilar ones should not. But clustering also achieves data compression: instead of identifying each data point individually, we can identify points by the cluster to which they belong, ending up with a simpler and shorter description of the data. Rate-distortion theory (4, 5) formulates precisely the tradeoff between these two considerations, searching for assignments to clusters such that the number of bits used to describe the data are minimized while the average similarity between each data point and its cluster representative (or prototype) is maximized. A well known limitation of this formulation (as in most approaches to clustering) is that one needs to specify the similarity measure in advance, and quite often this choice is made arbitrarily. Another issue, which attracts less attention, is that the notion of a representative or "cluster prototype" is inherent to this formulation, although it is not always obvious how to define this concept. Our approach provides plausible answers to both these concerns, with further interesting consequences.

Theory

Theoretical Formulation. Imagine that there are N elements ($i = 1, 2, \ldots, N$) and N_c clusters ($C = 1, 2, \ldots, N_c$) and that we have assigned elements i to clusters C according to some probabilistic rules, $P(C \mid i)$, that serve as the variables in our analysis.[†] If we reach into a cluster and pull out elements at random, we would like these elements to be as similar to one another as possible. Similarity usually is defined among pairs of elements (e.g., the closeness of points in some metric space), but as noted below we also can construct more collective measures of similarity among $r > 2$ elements; perhaps surprisingly we will see that that this more general case can be analyzed at no extra cost. Leaving aside for the moment the question of how to measure similarity, let us assume that computing the similarity among r elements $i_1, i_2, \ldots, i_r$ returns a similarity measure $s(i_1, i_2, \ldots, i_r)$. The average similarity among elements chosen out of a single cluster is

$$s(C) = \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_r=1}^{N} P(i_1|C)P(i_2|C)\cdots P(i_r|C)s(i_1, i_2, \ldots, i_r), \quad [1]$$

where $P(i|C)$ is the probability to find element i in cluster C. This average similarity corresponds to a scenario where one chooses the elements $\{i_1 i_1, i_2, \ldots, i_r\}$ at random out of a cluster C, independently of each other; other formulations also might be plausible. From Bayes' rule we have $P(i \mid C) = P(C \mid i)P(i)/P(C)$, where $P(C)$ is the total probability of finding any element in cluster C, $P(C) = \Sigma_i P(C \mid i)P(i)$. In many cases the elements i occur with equal probability so that $P(i) = 1/N$. We further consider this case for simplicity, although it is not essential. The intuition about the "goodness" of the clustering is expressed through the average similarity over all of the clusters

$$\langle s \rangle = \sum_{C=1}^{N_c} P(C)s(C). \quad [2]$$

For the special case of pairwise "hard" clustering, we obtain $\langle s \rangle_h = (1/N)\Sigma_{C,i,j}\,(1/|C|)s(i, j)$, where $|C|$ is the size of cluster C. This simpler form was shown in ref. 6 to satisfy basic invariance and robustness criteria.

The task then is to choose the assignment rules $P(C \mid i)$ that maximize $\langle s \rangle$, while, as in rate-distortion theory, simultaneously compressing our description of the data as much as possible. To implement this intuition we maximize $\langle s \rangle$ while constraining the information carried by the cluster identities (5)

$$I(C; i) = \frac{1}{N} \sum_{i=1}^{N} \sum_{C=1}^{N_c} P(C \mid i)\log\left[\frac{P(C \mid i)}{P(C)}\right]. \quad [3]$$

Conflict of interest statement: No conflicts declared.

Abbreviations: ESR, environmental stress response; GO, Gene Ontology.

*To whom correspondence should be addressed. E-mail: nslonim@princeton.edu.

[†]Conventionally, one distinguishes "hard" clustering, in which each element is assigned to exactly one cluster, and "soft" clustering in which the assignments are probabilistic, described by a conditional distribution $P(C \mid i)$; we consider here the more general soft clustering with hard clustering emerging as a limiting case.

Table 1. Examples of clusters in three different data sets

Cluster	Members	Description
Genes		
C_{18}	RPS10A, RPS10B, RPS11A, RPS11B, RPS12	Proteins of the small ribosomal subunit
C_{15}	FRS1, KRS1, SES1, TYS1, VAS1	Enzymes that attach amino acids to tRNA
C_4	PGM2, UGP1, TSL1, TPS1, TPS2	Enzymes involved in the trehalose anabolism pathway
Stocks		
C_{17}	Wal-Mart, Target, Home Depot, Best Buy, Staples	
C_{12}	Microsoft, Apple Comp., Dell, HP, Motorola	
C_2	NY Times, Tribune Co., Meredith Corp., Dow Jones & Co., Knight-Ridder Inc.	
Movies		
C_{12}	*Snow White*, *Cinderella*, *Dumbo*, *Pinocchio*, *Aladdin*	
C_1	*Psycho*, *Apocalypse Now*, *The Godfather*, *Taxi Driver*, *Pulp Fiction*	
C_7	*Star Wars*, *Return of the Jedi*, *The Terminator*, *Alien*, *Apollo 13*	

For each cluster, a sample of five typical items is presented. All clusters were found through the same automatic procedure.

Thus, our mathematical formulation of the intuitive clustering problem is to maximize the functional

$$\mathcal{F} = \langle s \rangle - TI(C; i), \qquad [4]$$

where the Lagrange multiplier T enforces the constraint on $I(C; i)$. Notice that, as in other formulations of the clustering problem, $\mathcal{F}$ resembles the free energy in statistical mechanics, where the temperature T specifies the tradeoff between energy and entropy like terms.

This formulation is intimately related to conventional rate-distortion theory. In rate-distortion clustering, one is given a fixed number of bits with which to describe the data, and the goal is to use these bits so as to minimize the distortion between the data elements and some representatives of these data. In practice, the bits specify membership in a cluster, and the representatives are prototypical or average patterns in each cluster. Here we see that we can formulate a similar tradeoff with no need to introduce the notion of a representative or average; instead, we measure directly the similarity of elements within each cluster; moreover, we can consider collective rather than pairwise measures of similarity. A more rigorous treatment detailing the relation between Eq. **4** and the conventional rate-distortion functional will be presented elsewhere.

Optimal Solution. In general it is not possible to find an explicit solution for the $P(C \mid i)$ that maximize $\mathcal{F}$. However, if we assume that $\mathcal{F}$ is differentiable with respect to the variables $P(C \mid i)$, equating the derivative to zero yields after some algebra a set of implicit, self-consistent equations that any optimal solution must obey:

$$P(C|i) = \frac{P(C)}{Z(i; T)} \exp\left\{ \frac{1}{T} [r s(C; i) - (r-1) s(C)] \right\}, \qquad [5]$$

where $Z(i; T)$ is a normalization constant and $s(C; i)$ is the expected similarity between i and $r-1$ members of cluster C

$$s(C; i) = \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_{r-1}=1}^{N} P(i_1|C) P(i_2|C) \cdots P(i_{r-1} \mid C) s(i_1, i_2, \cdots, i_{r-1}, i). \qquad [6]$$

The derivation of these equations from the optimization of $\mathcal{F}$ is reminiscent of the derivation of the rate-distortion (5) or information bottleneck (7) equations. This simple form is valid when the similarity measure is invariant under permutations of the arguments. In the more general case we have

$$P(C \mid i) = \frac{P(C)}{Z(i; T)} \exp\left\{ \frac{1}{T} \left[\sum_{r'=1}^{r} s(C; i^{(r')}) - (r-1) s(C) \right] \right\}, \qquad [7]$$

where $s(C; i^{(r')})$ is the expected similarity between i and $r-1$ members of cluster C when i is the r' argument of s.

An obvious feature of Eq. **5** is that element i should be assigned to cluster C with higher probability if it is more similar to the other elements in the cluster. Less obvious is that this similarity has to be weighed against the mean similarity among all of the elements in the cluster. Thus, our approach automatically embodies the intuitive principle that "tightly knit" groups are more difficult to join. We emphasize that we did not explicitly impose this property, but rather it emerges directly from the variational principle of maximizing $\mathcal{F}$; most other clustering methods do not capture this intuition.

The probability $P(C \mid i)$ in Eq. **5** has the form of a Boltzmann distribution, and increasing similarity among elements of a cluster plays the role of lowering the energy; the temperature T sets the scale for converting similarity differences into probabilities. As we lower this temperature, there is a sequence of "phase transitions" to solutions with more distinct clusters that achieve greater mean similarity in each cluster (8). For a fixed number of clusters, reducing the temperature yields more deterministic $P(C \mid i)$ assignments.

Algorithm. Although Eq. **5** is an implicit set of equations, we can turn this self-consistency condition into an iterative algorithm that finds an explicit numerical solution for $P(C \mid i)$ that corresponds to a (perhaps local) maximum of $\mathcal{F}$. In Fig. 5, which is published as supporting information on the PNAS web site, we present pseudocode for the algorithm in the case $r = 2$. 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 that maximizes $\mathcal{F} = \langle s \rangle - TI(C; i)$. We emphasize that we use 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 the different alternatives.

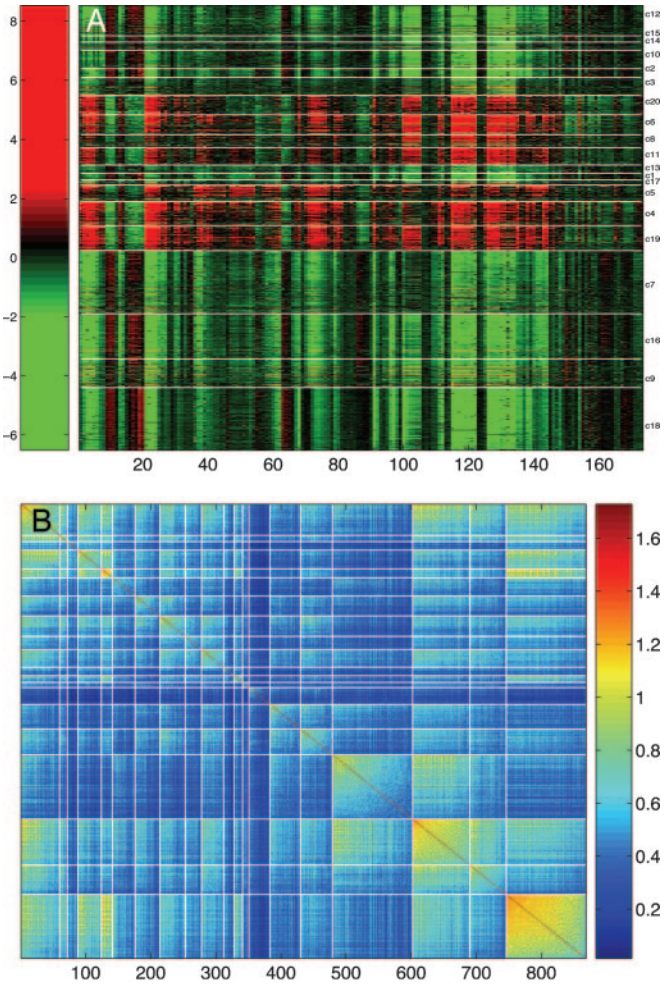

Fig. 1. ESR data and information relations. (*A*) Expression profiles of the ≈900 genes in the yeast ESR module across the 173 microarray stress experiments (12). (*B*) Mutual information relations (in bits) among the ESR genes. In both *A* and *B* the genes are sorted according to the solution with 20 clusters and a relatively saturated ⟨*s*⟩. Inside each cluster, genes are sorted according to their average mutual information relation with other cluster members.

Information as a Similarity Measure. In formulating the clustering problem as the optimization of $\mathcal{T}$, we have used, as in rate-distortion theory, the generality of information theory to provide a natural measure for the cost of dividing the data into more clusters, but the similarity measure remains arbitrary and commonly is believed to be problem specific. Is it possible to use information theory to address this issue as well? To be concrete, consider the case where the elements i are genes and we are trying to measure the relation between gene expression patterns across a variety of conditions $\mu = 1, 2, \ldots, M$; gene i has expression level $e_i(\mu)$ under condition μ. We imagine that there is some real distribution of conditions that cells encounter during their lifetime, and an experiment with a finite set of conditions provides samples out of this distribution. Then, for each gene we can define the probability density of expression levels

$$P_i(e) = \frac{1}{M}\sum_{\mu=1}^{M}\delta(e - e_i(\mu)), \qquad [8]$$

which should become smooth as $M \to \infty$. Similarly we can define the joint probability density for the expression levels of r genes $i_1, i_2, \ldots, i_r$

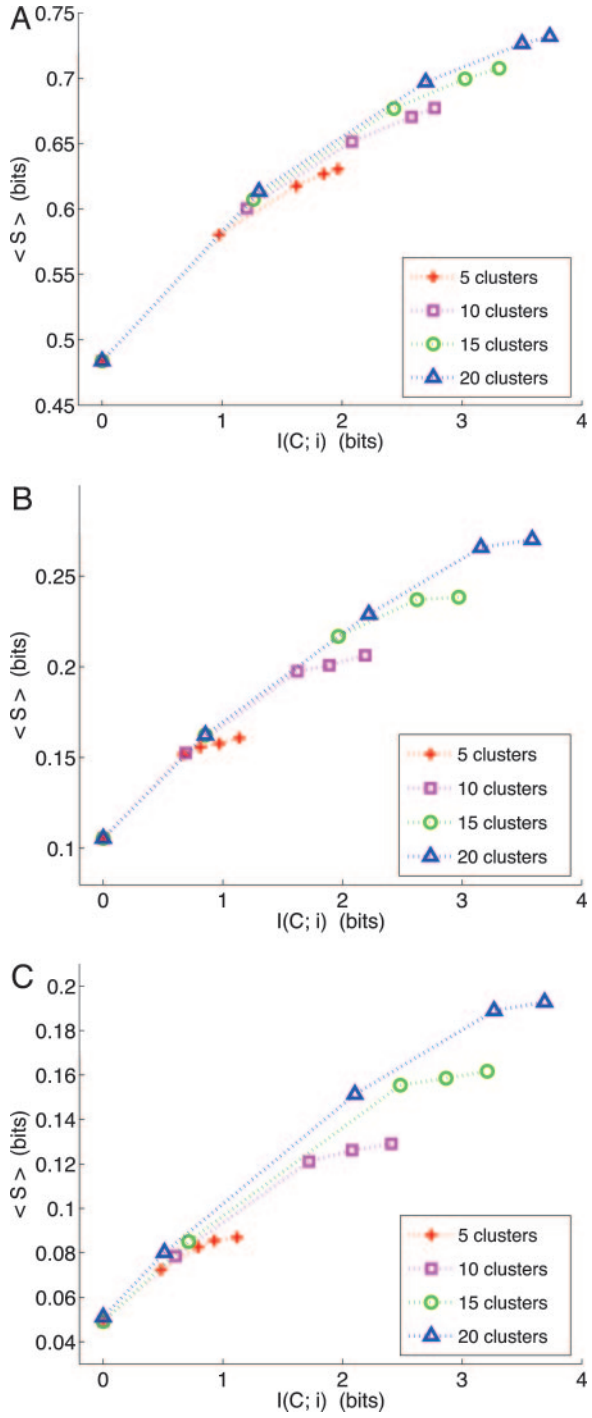

Fig. 2. Tradeoff curves in all three applications. Each curve describes the solutions obtained for a particular number of clusters. Different points along each curve correspond to different local maxima of F at different T values. (*A*) Tradeoff curves for the ESR data with $1/T = \{5, 10, 15, 20, 25\}$. In Fig. 4, we explore the possible hierarchical relations between the four saturated solutions at $1/T = 25$. (*B*) Tradeoff curves for the Standard & Poor's 500 data with $1/T = \{15, 20, 25, 30, 35\}$. (*C*) Tradeoff curves for the EachMovie data with $1/T = \{20, 25, 30, 35, 40\}$.

$$P_{i_1 i_2 \cdots i_r}(e_1, e_2, \ldots, e_r) = \frac{1}{M}\sum_{\mu=1}^{M}\delta(e_1 - e_{i_1}(\mu))\delta(e_2 - e_{i_2}(\mu))\cdots\delta(e_r - e_{i_r}(\mu)). \qquad [9]$$

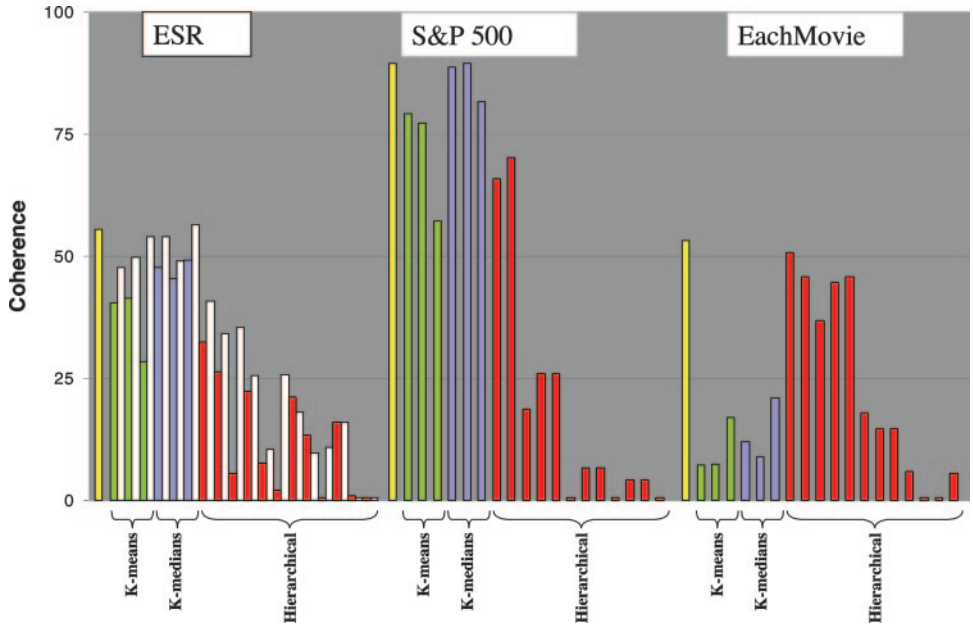

Fig. 3. Comparison of coherence results of our approach (yellow) with conventional clustering algorithms (17). Green, *K*-means; blue, *K*-medians; red, hierarchical. For the hierarchical algorithms, four different variants are tried as follows: (from left to right) complete, average, centroid, and single linkage. For every algorithm, three different similarity measures are applied as follows: Pearson correlation (left), absolute value of Pearson correlation (middle), and Euclidean distance (right). The white bars in the ESR data correspond to applying the algorithm to 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, $N_c = 5, 10, 15, 20$. For the ESR data, coherence is measured with respect to each of the three GOs, and the results are averaged.

Given the joint distributions of expression levels, information theory provides natural measures of the relations among genes. For $r = 2$, we can identify the relatedness of genes i and j with the mutual information between the expression levels

$$s(i,j) = I_{i,j} = \int de_1 \int de_2 P_{ij}(e_1, e_2) \log_2 \left[\frac{P_{ij}(e_1, e_2)}{P_i(e_1) P_j(e_2)} \right] \text{ bits.} \quad [10]$$

This measure is naturally extended to the multiinformation among multiple variables (9), or genes

$$I^{(r)}_{i_1,i_2,\ldots,i_r} = \int d^r e \, P_{i_1 i_2 \cdots i_r}(e_1, e_2, \ldots, e_r) \times \log_2 \left[\frac{P_{i_1 i_2 \cdots i_r}(e_1, e_2, \ldots, e_r)}{P_{i_1}(e_1) P_{i_2}(e_2) \cdots P_{i_r}(e_r)} \right] \text{ bits.} \quad [11]$$

We recall that the mutual information is the unique measure of relatedness between a pair of variables that obeys several simple and desirable requirements independent of assumptions about the form of the underlying probability distributions (4). In particular, the mutual (and multi-) information is independent of invertible transformations on the individual variables. For example, the mutual information between the expression levels of two genes is identical to the mutual information between the log of the expression levels: there is no need to find the "right" variables with which to represent the data. The absolute scale of the information measure also has a clear meaning. For example, if two genes share more than one bit of information, then the underlying biological mechanisms must be more subtle than just turning expression on and off. In addition, the mutual information reflects any type of dependence among variables, whereas ordinary correlation measures typically ignore nonlinear dependences.

Although these theoretical advantages are well known, in practice information theoretic quantities are notoriously difficult to estimate from finite data. For example, although the distributions in Eqs. **8** and **9** become smooth in the limit of many samples ($M \to \infty$), with a finite amount of data one needs to regularize or discretize the distributions, and this process could introduce artifacts. Although there is no completely general solution to these problems, we have found that in practice the difficulties are not as serious as one might have expected. By using an adaptation of the "direct" estimation method originally developed in the analysis of neural coding (10), we have found that one can obtain reliable estimates of mutual (and sometimes multi-) information values for a variety of data types, including gene expression data (11). In particular, experiments which explore gene expression levels under >100 conditions are sufficient to estimate the mutual information between pairs of genes with an accuracy of ≈0.1 bits.[‡]

To summarize, we have suggested a purely information-theoretic approach to clustering and categorization: relatedness among elements is defined by the mutual (or multi-) information, and optimal clustering is defined as the best tradeoff between maximizing this average relatedness within clusters and minimizing the number of bits required to describe the data. The result is a formulation of clustering that trades bits of similarity against bits of descriptive power, with no further assumptions. A freely available web implementation of the clustering algorithm and the mutual information estimation procedure is available from the web site of the Lewis–Sigler Institute for Integrative Genomics.

Results

Gene Expression. As a first test case we consider experiments on the response of gene expression levels in yeast to various forms of environmental stress (12). Previous analysis identified a group of ≈300 stress-induced and ≈600 stress-repressed genes with "nearly identical but opposite patterns of expression in response to the environmental shifts" (13), and these genes were termed the environmental stress response (ESR) module. In fact, based on this observation, these genes were excluded from recent further analysis of the entire yeast genome (14). Nonetheless, as we shall see next, our approach automatically reveals further rich and meaningful substructure in these data.

As seen in Fig. 1*A*, differences in expression profiles within the

[‡]It should be noted that in applications where there is a natural similarity measure it might be advantageous to use this measure directly. Furthermore, in situations where the number of observations is not sufficient for nonparametric estimates of the information relations, other heuristic similarity measures should be employed, or one could use parametric models for the underlying distributions. Notice, though, that these alternative measures can be incorporated into the algorithm in Fig. 5.

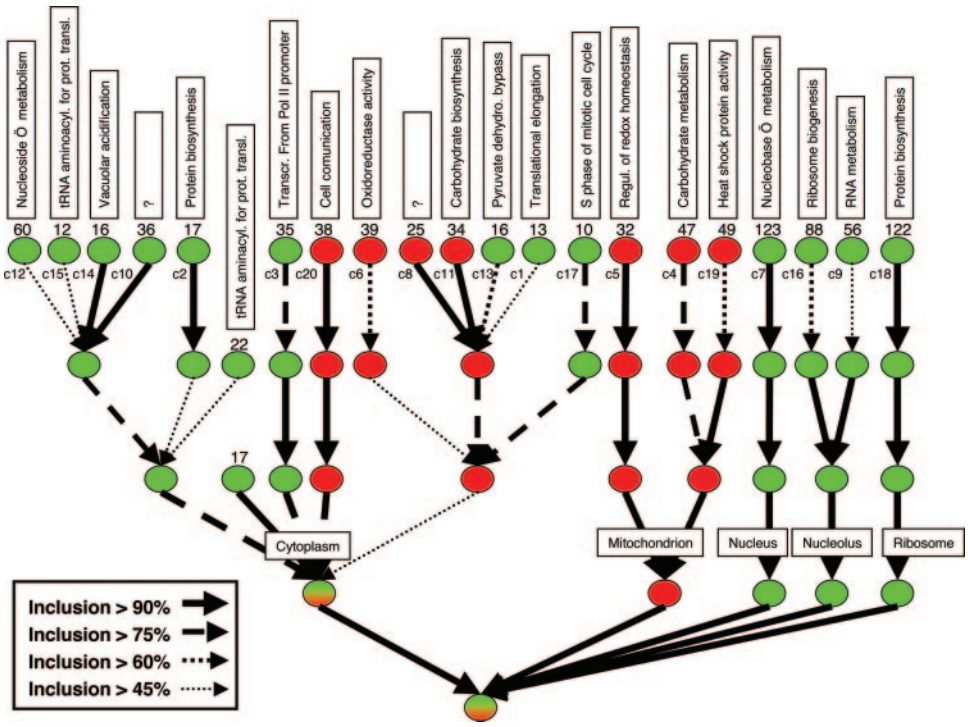

Fig. 4. Relations between the optimal solutions with $N_c = \{5, 10, 15, 20\}$ at $1/T = 25$ for the ESR data. Every cluster is connected to the cluster in the next, less detailed, partition that absorbs its most significant portion. The edge type indicates the level of inclusion. The independent solutions form an approximated hierarchical structure. At the upper level the clusters are sorted as in Fig. 1. The number above every cluster indicates the number of genes in it, and the text title corresponds to the most enriched GO biological–process annotation in this cluster. The titles of the five clusters at the lower level are their most enriched GO cellular-component annotation. Most clusters were enriched with more than one annotation; hence, the short titles sometimes are too concise. Red and green clusters represent clusters with a clear majority of stress-induced or stress-repressed genes, respectively.

ESR module indeed are relatively subtle. However, when considering the mutual information relations (Fig. 1*B*), a relatively clear structure emerges. We have solved our clustering problem for $r = 2$ and various numbers of clusters and temperatures. The resulting concave tradeoff curves between $\langle s \rangle$ and $I(C; i)$ are shown in Fig. 2*A*. We emphasize that we generate not a single solution, but a whole family of solutions describing structure at different levels of complexity. With the number of clusters fixed, $\langle s \rangle$ gradually saturates as the temperature is lowered and the constraint on $I(C; i)$ is relaxed. For the sake of brevity, we focused our analysis on the four solutions for which the saturation of $\langle s \rangle$ is relatively clear ($1/T = 25$). At this temperature, ≈85% of the genes have nearly deterministic assignments to one of the clusters [$P(C \mid i) > 0.9$ for a particular C]. As an illustration, 3 of the 20 clusters found at this temperature are in fact the clusters presented in Table 1.

We have assessed the biological significance of our results by considering the distribution of gene annotations across the clusters and estimating the corresponding clusters' coherence[§] with respect to all three Gene Ontologies (GOs) (16). Almost all of our clusters were significantly enriched in particular annotations. We compared our performance with 18 different conventional clustering algorithms that are routinely applied to this data type (17). We used the clustering software available at http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster to implement the conventional algorithms. In Fig. 3 we see that our clusters obtained the highest average coherence, typically by a significant margin. Moreover, even when the competing algorithms cluster the $\log_2$ of expression (ratio) profiles, a common regularization used in this application with no formal justification, our results are comparable with or superior to all of the alternatives.

Instead of imposing a hierarchical structure on the data, as done in many popular clustering algorithms, here we directly examine the relations between solutions at different numbers of clusters that were found independently.[¶] In Fig. 4, we see that an approximate hierarchy emerges as a result rather than as an implicit assumption, where some functional modules (e.g., the "ribosome cluster", C_{18}) are better preserved than others.

Our attention is drawn also to the cluster C_7, which is found repeatedly at different numbers of clusters. Specifically, at the solution with 20 clusters, among the 114 repressed genes in C_7, 69 have an uncharacterized molecular function; this level of concentration has a probability of $\approx 10^{-15}$ to have arisen by chance. 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. At least for this cluster, our results indicate a different scenario where a significant portion of tightly coexpressed genes remain uncharacterized to date.

Stock Prices. To emphasize the generality of our approach we consider a very different data set, the day-to-day fractional changes in price of the stocks in the Standard & Poor's 500 list (available at www.standardandpoors.com), during the trading days of 2003. We cluster these data exactly as in our analysis of gene expression data. The resulting tradeoff curves are shown in Fig. 2*B*, and again we focus on the four solutions where $\langle s \rangle$ already saturates.

To determine the coherence of the ensuing clusters we used the Global Industry Classification Standard (available at http://wrds.wharton.upenn.edu), which classifies companies at four different levels: sector, industry group, industry, and subindustry. Thus, each company is assigned four annotations, which are organized in a hierarchical tree, somewhat similar to the GO hierarchical annotation (16).

As before, our average coherence performance is comparable

[§]Specifically, the coherence of a cluster (14) is defined as the percentage of elements in this cluster that are annotated by an annotation that was found to be significantly enriched in this cluster ($P < 0.05$, with the Bonferroni correction for multiple hypotheses). See the technical report (15) for a detailed discussion regarding the statistical validation of our results.

[¶]In standard agglomerative or hierarchical clustering one starts with the most detailed partition of singleton clusters and obtains new solutions through merging of clusters. Consequently, one must end up with a tree-like hierarchy of clustering partitions, regardless of whether the data structure actually supports this description.

with or superior to all of the other 18 clustering algorithms we examined (Fig. 3). Almost all our clusters, at various levels of N_c, exhibit a surprisingly high degree of coherence with respect to the "functional labels" that correspond to the different (sub-) sectors of the economy. The four independent solutions, at $N_c = \{5, 10, 15, 20\}$ and $1/T = 35$, naturally form an approximate hierarchy (15).

We have analyzed in detail the results for $N_c = 20$ and $1/T = 35$ where selections from three of the derived clusters are shown in Table 1. Eight of the clusters are found to be perfectly (100%) coherent, capturing subtle differences between industrial sectors. For example, two of the perfectly coherent clusters segregate companies into either investment banking and asset management (e.g., Merrill Lynch) or commercial regional banks (e.g., PNC). Even in clusters with less than perfect coherence, we are able to observe and explain relationships between intracluster companies above and beyond what the annotations may suggest. For example, one cluster is enriched with "Hotel Resorts and Cruise Line" companies at a coherence level of 30%. Nonetheless, the remaining companies in this cluster also seem to be tied with the tourism industry, like the Walt Disney Co., banks that specialize in credit card issuing, and so on.

Movie Ratings. Finally, we consider a third test case of yet another different nature: movie ratings provided by >70,000 viewers (the EachMovie database, www.research.digital.com/SRC/eachmovie). Unlike the previous cases, the data here is already naturally quantized because only six possible ratings were permitted.

We proceed as before to cluster the 500 movies that received the maximal number of votes. The resulting tradeoff curves are presented in Fig. 2*C*. Few clusters are preserved among the solutions at different numbers of N_c, suggesting that a hierarchical structure may not be a natural representation of the data. Cluster coherence was determined with respect to the genre labels provided in the database: action, animation, art-foreign, classic, comedy, drama, family, horror, romance, and thriller. Fig. 3 demonstrates that our results are superior to all of the other 18 standard clustering algorithms.

We have analyzed in detail the results for $N_c = 20$ and $1/T = 40$ where, once again, selections from three of the derived clusters are shown in Table 1. The clusters indeed reflect the various genres but also seem to capture subtle distinctions between sets of movies belonging to the same genre. For example, two of the clusters are both enriched in the action genre, but one group consists mainly of science-fiction movies and the other consists of movies in contemporary settings.

Details of all three applications are given in a separate technical report (15).

Discussion

Measuring the coherence of clusters corresponds to asking whether the automatic, objective procedure embodied in our optimization principle does indeed recover the intuitive labeling constructed by human hands. Our success in recovering functional categories in different systems using exactly the same principle and practical algorithm is encouraging. It should be emphasized that our approach is not a model of each system and that there is no need for making data-dependent decisions in the representation of the data or in the definition of similarity.

Most clustering algorithms embody, perhaps implicitly, different models of the underlying statistical structure.[‖] In principle, more accurate models should lead to more meaningful clusters. However, the question of how to construct an accurate model obviously is quite involved, raising further issues that often are addressed arbitrarily before the cluster analysis begins. Moreover, as is clear from Fig. 3, an algorithm or model that is successful in one data type might fail completely in a different domain; even in the context of gene expression, successful analysis of data taken under one set of conditions does not necessarily imply success in a different set of conditions, even for the same organism. Our use of information theory allows us to capture the relatedness of different patterns independent of assumptions about the nature of this relatedness. Correspondingly, we have a single approach that achieves high performance across different domains.

Finally, our approach can succeed where other methods would fail qualitatively. Conventional algorithms search for linear or approximately linear relations among the different variables, whereas our information-theoretic approach is responsive to any type of dependencies, including strongly nonlinear structures. In addition, although the cluster analysis literature has focused thus far on pairwise relations and similarity measures, our approach sets a sound theoretical framework for analyzing complex data based on higher-order relations. Indeed, it was recently demonstrated, both in principle (18) and in practice (19), that in some situations the data structure is obscured at the pairwise level but clearly manifests itself only at higher levels. The question of how common such data are, as well as the associated computational difficulties in analyzing such higher-order relations, is yet to be explored.

[‖]For example, the *K*-means algorithm corresponds to maximizing the likelihood of the data on the assumption that these are generated through a mixture of spherical Gaussians.

We thank O. Elemento and E. Segal for their help in connection with the analysis of the ESR data and C. Callan, D. Botstein, N. Friedman, M. Rothschild, R. Schapire, and S. Tavazoie for helpful comments on earlier versions of this manuscript. This work was supported in part by National Institutes of Health Grant P50 GM071508. G.T. was supported by the Burroughs-Wellcome Graduate Training Program in Biological Dynamics.

1. Brown, P. O. & Botstein, D. (1999) *Nat. Genet.* **21,** 33–37.
2. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 14863–14868.
3. Jain, A. K., Murty, M. N. & Flynn, P. J. (1999) *ACM Comput. Surv.* **31,** 264–323.
4. Shannon, C. E. (1948) *Bell Sys. Tech. J.* **27,** 379–423, 623–656.
5. Cover, T. M. & Thomas, J. A. (1991) *Elements of Information Theory* (Wiley, New York).
6. Puzicha, J., Hofmann, T. & Buhmann, J. M. (2000) *Pattern Recognit.* **33,** 617–634.
7. Tishby, N., Pereira, F. C. & Bialek, W. (1999) in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, eds. Hajek, B. & Sreenivas, R. S. (Univ. of Illinois, Urbana), pp. 368–377.
8. Rose, K. (1998) *Proc. IEEE* **86,** 2210–2239.
9. Studen, M. & Vejnarova, J. (1998) in *Learning in Graphical Models*, ed. Jordan, M. I. (Kluwer Academic, Dordrecht, The Netherlands), pp. 261–298.
10. Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R. & Bialek, W. (1998) *Phys. Rev. Lett.* **80,** 197–200.
11. Slonim, N., Atwal, G. S., Tkačik, G. & Bialek, W. (2005) http://arxiv.org/abs/cs.IT/0502017.
12. 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.
13. Gasch, A. P. (2002) in *Topics in Current Genetics*, eds. Hohmann, S. & Mager, P. (Springer, Heidelberg), Vol. 1, pp. 11–70.
14. Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. & Friedman, N. (2003) *Nat. Genet.* **34,** 166–176.
15. Slonim, N., Atwal, G. S., Tkačik, G. & Bialek, W. (2005) http://arxiv.org/abs/q-bio.QM/0511042.
16. 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) *Nat. Genet.* **25,** 25–29.
17. de Hoon, M. J. L., Imoto, S., Nolan, J. & Miyano, S. (2004) *Bioinformatics* **20,** 1453–1454.
18. Schneidman, E., Still, S., Berry, M. J., II, & Bialek, W. (2003) *Phys. Rev. Lett.* **91,** 238701(4).
19. Bowers, P. M., Cokus, S. J., Eisenberg, D. & Yeates, T. O. (2004) *Science* **306,** 2246–2249.