Potts Ferromagnets on Coexpressed Gene Networks: Identifying Maximally Stable Partitions

Himanshu Agrawal* and Eytan Domany†

Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel

(Received 2 December 2002; published 17 April 2003)

Clustering gene expression data by exploiting phase transitions in granular ferromagnets requires transforming the data to a granular substrate. We present a method using the recently introduced homogeneity order parameter Λ [H. Agrawal, Phys. Rev. Lett. **89**, 268702 (2002)] for optimizing the parameter controlling the "granularity" and thus the stability of partitions. The model substrates obtained for gene expression data have a highly granular structure. We explore properties of phase transition in high q ferromagnetic Potts models on these substrates and show that the maximum of the width of superparamagnetic domain, corresponding to maximally stable partitions, coincides with the minimum of Λ.

DOI: 10.1103/PhysRevLett.90.158102 PACS numbers: 89.75.Hc, 02.70.Uu, 05.50.+q, 89.75.Fb

Recent developments in gene chip technology [1,2] have revolutionized biomedical sciences. These chips are being used extensively for studying molecular aspects of diseases, especially of various cancers [3,4]. They are used for simultaneous measurement of the expression levels of large numbers of genes. The quantities measured in an expression profiling experiment are light intensity ratios (or differences) that reflect the expression levels of genes in the tissue sample that is used. A typical experiment using 20–100 tissue samples generates a few hundred thousand to a million data points corresponding to the expression levels of 10 000 or more genes in each sample.

Extraction of meaningful information from gene expression data is a complex task because of the large volume of the data and the expected complexity of its structure and organization. To address this problem in an unbiased way several specialized methods, broadly known as "clustering techniques," have been developed during the past few years [5–7]. All these techniques, though differing in details, in essence try to identify genes (or samples) that behave similarly across the samples (or genes) and classify them as belonging to one group or cluster. Among the clustering techniques that employ the concepts of physics, the one that succeeds in correctly clustering most of the types of data [8] is the "superparamagnetic" clustering method [6]. This method exploits the properties of phase transitions in disordered Potts ferromagnets.

Ferromagnetic Potts systems have been studied extensively in the past [9]. The mapping of raw gene expression data to a Potts ferromagnet requires extensive chip dependent processing and filtering for selecting genes having high information content. Details of these procedures can be found with sources of the data [3,4]. The preprocessing gives an expression matrix with N rows and D columns. Each row represents a gene and each column a sample. The entries of this matrix are $\log_2$-transformed expression levels. Its rows are centered and normalized to have a mean of zero and standard deviation of unity.

For constructing the Potts ferromagnet each of the N genes is viewed as a point in D-dimensional metric space with its coordinates given by its expression levels in the D samples. A Potts spin s_i is assigned to each point i and neighbors of each of the points are identified (as described later). s_i takes q different integer values ranging from 1 to q. The spins at neighboring points i and j interact via a short range ferromagnetic coupling $J_{\langle i,j\rangle}$,

$$J_{\langle i,j\rangle} = \frac{1}{\hat{K}} \exp\left\{-\frac{1}{2}\left(\frac{d_{\langle i,j\rangle}}{\bar{d}}\right)^2\right\}, \tag{1}$$

where the angular brackets denote neighbor pairs, $d_{\langle i,j\rangle}$ is the Euclidean distance between points i and j, $\bar{d}$ is the mean distance between interacting neighbors, and $\hat{K}$ is the average number of interacting neighbors of a point [6]. The interaction between spins that are not neighbors is set to zero. The Hamiltonian of the system is given by

$$\mathcal{H}(S) = -\sum_{\langle i,j\rangle} J_{\langle i,j\rangle}\delta_{s_i,s_j}, \tag{2}$$

where the summation is over interacting neighbors, $S \equiv \{s_1, \ldots, s_N\}$ is the state of the system, and the delta function $\delta_{s_i,s_j} = 1$ if $s_i = s_j$ and zero otherwise.

Disordered ferromagnetic Potts systems have been studied using Monte Carlo simulations [10] and their mean-field behavior is well understood, especially in the context of the clustering problem [6]. Such a system may have three different phases, viz., ferromagnetic, paramagnetic, and superparamagnetic, depending on the temperature and interactions. The system is ferromagnetic at low temperatures T and paramagnetic at high. On increasing the temperature from zero, the system passes from a ferromagnetic to a paramagnetic state either directly in a single transition or via an intermediate superparamagnetic phase through two transitions. In both cases the transitions are first order. In the second case a transition from a ferromagnetic to a superparamagnetic phase occurs at a temperature $T_{\rm fs}$ which is followed, at a higher temperature $T_{\rm sp}$, by a transition from the superparamagnetic to the paramagnetic phase.

This path is of considerable interest in the study of disordered systems, especially in the context of data clustering as clusters of aligned spins automatically divide the data into its natural classes and a clear hierarchical structure among the classes emerges on varying the temperature.

The path taken by the phase transition depends on the nature of the interactions and the definition of interacting neighbors. Unlike regular lattices, disordered systems do not have an intrinsic definition of neighbors. Some "reasonable" definition must be imposed externally. The superparamagnetic clustering method uses the K mutual neighbor criterion with some heuristic modifications [6]. In the K mutual neighbor criterion, for an *a priori* selected integer K, two points i and j are defined to be neighbors if and only if j is among the closest K in the distance-ordered list of nearest neighbors of i, and vice versa. If the resulting edge set does not place all points into one connected component, the edges of a minimum spanning tree, a tree spanning all points and having minimum total edge length, are added to it. A single connected component is required to ensure that the ground state is ferromagnetic.

The procedure described above yields $N-1$ different sets of definitions of neighbors, each corresponding to one possible value of K, $K \in [1, N-1]$. Each one of these sets can be viewed as a graph $G(K)$ with edge couplings given by Eq. (1). In the Potts models on each of these graphs the path taken by the phase transition is different with differing transition temperatures and compositions of correlated "ferromagnetic grains." We define $\hat{G}(K)$ as the graph obtained by only K mutual neighbor criterion.

For very small K, say $K = 1$ or 2, each point has a very small number of mutual neighbors and the system forms a single connected component only through bonds generated by the minimal spanning tree. These constitute a significant fraction of the bonds in the system and nontrivially alter its thermal behavior. Thus, in this case the path taken by the system on increasing the temperature and the range of temperatures $\Delta T = T_{\rm sp} - T_{\rm fs}$ in which superparamagnetic phase occurs are effects of superimposing the minimal spanning tree. As K is increased this effect rapidly decreases and becomes vanishingly small at K near $z_{\rm max}^{\rm mst}$, the largest degree in the minimum spanning tree. For large K each vertex has many neighbors and for $K = N-1$ all the vertices are mutually connected. In these cases the system behaves as an ordered lattice with weak bond strength disorder and goes in a single transition from a fully ordered (single cluster) to a fully disordered state (N uncorrelated spins). Thus, for large K there is either no superparamagnetic phase or a very small one.

At some intermediate values of K a phase transition through a sizable superparamagnetic phase, i.e., one with a large ΔT, may occur. If the system has high granularity (i.e., is highly inhomogeneous) a superparamagnetic phase ($\Delta T > 0$) is expected to occur for a wide range $[K', K'']$ of values of K. It is desirable to identify that subrange $[K_1, K_2] \subset [K', K'']$ for which ΔT is the largest. In this subrange the granularity of the data, if any, is maximally enhanced. Alternatively, in this subrange the clusters are expected to be well separated with graphs consisting of loosely connected "elements."

Quantitative characterization of inhomogeneity in these systems, that allows comparison for different K, is a nontrivial task. This is because these systems are inhomogeneous at two levels, viz., (i) bond strength inhomogeneity and (ii) inhomogeneity of the number of neighbors of each spin. Note, however, that on average the strength of bonds added on going from $G(K)$ to $G(K+1)$ does not differ significantly from that of bonds added on going from $G(K-1)$ to $G(K)$. As a result, the difference in the inhomogeneity of systems corresponding to neighboring values of K is largely due to topological variation. Thus, for comparing the inhomogeneity of the systems the effect of bond strength inhomogeneity can be ignored and all couplings $J_{\langle i,j\rangle}$ can be set to the same strength. This equates the total inhomogeneity to the inhomogeneity of the number of bonds. It is equivalent to the averaging over bond strength disorder.

A quantitative measure of topological inhomogeneity has recently been introduced by Agrawal [11] using a "homogeneity order parameter" Λ determined as follows. In a graph $H(K)$, measure the probability $F(z)$ that a vertex has *at least* z neighbors. If z varies in the range $[z_{\rm min}, z_{\rm max}]$, $F(z)$ decreases monotonically from unity for $z \le z_{\rm min}$ to zero for $z > z_{\rm max}$. Let $c = 1 + z$ be the size of the "droplet" formed by the vertex and its neighbors and $\tilde{F}(c)$ be the corresponding distribution, $\tilde{F}(c) = F(z)$. The parameter Λ is defined [11] as the area enclosed below the $\tilde{F}(c)$ curve between zero and $c_{\rm max}$ normalized by $c_{\rm max}$. It is readily calculated using the trapezoidal rule and equals

$$\Lambda(K) = \frac{1 + \bar{z} + [1 - P(z_{\rm max})]/2}{1 + z_{\rm max}}, \tag{3}$$

where $\bar{z}$ is the mean connectivity and $P(z_{\rm max})$ is the probability of finding a vertex with $z_{\rm max}$ neighbors.

Since we are interested in comparing the inhomogeneity exhibited by $N-1$ graphs $G(K)$ or $\hat{G}(K)$, we define the relative version of the homogeneity order parameter as

$$\Lambda^\star(K) = [\Lambda(K) - \Lambda_{\rm min}]/[1 - \Lambda_{\rm min}], \tag{4}$$

where $\Lambda_{\rm min} = \min[\Lambda(1), \ldots, \Lambda(N-1)]$. For a homogeneous graph (one having $z_{\rm min} = z_{\rm max}$) we have $\Lambda^\star = 1$. The more spread out is the distribution of the number of neighbors, the smaller is the value of $\Lambda^\star$.

Figure 1 shows the variation of the relative homogeneity $\Lambda^\star(K)$ of the interaction neighborhood $\hat{G}(K)$ for different values of K, obtained using an expression matrix having 1270 genes and 36 samples taken from colon cancer data [3]. Similar behavior was observed in other expression data sets also [11]. Here we use $\hat{G}(K)$ because

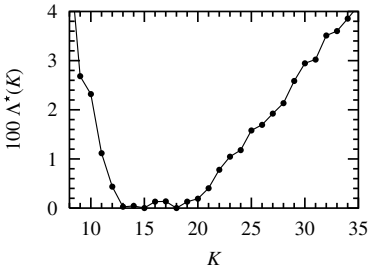

FIG. 1. Variation of $\Lambda^\star(K)$ with K for graphs $\hat{G}(K)$ constructed from colon cancer data (see the text for details).

the effect of the minimum spanning tree, in clustering, becomes negligible at $K = K'$ near $z_{\max}^{\text{mst}}$, equating the total inhomogeneity of $G(K)$ and $\hat{G}(K)$ for $K > K'$. The topological inhomogeneity of the minimum spanning tree, however, significantly influences that of $G(K)$ and, in clustering, $\Lambda(K)$ of $G(K)$ does not represent its total inhomogeneity for K considerably larger than K'. Figure 1 clearly shows a *flat* minimum of $\Lambda^\star(K)$ in the range $K_1 = 13 \leq K \leq 20 = K_2$ corresponding to maximally inhomogeneous interaction neighborhoods. In this range of K the graphs $\hat{G}(K)$ were earlier observed to have small-world and scale-free structure [11]. The same topological characteristics were observed for both $G(K)$ and $\hat{G}(K)$ in the present case also. Furthermore, the presence of a *flat* minimum extending over a range of values of K, as opposed to a sharp one at a specific value of K, is an extremely significant and important feature in the context of clustering: it implies that the data have natural partitions that are stable against small perturbations. The figure shows that $\Lambda^\star(K)$ is not a very smooth function of K. This is an effect of finite size (small N). As a result, the range of K housing the minimum of $\Lambda^\star(K)$ must be estimated by taking the fluctuations into account. Usually this range occurs for $\Lambda^\star(K) \lesssim 0.01$.

The following are the supporting simulation results. We used the expression matrix obtained from colon cancer data described earlier [3]. Using this matrix we constructed disordered Potts ferromagnets for $K \in [1, 100]$ with interactions given by Eq. (1) between neighboring spins. We used $q = 20$ spin states because for large q the results were shown to be largely independent of q [6]. We carried out extensive Monte Carlo simulations of these systems using the Swendsen-Wang algorithm [12]. For each K we sampled thermal variation of energy per spin e, magnetization m, and their fluctuations, with

$$m = \frac{(N_{\max}/N)q - 1}{q - 1}, \tag{5}$$

where $N_{\max} = \max[N_1, \ldots, N_q]$, and N_a is the number of spins with value a.

For each K, T_{fs} was taken as the temperature at which susceptibility shows a prominent peak. Similarly, the temperature at which susceptibility decays sharply to almost zero is the ideal value of T_{sp}. This point, however, cannot be identified from the data for most values of K because around this point susceptibility does not have a sharp variation. However, we noted that fluctuations in energy have a prominent peak at this temperature and used it to identify T_{sp} (with supporting decay observed in susceptibility). Since the system is finite (i.e., the number of genes is fixed), an analysis leading to better estimates of transition temperatures is not possible. The thermal variation of fluctuations in magnetization and energy for several different values of K is shown in Fig. 2. The figure clearly shows the smearing of transition points due to large fluctuations in both quantities. A significant observation from this figure, however, is that smearing of the (ideal) sharp decay of susceptibility marking T_{sp} is compensated by a prominent peak of the specific heat at the corresponding location. The reverse is the case for T_{fs}.

The variation of transition temperatures T_{fs} and T_{sp} with K is shown in the K-T phase diagram in Fig. . The figure clearly shows that T_{fs} increases very rapidly on increasing K for small values of K and then seems to saturate for large K. On the other hand, T_{sp} initially increases rapidly on increasing K, achieves a peak, stays at it for some K, and then starts decreasing as K is increased further. For large values of K, T_{sp} also seems

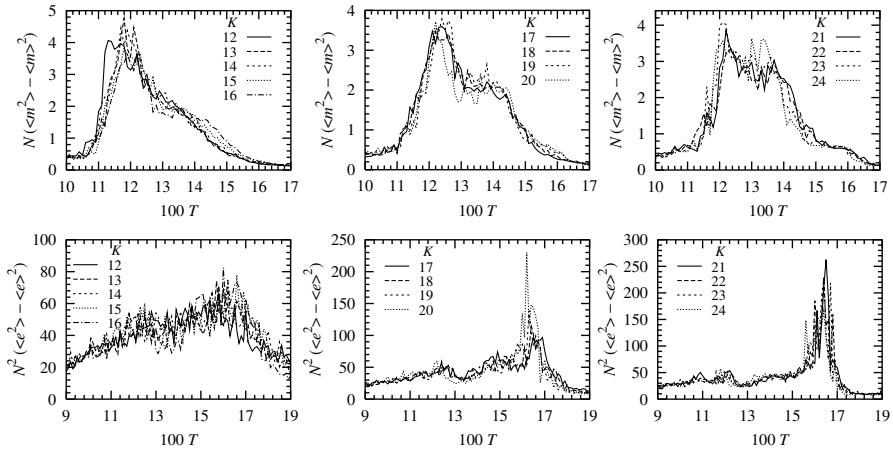

FIG. 2. Variation of susceptibility $\chi T = \langle m^2 \rangle - \langle m \rangle^2$ (top row) and specific heat $C_v T^2 = \langle e^2 \rangle - \langle e \rangle^2$ (bottom row) with temperature T for different values of K in the range $[K_a, K_b]$ corresponding to a maximum of ΔT in Potts ferromagnets constructed from colon cancer data (see details in text). The K range has been decomposed in three to show different patterns of behavior.

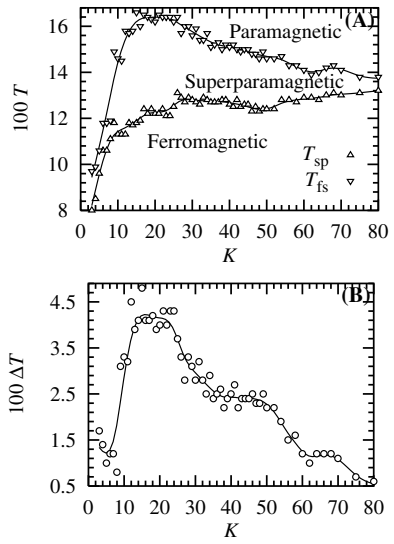

FIG. 3. (a) The K-T phase diagram for $q = 20$ state Potts ferromagnets constructed from colon cancer data (see details in text). (b) Variation of the width ΔT of the superparamagnetic domain with K. Solid lines are natural smoothing splines.

to plateau off, at the same value at which T_{fs} saturates. The saturation is a consequence of the form of interaction and also of the fact that the mean-field solution is exact for large K.

Variation of the width of the superparamagnetic phase in the K-T phase diagram is shown in Fig. 3(a). Although the data are very noisy, the overall trend is quite clear from the figure. For small K the width decreases sharply and attains a minimum on increasing K. This is a consequence of superimposing the minimum spanning tree which contributes a significant number of bonds to the system in this range of K. The effect of the minimum spanning tree becomes negligible after $K \gtrsim 8$, which is the location of the minimum of ΔT. As K is increased further, ΔT increases rapidly and attains a peak which seems to be flat and persists for a small range of values of K. On increasing K further, ΔT decays and for large K appears to go to zero. For $K = N - 1$, ΔT is expected to be nearly zero because total interaction of each point is nearly the mean-field value.

Figure 3(a) shows that the maximum of ΔT occurs in the range $K_a \approx 12 \leq K \leq 24 \approx K_b$. This range maps, with slightly wider ends, on the range $K_1 = 13 \leq K \leq 20 = K_2$ corresponding to the minimum of $\Lambda^\star(K)$. $[K_a, K_b]$ is the superset of $[K_1, K_2]$ because while $[K_1, K_2]$ incorporates full averaging over bond strength disorder, $[K_a, K_b]$ does not. This is because disorder averaging is not relevant for clustering.

The results presented above lead to a few important conclusions concerning disordered Potts ferromagnets as well as clustering of gene expression data. First, they show that in a given set of disordered ferromagnetic Potts systems, the superparamagnetic phase occurs for the widest range of temperatures in those that have maximum inhomogeneity (or granularity). The corresponding graphs are also the optimal ones for clustering. Second, the effect of the minimum spanning tree on clustering vanishes at $K = K'$ near $z_{\max}^{\text{mst}}$. As superparamagnetic clustering is done at $K > K'$, its solutions are different from those obtained using methods based primarily on the minimum spanning tree. Third, the range of temperature in which the superparamagnetic phase is observed depends primarily on the overall topological inhomogeneity of the interaction neighborhood provided that the interactions are short ranged. The next conclusion is a consequence, specifically, of the presence of a *flat* minimum in a range of K in the relative order parameter. It implies that the clustering solutions obtained by superparamagnetic clustering are robust against noise inherent in the data.

H. A. would like to thank Sven Lübeck for useful discussions. This work was supported by the Israel Science Foundation, the Germany-Israel Science Foundation, the Minerva Foundation, and NIH Grant No. 5 P01 CA 65930-06.

*Electronic address: feagrawa@wicc.weizmann.ac.il
†Electronic address: fedomany@wicc.weizmann.ac.il

[1] E. S. Lander, Nat. Genet. **21**, 3 (1999).
[2] D. Gerhold, T. Rushmore, and C. T. Caskey, Trends Biochem. Sci. **24**, 168 (1999).
[3] D. A. Notterman, U. Alon, A. J. Sierk, and A. J. Levine, Cancer Res. **61**, 3124 (2001).
[4] A. A. Alizadeh *et al.*, Nature (London) **403**, 506 (2000); S. M. Dhanasekaran *et al.*, *ibid.* **412**, 822 (2001); A. Bhattacharjee *et al.*, Proc. Natl. Acad. Sci. U.S.A. **98**, 13 790 (2001); D. G. Beer *et al.*, Nature Med. (NY) **8**, 816 (2002).
[5] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, Proc. Natl. Acad. Sci. U.S.A. **95**, 14863 (1998).
[6] M. Blatt, S. Wiseman, and E. Domany, Phys. Rev. Lett. **76**, 3251 (1996); Neur. Compt. **9**, 1805 (1997); S. Wiseman, M. Blatt, and E. Domany, Phys. Rev. E **57**, 3767 (1998).
[7] A. Brazma and J. Vilo, FEBS Lett. **480**, 17 (2000).
[8] E. Domany, physics/0206056.
[9] F. Y. Wu, Rev. Mod. Phys. **54**, 235 (1982); Erratum, **55**, 315 (1983).
[10] M. S. S. Challa, D. P. Landau, and K. Binder, Phys. Rev. B **34**, 1841 (1986); S. Wiseman and E. Domany, Phys. Rev. E **51**, 3074 (1995); S. Chen, A. M. Ferrenberg, and D. P. Landau, Phys. Rev. E **52**, 1377 (1995).
[11] H. Agrawal, Phys. Rev. Lett. **89**, 268702 (2002).
[12] S. Wang and R. H. Swendsen, Physica (Amsterdam) **167A**, 565 (1990).