Significance Analysis and Statistical Mechanics: An Application to Clustering

Marta Łuksza,[1] Michael Lässig,[2] and Johannes Berg[2]
[1]Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, 14195 Berlin, Germany
[2]Institut für Theoretische Physik, Universität zu Köln, Zülpicher Straße 77, 50937 Köln, Germany

(Received 27 November 2009; published 23 November 2010)

This Letter addresses the statistical significance of structures in random data: Given a set of vectors and a measure of mutual similarity, how likely is it that a subset of these vectors forms a cluster with enhanced similarity among its elements? The computation of this cluster p value for randomly distributed vectors is mapped onto a well-defined problem of statistical mechanics. We solve this problem analytically, establishing a connection between the physics of quenched disorder and multiple-testing statistics in clustering and related problems. In an application to gene expression data, we find a remarkable link between the statistical significance of a cluster and the functional relationships between its genes.

DOI: 10.1103/PhysRevLett.105.220601 PACS numbers: 05.10.Gg, 02.50.−r, 05.40.−a, 07.05.Kf

Clustering is a heavily used method to group the elements of a large data set by mutual similarity. It is usually applied without information on the mechanism producing similar data vectors. Any clustering depends on two ingredients: a notion of similarity between elements of the data set, which leads to a scoring function for clusters, and an algorithmic procedure to group elements into clusters. Diverse methods address both aspects of clustering: similarities can be defined by Euclidean or by information-theoretic measures [1,2], and there are many different clustering algorithms ranging from classical k means [3] and hierarchical clustering [4] to recent message-passing techniques [5].

An important aspect of clustering is its statistical significance, which poses a conceptual problem beyond scoring and algorithmics. First, we have to distinguish "true" clusters from spurious clusters, which occur also in random data. An example is the starry sky: true clusters are galaxies with their stars bound to each other by gravity, but there are also spurious constellations of stars which are in fact unrelated and may be far from one another. Second, clustering procedures generally produce different and competing results, since their scoring functions depend on free parameters. The most important scoring parameter weighs the number versus the size of clusters and is contained explicitly (e.g., the number k in k-means clustering) or implicitly (e.g., the temperature in superparamagnetic [6] and information-based clustering [2]) in all clustering procedures. Choosing smaller values of k will give fewer, but larger clusters with lower average similarity between elements. Larger values of k will result in more, but smaller clusters with higher average similarity. None of these choices is *a priori* better than any other: both tight and loose clusters may reflect important structural similarities within a data set.

Addressing the cluster significance problem requires a statistical theory of clustering, which is the topic of this Letter. Our aim is not to propose a new method for clustering, but to distinguish significant clusters from insignificant ones. The key result of the Letter is the analytic computation of the so-called cluster p value $p(S)$, defined as the probability that a random data set contains a cluster with a similarity score larger than S. This result provides a conceptual and practical improvement over current methods of estimating p values by simulation of an ensemble of random data sets, which are computationally intensive [7] and, hence, often omitted in practice.

Our approach is based on an intimate connection between cluster statistics and the physics of disordered systems. The score S of the highest-scoring cluster in a set of random vectors is itself a random variable, whose cumulative probability distribution defines the p value $p(S)$. For significance analysis, we are specifically interested in the large-S tail of this distribution. Our calculation employs the statistical mechanics of a system whose Hamiltonian is given by (minus) the similarity score function. In this system, $\log p(S)$ is the entropy of all data vector

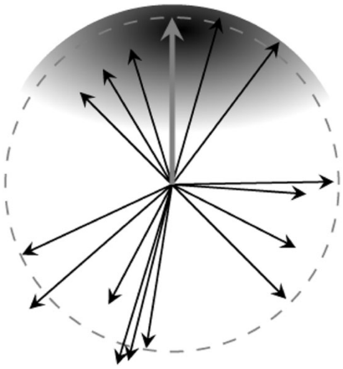

FIG. 1. Clustering a set of random vectors. In a set of randomly chosen vectors, subsets of highly similar elements can arise by chance. Here a cluster is shown with its center of mass pointing upwards, and the shading indicates score contributions. Large clusters with high similarity among their elements occur only in exponentially rare configurations of the random vectors.

configurations with energy below $-S$. We evaluate this entropy in the thermodynamic limit, where both the number of random vectors and the dimension of the vector space are large. In this limit, the overlap of a data vector with a cluster center is a sum of many variables; the resulting thermodynamic potentials can then be expressed in terms of averages over Gaussian ensembles.

High-scoring clusters have to be found in each fixed configuration of the random data vectors, which act as quenched disorder for the statistics of clusterings. It turns out that the disorder generates correlations between the scores of clusters centered on different directions of the data vector space. These correlations, which become particularly significant in high-dimensional data sets, show that clustering is an intricate multiple-testing problem: spurious clusters may appear in many different directions of the data vectors. Here, we illustrate our results by a simple biophysical example: analysis of gene coexpression clusters. In this case, high-dimensional data vectors are generated by multiple measurements of a gene under different experimental conditions. The link between quenched disorder and multiple testing statistics is more generic than clustering, as discussed in the conclusion.

Distribution of data vectors and scoring.—We consider an ensemble of N vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, which are drawn independently from a distribution $P_0(\mathbf{x})$. We are specifically interested in data vectors with a large number of components, M. Clusters of such vectors are generically supported by multiple vector components, which is the source of the intricate cluster statistics discussed in this Letter. We assume that the distribution $P_0(x)$ factorizes in the vector components, $P_0(\mathbf{x}) = p_0(x_1) \ldots p_0(x_M)$ (this assumption can be relaxed; see below). Such null models are, of course, always simplifications, but they are useful for significance estimates in empirical data (an example is a p value of sequence alignments [8]).

A subset of these vectors forms a cluster. The clustered vectors are distinguished by their mutual similarity or, equivalently, their similarity to the center $\mathbf{z}$ of the cluster; see Fig. 1. We consider a simple similarity measure of vectors, the Euclidean scalar product: each vector $\mathbf{x}$ contributes a score

$$s(\mathbf{x}|\mathbf{z}, \mu) = \frac{1}{\sqrt{M}} \mathbf{x} \cdot \mathbf{z} - \mu. \qquad (1)$$

The scoring parameter μ acts as a threshold; vectors $\mathbf{x}$ with an insufficient overlap with the cluster center $\mathbf{z}$ result in a negative score contribution. The squared length of the cluster centers is normalized to $\mathbf{z} \cdot \mathbf{z} = M$.

A cluster can now be defined as a subset of positively scoring vectors. The cluster score is the sum of contributions from vectors in the cluster,

$$S(\mathbf{x}_1, \ldots, \mathbf{x}_N|\mathbf{z}, \mu) = \sum_{i=1}^{N} \max[s(\mathbf{x_i}|\mathbf{z}, \mu), 0]. \qquad (2)$$

Large values of μ result in clusters whose elements have a large overlap; small values result in looser clusters. The total score is determined both by the number of elements and by their similarities to the cluster center; that is, tighter clusters with fewer elements can have scores comparable to those of looser but larger clusters. Both the direction $\mathbf{z}$ and width parameter μ of clusters are *a priori* unknown.

Cluster score statistics.—To describe the statistics of an arbitrary cluster score $S(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ for vectors drawn independently from the distribution $P_0(\mathbf{x})$, we consider the partition function

$$Z(\beta) = \prod_{i=1}^{N} \int d\mathbf{x}_i P_0(\mathbf{x}_i) e^{\beta S(\mathbf{x}_1,\ldots,\mathbf{x}_N)} = \int dS p(S) e^{\beta S}. \qquad (3)$$

The second step collects all configurations of vectors $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ with a cluster score S, so $p(S)$ denotes the density of states as a function of score S. Asymptotically for large N, this density can be extracted from $Z(\beta)$ as

$$\log p(S) \simeq N\Omega(s) - \tfrac{1}{2}\log(gN). \qquad (4)$$

Here $\Omega(s)$ is the entropy as a function of the score per element, $s \equiv S/N$, which is the Legendre transform of the reduced free-energy density $\beta f(\beta) = -\log Z(\beta)/N$, i.e., $\Omega(s) = -\max_\beta[\beta f(\beta) + \beta s] \equiv -\beta^* f(\beta^*) - \beta^* s$. The prefactor g of the subleading term is given by $g = 2\pi|(\partial^2/\partial\beta^2)\beta f(\beta)|_{\beta=\beta^*}$. The p value of a cluster score S is defined as the probability $\int_S^\infty dS' p(S')$ to find a score larger or equal to S. Inserting (4) shows that this p value equals $p(S)$ up to a proportionality factor of order 1.

Clusters in a fixed direction.—As a first step, and to illustrate the generating function (3), we compute the distribution of scores for clusters with a fixed center $\mathbf{z}$. We assume that the null distribution p_0 for vector components has finite moments, we set the first two moments to 0 and 1 without loss of generality, and we choose $\mathbf{z}$ to lie in some direction which has nonzero overlap with a finite fraction of all M directions. Hence, the overlap $x_i \equiv \mathbf{x}_i \cdot \mathbf{z}$ is approximately Gaussian-distributed by the central limit theorem. The generating function (3) gives

$$-\beta f_c(\beta, \mu) = \log[(1 - H(\mu)) + e^{(\beta^2/2)-\beta\mu} H(\mu - \beta)], \qquad (5)$$

where the index c denotes the evaluation for a fixed cluster center and $H(x) = \int_x^\infty dx G(x)$ is the cumulative distribution function of the Gaussian $G(x) = \exp(-x^2/2)/\sqrt{2\pi}$. The result is an integral over the component $x \equiv \mathbf{x} \cdot \mathbf{z}$ of a data vector in the direction of the cluster center: below the score threshold μ, the component gives zero score, which contributes the cumulative distribution $\int_{-\infty}^{\mu} dx G(x)$ to the partition function. Above the score threshold, the component gives a positive score, which generates a contribution of $\int_\mu^\infty dx G(x) \exp\{\beta s(x|\mu)\}$. The resulting score distribution is given by (4), $\log p_c(S) = N\Omega(s = S/N) -(1/2)\log(g_c N)$; see Fig. 2(a).

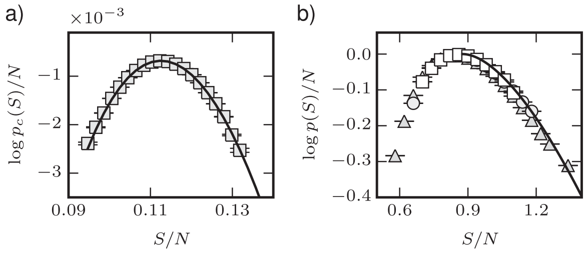

FIG. 2. Cluster score distributions in random data for fixed and optimal cluster directions. Analytical distributions $p(S)$ (solid lines) are plotted against the score per element, $s = S/N$, and are compared to normalized histograms obtained from numerical experiments with 10^6 samples (symbols). (a) Distribution $p_c(S)$ of the cluster score (2) for a fixed cluster center and data sets of $N = 6000$ vectors with $M = 70$, with the parameter $\mu = 0.1\sqrt{M}$. Error bars show the standard error due to the finite size of the sample. (b) Distribution of the maximum cluster score (6) with the parameter $\mu = 0.1\sqrt{M}$ for $N = 40$ (triangles), $N = 80$ (circles), and $N = 120$ (squares), keeping $M/N = 0.5$ fixed.

Maximal scoring clusters.—To gauge the statistical significance of high-scoring clusters in actual data sets, we need to know the distribution of the maximum cluster score in random data. The maximum cluster score is, in turn, implicitly related to the optimal cluster direction in a data set: for a given subset of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$, the maximal cluster score is reached if the direction of the center $\mathbf{z}$ coincides with the direction of the "center of mass," $\mathbf{x}_{\rm av} = (\mathbf{x}_1 + \ldots + \mathbf{x}_k)/k$. However, adding or removing vectors shifts the center of mass $\mathbf{x}_{\rm av}$ of the cluster and changes the score of each vector. Thus, finding the maximum score for a given data set,

$$S_{\max}(\mathbf{x}_1, \ldots, \mathbf{x}_N|\mu) = \max_{\mathbf{z}} S(\mathbf{x}_1, \ldots, \mathbf{x}_N|\mathbf{z}, \mu), \tag{6}$$

is a hard algorithmic problem, in particular, for large dimensions M. We calculate the distribution of $S_{\max}$ for independent random vectors from the generating function (3) with the integral representation

$$e^{\beta S_{\max}(\mathbf{x}_1,\ldots,\mathbf{x}_N|\mu)} = \lim_{\beta'\to\infty}\left[\int d\mathbf{z}\, e^{\beta' S(\mathbf{x}_1,\ldots,\mathbf{x}_N|\mathbf{z},\mu)}\right]^{\beta/\beta'} \tag{7}$$

for the statistical weight of a configuration $\mathbf{x}_1, \ldots, \mathbf{x}_N$. For large values of the auxiliary variable β', only directions $\mathbf{z}$ with a high cluster score $S(\mathbf{x}_1, \ldots, \mathbf{x}_N|\mathbf{z}, \mu)$ contribute to this integral over cluster directions $\mathbf{z}$, and the maximum over the cluster score (6) is reproduced in the limit $\beta' \to \infty$. We obtain

$$-\beta f(\beta, \mu) = \min_a\left[-\beta f_c\left(\beta, \mu - \frac{a}{2}\right) + \frac{M}{2N} \times \log\left(\frac{a+\beta}{a}\right)\right]. \tag{8}$$

This expression is to be understood in the asymptotic limit $N \to \infty$ with M/N kept fixed. The result (8) involves a variation over a, which, compared to the corresponding expression (5) for a fixed cluster center, generates an effective shift $a/2$ in the score cutoff μ and an additional entropy-like term. The calculation uses the so-called replica trick [9–11], representing the power $n = \beta/\beta'$ of the integral in (7) by a product of n copies (replicas). The calculation proceeds for integer values of n, and the limit $n \to 0$ ($\beta' \to \infty$) is taken by analytic continuation. A key ingredient is the average overlap $q = \langle \mathbf{z} \cdot \mathbf{z}' \rangle / M$ between directions of different cluster centers for the same configuration of data vectors at finite temperature $1/\beta'$. We find a unique ground state (i.e., $q \to 1$ for $\beta' \to \infty$) and a low-temperature expansion

$$q = 1 - \frac{a}{\beta'} + O\left(\frac{1}{\beta'^2}\right) \tag{9}$$

of the average overlap, similar to the case of directed polymers in a random potential [12], which arises in the statistics of sequence alignment [13]. Thus, the effect of center optimization on the free-energy density (8) and on cluster p values is related to the fluctuations between subleading cluster centers for the same random data set.

This solution determines the asymptotic form of the distribution of the maximum cluster score $S_{\max} = S$ as given by (4), $\log p(S) = N\Omega(s) + O(\log N)$. Figure 2(b) shows this result together with numerical simulations for several values of M and N, producing good agreement already for moderate N. According to (8), the effect of center optimization on score statistics increases with M and decreases with N. For small M/N, we expand the solution in N for fixed large M and obtain $-\beta f(\beta, \mu) = -\beta f_c(\beta, \mu) + (M/2N)\log N + \text{const}$, which leads to a distribution of maximum cluster scores,

$$\begin{aligned}\log p(S) &= \log p_c(S) + \frac{M}{2}\log N \\ &= N\Omega_c(s) + \frac{M-2}{2}\log N,\end{aligned} \tag{10}$$

up to terms of order N^0.

The presented calculation for the maximal cluster score distribution can be generalized to null distributions P_0 with arbitrary correlations between vector components $x^1, \ldots, x^M$ [14].

The free-energy density (8) was derived under the assumption of replica-symmetry (RS) [9], implying that only a single direction $\mathbf{z}$ yields the maximal score. This is appropriate for high-scoring clusters, since they occur in exponentially rare configurations of the random vectors, for which a second cluster direction with the same score would be even more unlikely. On the other hand, RS is known to be violated in the case $\beta = 0$, which describes clusters in typical configurations of the random vectors. This case has been studied before in the context of unsupervised learning in neural networks [10]. RS is also likely to be broken for $\beta < 0$, which describes configurations

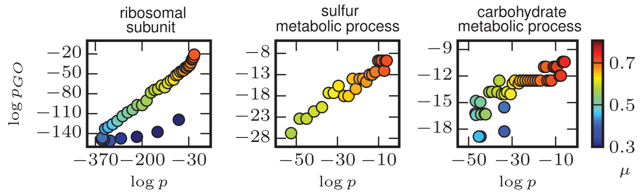

FIG. 3 (color). Statistical significance of clusters is correlated with functional annotation for yeast expression data. The diagrams show the significance p_{GO} of gene annotation terms vs the cluster score significance, traced over a range of the scoring parameter μ (shown by the color scale) of three representative clusters involved in translation (ribosomal genes), the sulfur metabolic process, and the carbohydrate metabolic process.

with score maxima biased towards values lower than in typical configurations. The limit $\beta \to -\infty$ is relevant to the problem of sphere packing in high dimensions, for which currently only loose bounds are known.

Application to clusters in gene expression data.—Clusters with high statistical significance may contain elements with a common mechanism causing their similarity. Here we test the link between our p value and the biological function of clusters in a data set of gene expression in yeast [15,16]. We trace several high-scoring clusters over the range of μ where they give a positive score. As μ increases, the cluster opening angle decreases (see Fig. 1), leading to a tighter, smaller cluster. The cluster p value also changes continuously, and the genes contained in the cluster also change. We ask if specific functional annotations [gene ontology (GO) terms] appear repeatedly in the genes of a cluster, and how likely it is for such a functional enrichment to arise by chance. We compute the p value $p_{\mathrm{GO}}(C)$ of the most significantly enriched GO term in a cluster C, using parent-child enrichment analysis [18] with a Bonferroni correction. A cluster with small $p_{\mathrm{GO}}(C)$ is thus significantly enriched in at least one GO annotation, which points to a functional relationship between its genes. As shown in Fig. 3, the parameter dependence of the cluster score significance $p[S(C)]$ and the significance $p_{\mathrm{GO}}(C)$ of gene annotation terms is strikingly similar. The statistical measure based on cluster score p values is thus a good predictor of functional coherence of its elements.

Conclusions.—We have established a link between quenched disorder physics and the multiple-testing statistics in clustering. This connection applies to a much broader class of problems, which involve the parallel testing of an exponentially large number of hypotheses on a single data set. Examples include imaging data (e.g. functional magnetic resonance imaging) and the analysis of next-generation sequencing data. If the scores of different hypotheses are correlated with each other, the distribution of the maximal score is not described by a known universality class of extreme value statistics. It may still be computable by the methods used here: the state space of the problem is the set of all hypotheses tested (here the centers of all clusters), and configurations of data vectors generated by a null model act as quenched random disorder.

Many thanks to M. Vingron for discussions and B. Nadler, M. Schulz, and E. Szczurek for their comments on the manuscript. This work was supported by Deutsche Forschungsgemeinschaft (DFG) grants GRK1360 (M. Łuksza), BE 2478/2-1 (J. Berg) and SFB 680.

[1] S. Still and W. Bialek, J. Neural. Comput. **16**, 2483 (2004).
[2] N. Slonim, G. S. S. Atwal, G. Tkačik, and W. Bialek, Proc. Natl. Acad. Sci. U.S.A. **102**, 18297 (2005).
[3] J. B. MacQueen, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press, Berkeley, 1967), Vol. 1, pp. 281–297.
[4] J. H. Ward, J. Am. Stat. Assoc. **58**, 236 (1963).
[5] B. J. Frey and D. Dueck, Science **315**, 972 (2007).
[6] M. Blatt, S. Wiseman, and E. Domany, Phys. Rev. Lett. **76**, 3251 (1996).
[7] R. Suzuki and H. Shimodaira, http://www.is.titech.ac.jp/~shimo/prog/pvclust/ (2009).
[8] S. Karlin and S. F. Altschul, Proc. Natl. Acad. Sci. U.S.A. **87**, 2264 (1990).
[9] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[10] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, England, 2001).
[11] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).
[12] D. A. Huse and C. L. Henley, Phys. Rev. Lett. **54**, 2708 (1985).
[13] T. Hwa and M. Lässig, Phys. Rev. Lett. **76**, 2591 (1996).
[14] M. Łuksza, M. Lässig, and J. Berg (to be published).
[15] A. P. Gasch *et al.*, Mol. Biol. Cell **11**, 4241 (2000).
[16] The data set contains expression levels from 173 samples for $N = 6152$ genes. Raw expression levels were log-transformed and mean-centered, first by gene (setting the average expression level of a gene to zero) and then by sample (setting the average expression level in a sample to zero). Since expression levels may be correlated across samples (for example, in successive expression levels of a time course), we perform a principal component analysis [17]. We restrict our analysis to the leading $M = 70$ eigenvectors of the yeast data set, which account for over 95% of the gene expression variance.
[17] I. T. Jolliffe, *Principal Component Analysis* (Springer, New York, 2002), 2nd ed.
[18] S. Grossmann, S. Bauer, P. N. Robinson, and M. Vingron, Bioinformatics **23**, 3024 (2007).