Significance-based clustering: a statistical mechanics approach

Marta Łuksza[1], Michael Lässig[2], and Johannes Berg[3]
[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*
[3]*Physikalisches Institut, Albert-Ludwigs-Universität Freiburg,*
Herrmann-Herder-Straße 3, 79104 Freiburg, Germany

(Dated: January 1, 1970)

Detecting sets of mutually similar elements in data is a core problem in data analysis. Applications from gene expression analysis to image recognition rely on clustering algorithms. Here, we consider the problem of clustering in random data: Given a set of randomly distributed vectors, how likely do some of them form a cluster with a given similarity among its elements? This *cluster p-value* is crucial to assess the statistical significance of clusters found in real data: a cluster with properties that rarely occur at random more likely involves a functional relationship between its elements. We use methods from the statistical mechanics of disordered systems to analytically solve the random clustering problem. In an application to gene expression data we find a remarkable link between the statistical significance of a cluster and the functional relationships between the genes it contains.

PACS numbers: 5.00.00 02.50.-r 07.05.Kf

Identifying groups of similar elements is a key problem in many areas of data analysis. An example is the detection of genes with similar expression levels measured over different conditions. While the aim of grouping objects into clusters is easily stated, clustering remains a challenge for both theory and practice.

Two ingredients are required to cluster data: a notion of similarity between the elements of the dataset, leading to a *scoring function* for clusters, and an *algorithmic procedure* to group elements into clusters. Diverse methods have been applied to both these aspects of clustering: similarities can be defined by Euclidean or by information-theoretic measures [1], and there are many different clustering algorithms from classical k-means [2] and hierarchical clustering [3], to recent message-passing techniques [4]. Moreover, even a single clustering procedure can produce different results, since the scoring function generally depends on free parameters. The most important scoring parameter weighs number versus size of clusters and is contained explicitly (e.g., the number k in k-means clustering) or implicitly (e.g., the temperature in superparamagnetic [5] and information-based clustering [1]) 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 dataset.

An important aspect is the *statistical significance of clusters*, which distinguishes "true" clusters from spurious clusters as would occur also in random data. An example is the starry sky: it shows true clusters such as galaxies with their stars bound to each other by gravity, but also constellations of stars which are in fact unrelated and may be far from one another. Even random datasets are unlikely to be completely devoid of clusters, which would require a statistically rare order. The problem of spurious clusters is exacerbated for high-dimensional data vectors, where clusters can lie in many different directions.

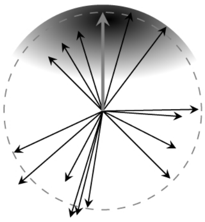

FIG. 1: **Clustering a set of random vectors.** In a set of randomly chosen vectors, subsets of vectors can arise whose elements share a large similarity among each other. Here a cluster is shown with its centre of mass pointing upwards and shading proportional to scores of cluster elements.

One thus has to distinguish between clusters that arise by chance, and statistically unexpected clusters which more likely involve a functional relationship between their elements. The *p-value of a cluster* quantifies this distinction: it is defined as the probability that a random data set contains a cluster with similarity score equal or higher than a given score S. The statistical significance of clusters is also crucial to address the problem of choosing scoring parameters: evaluating the significance of each cluster, we can base the scoring parameter choice on how unlikely the resulting clusters arise in random data. Cluster p-values can be estimated numerically by generating many random data sets and observing the frequency of a given score [6]. However, since this method is computationally intensive, clusterings are often reported without p-values.

These observations call for a statistical theory of clustering, which is the topic of this paper. Our aim is not to propose a new method for clustering, but to tell significant clusters from insignificant ones. The score of the maximal scoring cluster in a set of random vectors is a random variable, whose distribution we calculate. We use a mapping to a physical system, where the logarithm of the cluster p-value appears as an entropy. A problem of the statistical mechanics of disordered systems emerges: the direction of clusters is optimized for a fixed realisation of the random data vectors (quenched disorder).

Distribution of data vectors and scoring. In order to demonstrate our method on a concrete example, we consider an ensemble of N vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ in an M-dimensional space. Each vector is drawn independently from a distribution $P_0(\mathbf{x})$, with

$$P_0(\mathbf{x}) = \prod_{\nu=1}^{M} P(x^\nu) \ . \tag{1}$$

Correlations between vectors, or between their components x^ν can also be treated, but are not considered here. We assume that all moments of the distribution $P(x)$ exist, and without loss of generality take the first two moments to be zero and one, respectively.

Within a set of vectors $\{\mathbf{x_i}\}, i = 1, \ldots, N$, a subset of vectors may form a cluster. These vectors are distinguished by their mutual similarity, or equivalently, their similarity to the centre $\mathbf{z}$ of the cluster, see Fig. 1. We restrict the discussion to a simple similarity measure of vectors, the Euclidean scalar product: each vector $\mathbf{x}$ contributes a score

$$s(\mathbf{x}|\mathbf{z}, \mu) = \frac{1}{\sqrt{M}} \mathbf{x} \cdot \mathbf{z} - \mu \ . \tag{2}$$

The scoring parameter μ acts as a threshold; vectors $\mathbf{x}$ with an insufficient overlap with the cluster centre $\mathbf{z}$ result in a negative score contribution. The squared length of $\mathbf{z}$ is chosen to be M throughout, the same as that of typical vectors $\mathbf{x}$ from the ensemble (1).

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,

$$S(\mathbf{x}_1, \ldots, \mathbf{x}_N|\mathbf{z}, \mu) = \sum_{i=1}^{N} \max\left[s(\mathbf{x_i}|\mathbf{z}, \mu), 0\right] \ . \tag{3}$$

Large values of μ result in clusters whose elements have a large overlap, small values result in more loose clusters. The total score is determined both by the number of elements and by their similarities with the cluster centre, 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 unknown a priori.

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

$$\begin{aligned} Z(\beta) &= \prod_{i=1}^{N} \int d\mathbf{x}_i \, P_0(\mathbf{x}_i) \, e^{\beta S(\mathbf{x}_1, \ldots, \mathbf{x}_N)} \\ &= \int dS \, p(S) \, e^{\beta S} \ . \end{aligned} \tag{4}$$

The second step collects all configurations of vectors with the same cluster score S, so $p(S)$ denotes the probability density of S. In the language of statistical physics, $-S$ is the energy of a state $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ of the system, $f(\beta) \equiv -\log Z(\beta)/(\beta N)$ is the free energy density, and $p(S)$ is the density of states as a function of the energy. Asymptotically for large N, this density can be extracted from $Z(\beta)$ by

$$\log p(S) \simeq N\Omega(s) - \frac{1}{2} \log(gN) \ , \tag{5}$$

where $\Omega(s)$ is the *entropy* as a function of the score per element, $s \equiv S/N$ and is given by the Legendre transform of $\beta f(\beta)$, i.e. $\Omega(s) = -\max_\beta[\beta f(\beta) + \beta s]$. The prefactor g of the subleading term can be evaluated in terms of the second derivative of the free energy density.

The p-value of a cluster score S is defined as the probability to find a score larger or equal to S, which is given by the integral $\int_S^\infty dS' \, p(S')$. Inserting (5) shows that this log-p-value equals $\log p(S)$ up to a term of order $\log N$, which we neglect in the following.

Clusters in a fixed direction. We first illustrate the method on a toy problem and compute the distribution of scores for clusters with a fixed centre $\mathbf{z}$. We choose $\mathbf{z}$ to lie some direction which has non-zero overlap with a finite fraction of all M directions, so the overlap $x_i \equiv \mathbf{x}_i \cdot \mathbf{z}$ is Gaussian-distributed by the law of large numbers. The generating function (4) is then straightforward to evaluate, giving

$$-\beta f_c(\beta, \mu) = \log\left[(1 - H(\mu)) + e^{\frac{\beta^2}{2} - \beta\mu} H(\mu - \beta)\right] , \tag{6}$$

where the index c denotes evaluation for a fixed cluster centre and $H(x) = \int_x^\infty G(x)$ 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 centre: 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 (5), $\log p_c(S) = N\Omega\,(s = S/N) - (1/2)\log(g_c N)$ including sub-leading terms with $g_c = 2\pi|\frac{\partial^2}{\partial^2 \beta} \beta f(\beta)|$ evaluated at the maximum of $\beta f(\beta) + \beta s$ over β. The result is plotted in Fig. 2(a) together with

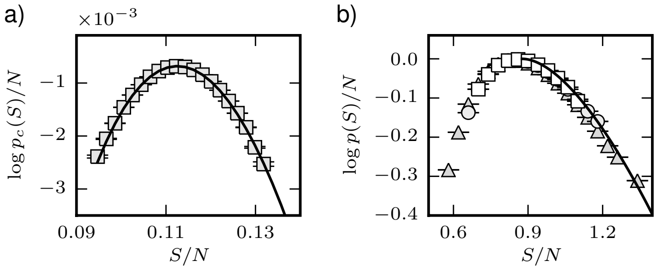

FIG. 2: **Cluster score distributions in random data for fixed and optimal cluster direction.** Analytical distributions $p(S)$ (solid lines) are plotted against the score per element, $s = S/N$, and are compared to normalized histograms obtained from numerical experiments with 10^6 samples (symbols). (a) Distribution $p_c(S)$ of the cluster score (3) for fixed cluster centre and datasets of $N = 6000$ vectors with $M = 70$, with 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 (7) with parameter $\mu = 0.1\sqrt{M}$ for $N = 40$ (triangles), $N = 80$ (circles) and $N = 120$ (squares), keeping $M/N = 0.5$ fixed. The analytical solution is valid asymptotically for large N, but good agreement with the numerics is seen already for moderate values of N.

a score distribution obtained from simulations of randomly generated data vectors, showing excellent agreement. The leading asymptotics of $p_c(S)$ can also be derived using large deviation statistics [7].

Maximal scoring clusters. To gauge the statistical significance of high-scoring clusters in actual datasets we need to know the distribution of the *maximum cluster score* in data. For a given subset of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$, the maximal cluster score is reached if the centre $\mathbf{z}$ is in the direction of the "centre of mass", $\mathbf{x}_{\text{av}} = (\mathbf{x}_1 + \ldots + \mathbf{x}_k)/k$. However, in the search for the maximal scoring cluster, we may add new vectors to a cluster or remove vectors from it. Each of these steps shifts the centre of mass $\mathbf{x}_{\text{av}}$ of the cluster and changes the score of each vector. In random datasets, competing clusters may appear in any combination of the M components of the data vectors. This makes the search for the maximal scoring cluster a hard statistical and algorithmic problem, in particular for large M: to locate the cluster with maximum score for a given set of vectors, one has to probe all possible cluster centre directions $\mathbf{z}$.

We now calculate the distribution of the maximum cluster score

$$S_{\max}(\mathbf{x}_1, \ldots, \mathbf{x}_N|\mu) = \max_{\mathbf{z}} S(\mathbf{x}_1, \ldots, \mathbf{x}_N|\mathbf{z}, \mu) \qquad (7)$$

for vectors chosen independently from the distribution $P_0(\mathbf{x})$. To evaluate the generating function (4), we use the integral representation

$$e^{\beta S_{\max}(\mathbf{x}_1,\ldots,\mathbf{x}_N|\mu)} = \lim_{\beta'\to\infty} \left[\int d\mathbf{z}\, e^{\beta' S(\mathbf{x}_1,\ldots,\mathbf{x}_N|\mathbf{z},\mu)} \right]^{\beta/\beta'} \qquad (8)$$

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, and the maximum over the cluster score (7) is reproduced in the limit $\beta/\beta' \to 0$. We perform this limit using techniques from the statistical mechanics of disordered systems [8–10], obtaining

$$-\beta f(\beta, \mu) = \min_a \left[-\beta f_c\left(\beta, \mu - \frac{a}{2}\right) + \frac{M}{2N} \log\left(\frac{a+\beta}{a}\right) \right] . \qquad (9)$$

This expression is to be understood in the asymptotic limit $N \to \infty$ with M/N kept fixed. The calculation uses the so-called replica-trick [8], representing the n-th power ($n = \beta/\beta'$) in (8) by n copies (replicas) of the integral over $\mathbf{z}$. The calculation proceeds for integer values of n, and the limit $n \to 0$ ($\beta' \to \infty$) is taken by analytic continuation. For finite values of β', different replicas $\mathbf{z}$ and $\mathbf{z}'$ have a finite overlap $q = \frac{1}{M}\sum_{\nu=1}^{M} z^\nu z'^\nu$. This overlap quantifies the thermal fluctuations of the cluster direction at fixed vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$. The volume of these fluctuations corresponds to the different alternative cluster centres at submaximal score. In the limit of large β', the volume of these fluctuations vanishes and $q = 1 - a/\beta'$ to leading order in β'. The result (9) involves a variation over a, which, compared to the corresponding expression (6) for fixed cluster centre, includes in an effective shift $a/2$ in the score cutoff μ and an additional entropy-like term. This solution determines the asymptotic form of the distribution of maximum cluster score $S_{\max} = S$ as given by (5), $\log p(S) = N\Omega(s) + O(\log N)$. This is plotted in Fig. 2(b) together with numerical simulations for several values of M and N, showing good agreement already for moderate values of N. According to (9), the effect of centre optimization on the score statistics increases with the number of data components, M, and decreases with the size of the dataset, N. For small values of M/N, we can expand the solution to leading order 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 given by

$$\log p(S) = \log p_c(S) + \frac{M}{2}\log N = N\Omega_c(s) + \frac{M-2}{2}\log N \qquad (10)$$

up to terms of order N^0. This expansion is appropriate for the situation $M \ll N$ frequently encountered in gene expression data (see below), where many expression levels are measured in comparatively few experimental conditions.

The free energy density (9) was derived using the replica-trick [8] under the assumption of replica-symmetry (RS), implying that only a single direction $\mathbf{z}$ yields the maximal score. This is appropriate for high-scoring clusters, which occur in *exponentially rare instances* of the random vectors, and a second cluster direction with the same score would be even more unlikely.

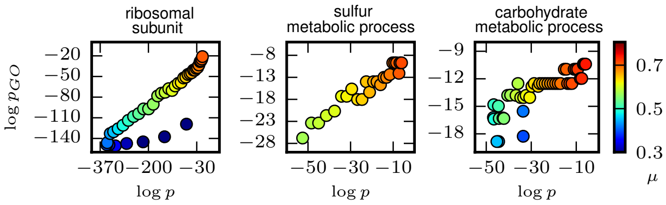

FIG. 3: **Statistical significance of clusters correlates with functional annotation for yeast expression data**. The significance p_{GO} of gene annotation terms vs. the cluster score significance, traced over a range of scoring parameter μ (shown by color-scale) of three representative clusters involved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

RS is known not to hold, however, in the special case $\beta = 0$, which describes clusters in *typical instances* of the random vectors. The case $\beta = 0$ has been studied before in the context of unsupervised learning in neural networks [9]. RS is also likely to be broken for $\beta < 0$, which describes instances with lower score maxima than those found in typical instances. The limit $\beta \to -\infty$ could be used to address the problem of sphere packing in high dimensions, for which currently only loose bounds are known.

Application to clusters in gene expression data. We expect that clusters which are unlikely to arise by chance consist of elements sharing some functional relationship. In clusters of stars referred to above, the relationship is gravitational influence, in gene expression clusters it is shared biochemical pathways or another functional property. If this expectation were borne out, the cluster significance calculated from (9) could be used to detect those clusters where a high degree of functional coherence is expected.

We test the link between cluster significance and functional relationships between cluster elements in yeast gene expression data [11] [15]. 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 cluster with a smaller number of elements. 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 [12] 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 thus is a good predictor of functional coherence of its elements.

These results open the way to a significance-based approach to clustering [13], where clusters are selected such that their statistical significance is maximal. Instead of detecting clusters on the basis of a single, global scoring parameter, each cluster can then be optimized individually for its statistical significance.

Many thanks to Martin Vingron for discussons and Boaz Nadler, Marcel Schulz and Ewa Szczurek for critically reading the manuscript. This work was supported by Deutsche Forschungsgemeinschaft (DFG) grants GRK1360 (M. Łuksza), BE 2478/2-1 (J. Berg) and SFB 680.

[1] Slonim, N., Atwal, G. S. S., Tkačik, G. & Bialek, W. (2005). Proc Natl Acad Sci U S A 102, 18297–18302.

[2] MacQueen, J. (1967). Proc 5th Berkeley Symp Math Stat Probab 1, 281–197.

[3] Ward, J. H. (1963). J Am Stat Assoc 58, 236–244.

[4] Frey, B. J. & Dueck, D. (2007). Science 315, 972–976.

[5] Blatt, M., Wiseman, S. & Domany, E. (1996). Phys. Rev. Lett 76, 3251–3254.

[6] Suzuki, R. & Shimodaira, H. (2009). http://www.is.titech.ac.jp/ shimo/prog/pvclust/

[7] Cover, T. M. & Thomas, J. A. (1991). Elements of Information Theory. John Wiley, New York.

[8] Mézard, M., Parisi, G. & Virasoro, M. A. (1987). Spin Glass Theory and Beyond. World Scientific, Singapore.

[9] Engel, A. (2001). Theor Comput Sci 265, 285–306.

[10] Gardner, E. & Derrida, B. (1988). J Phys A: Math Gen 21, 271–284.

[11] 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.

[12] Grossmann, S., Bauer, S., Robinson, P. N. & Vingron, M. (2007). Bioinformatics 23, 3024–3031.

[13] Łuksza, M., Lässig, M. & Berg, J. (2009). Significance-based clustering of gene-expression data. submitted.

[14] Jolliffe, I. T. (2002). Principal Component Analysis. Second edition, Springer, New York.

[15] The dataset contains expression levels from 173 experiments (different environmental conditions and time-course data) 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 experiment (setting the average expression level in an experiment over all genes to zero). Since expression levels may be correlated across experiments (for example, in successive expression levels of a time course), we perform a principal component analysis [14]. We restrict our analysis to the leading $M = 70$ eigenvectors of the the yeast dataset, which account for over 95% of the gene expression variance.