A Letters Journal Exploring the Frontiers of Physics

August 2007

EPL, **79** (2007) 38006 www.epljournal.org
doi: 10.1209/0295-5075/79/38006

1-D random landscapes and non-random data series

T. M. A. Fink[1,2(a)], K. Willbrand[3] and F. C. S. Brown[4]

[1] *Systems Biology and CNRS UMR 144, Institut Curie Paris 75005, France*
[2] *Theory of Condensed Matter, Cavendish Laboratory - Cambridge CB3 0HE, UK*
[3] *Laboratoire de Physique Statistique, Ecole Normale Supérieure - 75005 Paris, France*
[4] *Departement de Mathematique, Ecole Normale Supérieure - 75005 Paris, France*

received 4 May 2007; accepted in final form 11 June 2007
published online 17 July 2007

PACS 89.70.+c – Information theory and communication theory
PACS 87.14.Gg – DNA, RNA
PACS 89.75.-k – Complex systems

Abstract – We study the simplest random landscape, the curve formed by joining consecutive data points $f_1, \ldots, f_{N+1}$ with line segments, where the f_i are i.i.d. random numbers and $f_i \neq f_j$. We label each segment increasing (+) or decreasing (−) and call this string of +'s and −'s the up-down signature σ. We calculate the probability $P(\sigma(f))$ for a random curve and use it to bound the algorithmic information content of f. We show that f can be compressed by $k = \log_2 1/P(\sigma) - N$ bits, where k is a universal currency for comparing the amount of pattern in different curves. By applying our results to microarray time series data, we blindly identify regulatory genes.

Introduction. – Random landscapes are central to the disciplines of spin glasses, drainage networks, protein folding, neural networks and combinatorial optimisation [1,2]. Properties of these systems are related to simple questions about their landscapes: How many minima are there? What is the size of their basins of attraction? What is the pattern of rises and falls?

The large-scale analysis of data series has become increasingly important in the study of financial and biological systems. Identifying trends or pattern, for example, in currency exchange rates or microarray expression data can identify market inefficiencies or the regulatory function of a specific gene. Often the pattern in question is weak (close to random) and it must be identified from amongst a large ensemble of other random data series.

In this letter we show that that there are fruitful underlying connections between the dynamical properties of a 1-D landscape and the presence of pattern in a series of data. Considering a series as a sequence of increases and decreases provides a method of compressing a curve, in the sense that the size of the file needed to store instructions for generating the curve is less than it would be by storing the curve outright. We derive a formal relation between the up-down properties of a curve and the algorithmic information content (AIC) of the equivalent data series, or size of the smallest file needed to store it, which is the ultimate test of pattern. As a demonstration of its efficacy, we use our method to blindly identify regulatory genes from a classic yeast cell cycle microarray data set.

Random data and permutations. – We study the simplest form of random landscape, a sequence of $N+1$ identically and independently distributed random numbers. We connect pairs of consecutive data points with line segments to form a curve. If we assume that the probability that two points are identical is negligible, we can label these segments increasing (+) or decreasing (−). The $N+1$ points can thus be reduced to an up-down signature σ: a string of +'s and −'s of length N. The data points $0.2, 0.6, 0.9, 0.5, 0.3$, for example, have up-down signature $++--$.

Moving from real numbers to their rank order permutation, and from permutations to a signature of +'s and −'s, leads to a loss of information about the curve, but a gain of mathematical rigour and tractability. In particular, because the up-down properties of a data series depend on their relative values only, the distribution of signatures $P(\sigma)$ does not depend on the distribution from which the data is drawn. The up-down picture allows immediate insight into questions about the number of minima and the size of basins of attraction. Moreover, as we shall see, the amount that a curve can be compressed allows us to rigourously compare curves of different lengths or from different experiments.

(a) www.tcm.phy.cam.ac.uk/∼tmf20/

Table 1: Frequencies of permutations $C(\sigma)$ with given up-down signature σ of length N. Since $C(\sigma)$ is symmetric on interchanging $+$ and $-$, only the first half of the frequencies for $N=4$ and $N=5$ are shown. The probability $P(\sigma)$ that a random curve has signature σ is $C(\sigma)/(N+1)!$.

$N=1$		$N=2$		$N=3$		$N=4$		$N=5$		$N=5$ (cont.)	
σ	C	σ	C	σ	C	σ	C	σ	C	σ	C
$-$	1	$--$	1	$---$	1	$----$	1	$-----$	1	$-+---$	14
$+$	1	$-+$	2	$--+$	3	$---+$	4	$----+$	5	$-+--+$	40
		$+-$	2	$-+-$	5	$--+-$	9	$---+-$	14	$-+-+-$	61
		$++$	1	$-++$	3	$--++$	6	$---++$	10	$-+-++$	35
				$+--$	3	$-+--$	9	$--+--$	19	$-++--$	26
				$+-+$	5	$-+-+$	16	$--+-+$	35	$-++-+$	40
				$++-$	3	$-++-$	11	$--++-$	26	$-+++-$	19
				$+++$	1	$-+++$	4	$--+++$	10	$-++++$	5

Consider $N+1$ points with up-down signature σ. If we transform the data such that the ordering of the points remains the same, the up-down signature is unaffected. We can therefore replace each data point with its rank in ascending order (*e.g.* 0.5, 0.3, $0.8 \to 2$, 1, 3). This defines a permutation. By symmetry, every permutation is equally likely to occur, so the probability that $N+1$ random data points has up-down signature σ is equal to the probability $P(\sigma)$ that a random permutation has signature σ. Because permutations are easier to work with, we study the distribution of $C(\sigma)$, where $C(\sigma)$ denotes the number of permutations of $1, 2, \ldots, N+1$ with signature σ (see table 1). This is related to $P(\sigma)$ by

$$P(\sigma) = C(\sigma)/(N+1)!. \tag{1}$$

Unlike the up-down properties of random walks, the enumeration of permutations with a given up-down signature is an intricate problem (non-random random-walks are considered in [3]). It was first studied by André, who showed the probability of the alternating sequence $P(+-+-\ldots)$ is asymptotically $(2/\pi)^N$ [4]. The numbers $C(\sigma)$, which can be regarded as a generalisation of the Euler-Bernoulli numbers, have since been studied from various points of view [5], in particular via representations of the symmetric group [6], but many questions remain unresolved [7]. Questions about the number of minima come up in the study of spin glasses [8].

Calculating $C(\sigma)$. – In studying $C(\sigma)$, it is useful to let $(i_1, i_2, \ldots, i_n)$ be the signature with an island of i_1 pluses, followed by an island of i_2 minuses, *etc.*, where $i_1 + i_2 + \ldots + i_n = N$. For example, $(2,3,1) \equiv ++---+$. Then

$$C(i) = 1 \quad \text{and} \quad C(i,j) = \tbinom{i+j}{i}, \tag{2}$$

but there is no simple formula for $C(i_1, .., i_n)$ when $n \geqslant 3$. Instead we have the recursion relation

$$C(i_1, \ldots, i_n) = C(i_1 - 1, .., i_n) + .. + C(i_1, .., i_n - 1) \tag{3}$$

with boundary conditions $C(\ldots, i, 0, j, \ldots) = C(\ldots, i+j, \ldots)$ and $C(0, i, \ldots) = C(i, \ldots)$. This relates elements in each column of table 1 explicitly in terms of elements in the previous column.

We introduce an iterative method, based on (3), of calculating the distribution of permutations with signature σ. Let $\mathbf{c}_N(\sigma)$ denote the vector of frequencies of signatures of length N in lexicographical order (a column in table 1). Because $C(\sigma)$ is symmetric on interchanging $+$ and $-$, we only consider signatures beginning with $-$; $\mathbf{c}_3$, for example, is $(1,3,5,3)$. We can compute $\mathbf{c}_{N+1}$ from $\mathbf{c}_N$ using the equation

$$\mathbf{c}_{N+1} = T_N \, \mathbf{c}_N, \tag{4}$$

where T_N is a $2^N \times 2^{N-1}$ matrix generated as follows. Denote by 0_N, I_N, and I_N^r the $2^{N-1} \times 2^{N-1}$ zero, identity, and anti-diagonal identity matrices, where, *e.g.*, $I_2^r = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$. We first define the matrices S_N recursively,

$$S_{N+1} = \left[\begin{array}{c|c} S_N & \begin{matrix} 0_N \\ I_N \end{matrix} \\ \hline \begin{matrix} I_N & 0_N \end{matrix} & S_N \end{array}\right], \quad S_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{5}$$

and then set

$$T_N = S_N + \begin{bmatrix} 0_N \\ I_N^r \end{bmatrix}. \tag{6}$$

Equations (4)–(6) imply that each element of C_{N+1} is the sum of a small number (at most the number of islands n) of elements in C_N. Because the matrices T_N are sparse, our recursive definition yields an efficient method of computing the signature frequencies C.

$C(\sigma)$ and algorithmic information: 1 curve. – Here we show that curves with unusual signatures (small $C(\sigma)$) contain pattern in the sense of low AIC or Kolmogorov complexity [9,10]. The Kolmogorov complexity of a string is the length in bits of the shortest description of (or algorithm generating) that string, given a fixed, universal computer. Strings with obvious forms of pattern —such

as linear, periodic or exponential— have low Kolmogorov complexities, but so does a deterministic, chaotic signal —such as the logistic map. (Although in practice it might be difficult to detect the latter.) At most the Kolmogorov complexity is the Shannon information [9], the length of the trivial algorithm "print (f)"; strings with minimal descriptions on order of their length are said to be random. We relate the number of bits k by which a curve can be compressed (the difference between the Shannon and algorithmic information) to its signature σ, thereby giving an upper bound on the AIC.

Consider a random curve f formed from $N+1$ observations f_i drawn from the interval $(0,1]$ with precision $1/T$ (*i.e.*, there are T possibilities for each observation). We suppose T is large enough such that $P(f_i = f_j)$ is negligible for $i \neq j$. The number of bits H required to store the curve in a file without compression is

$$H(f) = -\sum_f T^{-(N+1)} \log_2 T^{-(N+1)} = \log_2 T^{N+1}, \quad (7)$$

which is the Shannon information.

Suppose we do not know the curve f itself, but only its permutation representation. In the case of two points f_1, f_2, the permutation $1,2$ tells us that $f_2 > f_1$; in f_1, f_2 phase space, the data series lies in the triangle above the diagonal. This conveys 1 bit of information. In the case of $N+1$ points, the unit cube can be divided into $(N+1)!$ equal volumes, and the permutation conveys $\log_2(N+1)!$ bits. Because the T^{N+1} possible curves are uniformly distributed over the $(N+1)!$ permutations, there are $T^{N+1}/(N+1)!$ curves in each. Therefore, given the permutation, we need $\log_2(T^{N+1}/(N+1)!)$ additional bits to specify the curve uniquely.

Now suppose we only know the up-down signature σ of f, which requires N bits to write down. It tells us that the curve belongs to a volume containing $C(\sigma)\frac{T^{N+1}}{(N+1)!}$ curves, and specifying which provides us with an algorithm for reconstructing f. Therefore the length of our up-down description is

$$I_\pm(f) = N + \log_2(C(\sigma)\ T^{N+1}/\,(N+1)!), \quad (8)$$

which is an upper bound on the Kolmogorov complexity of f. The number of bits k by which we can compress is

$$k(f) = H(f) - I_\pm(f) = \log_2 \frac{1}{P(\sigma)} - N = -\log_2 \frac{C(\sigma)}{\langle C\rangle}, \quad (9)$$

where $\langle C\rangle = (N+1)!/2^N$. In the case that k is negative, $I_\pm(f)$ is greater than $H(f)$ and $H(f)$ remains the (trivial) bound on f. Thus a single curve f is compressible if its signature satisfies $k(f) < 0$ or $C(\sigma) \leqslant \langle C\rangle$. Because this is an exact, rather than statistical, result, it applies even when the number of data points is small ($N \geqslant 2$).

$C(\sigma)$ and algorithmic information: M curves. – Often, however, we are confronted with the problem of identifying which, if any, of a large number of curves exhibits pattern. In this case we face the additional difficulty that some curves will be compressible by chance alone, just how much depending on the number of curves M. What is the typical reduction in bits of the most compressible of M random curves, $\langle k_M^{\max}\rangle$?

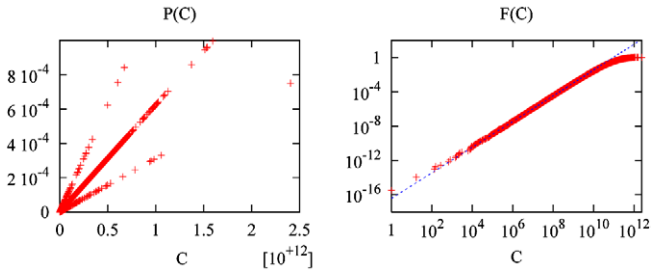


Fig. 1: The distribution $P(C)$ and cumulative distribution $F(C)$ for $N+1=18$ data points. Left: the 3 apparent lines are the result of degeneracies of 2, 4 and 6 in the number of signatures σ with a given frequency $C(\sigma)$. For any $C(\sigma)$, there is at least one other signature σ which has the same frequency C, namely, its reverse on exchange of $+$'s and $-$'s, giving degeneracy 2. Some signatures have degeneracy 4; others 6. For higher values of N the range of degeneracies will likely increase. Right: the cumulative distribution closely satisifies the plotted r.h.s. of eq. (10).

The exact distribution $P(C)$ and exact cumulative distribution $F(C)$ are plotted in fig. 1 for $N=17$. This, and the distributions for other values of N, suggest that $F(C)$ closely satisfies the power law

$$F(C) = \frac{1}{(N+1)!} \sum_{\sigma:\, C(\sigma)\leqslant C} C(\sigma) \simeq \delta \left(\frac{C}{\langle C\rangle}\right)^{3/2} \quad (10)$$

for $C \ll \langle C\rangle$, also plotted in fig. 1. The prefactor $\delta \simeq 0.38$ for $N=17$; it seems to grow slowly with N. Then the probability that M random curves all have frequencies greater than $C(\sigma)$ is $A(\sigma, M) = (1 - F(C(\sigma)))^M$. When $C(\sigma)$ is small, we can linearise this to obtain

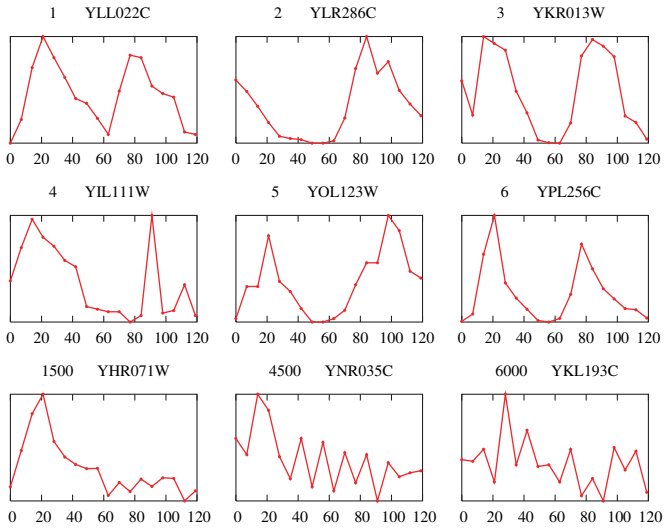
$$A_{\text{approx}}(\sigma, M) \simeq 1 - M\delta \left(\frac{C(\sigma)}{\langle C\rangle}\right)^{3/2}. \quad (11)$$

If a curve f has signature σ, then the quantity $A(\sigma, M)$ is the probability that f is the most unusual of M curves. We list the exact values of $A(\sigma, M)$ and those given by (11) for the top yeast cell cycle curves in fig. 2; they are in very close agreement.

If our confidence that $\sigma(f)$ is the least likely signature of M random curves is at least one half ($A \geqslant 1/2$), then by (11) σ satisfies $C(\sigma)/\langle C\rangle \leqslant (2M\delta)^{-2/3}$. Combining this with (9) gives the corresponding condition on k:

$$k \geqslant \langle k_M^{\max}\rangle \simeq 2/3 \log_2(2M\delta), \quad (12)$$

valid for $M \gg 1$ (recall that $\langle k_1^{\max}\rangle = 0$). This quantifies the more stringent test for significance we must make when



rank	σ	$\log_2(C)$	k	A	A_{approx}	name
1	(3,6,2,6)	24.4	11.1	0.978	0.978	YLL022C
2	(7,5,1,1,3)	24.6	10.9	0.973	0.973	YLR286C
3	(1,1,7,3,5)	24.9	10.6	0.962	0.963	YKR013W
4	(2,9,2,1,2,1)	25.0	10.5	0.956	0.957	YIL111W
5	(3,4,7,3)	25.6	9.87	0.917	0.922	YOL123W
6	(3,5,3,6)	25.7	9.81	0.915	0.918	YPL256C
1500	$(3,4,1_7,2,1)$	35.5	0.02	0.000	—	YHR071W
4500	$(1_2,3,1_{10},2)$	38.4	-2.91	0.000	—	YNR035C
6000	(1_{17})	41.1	-5.62	0.000	—	YKL193C

Fig. 2: The 6073 yeast cell cycle expression curves ranked by their compression in bits k. Here we show the top six genes and three genes from further down the list (rank 1500, 4500, 6000). Top: expression curves for the nine genes, as a function of time in minutes. Bottom: the compression in bits k for the nine genes. Also shown is signature σ (where, for example, the shorthand 1_7 means $1,1,1,1,1,1,1$); frequency of occurrence C; exact probability A that $C(\sigma)$ is the smallest of 6073 random f; A_{approx}, calculated using (11); and gene name.

considering M curves as opposed to one. In particular, we see that increasing the number of curves by a factor of two requires that a curve, to be as significant as before, be one bit more compressible.

Compression in bits is universal currency. – The number of bits k by which a curve can be compressed is the universal currency by which we can compare the significance of different curves, whether they differ in the number of points $N+1$, or the kind of pattern exhibited, or the associated experimental setup. The arithmetic compression k, rather than the geometric compression k/H, is the relevant quantity because the presence of pattern in a data series is piecewise independent.

Although the compression k allows us to compare curves of different lengths, the maximum possible compression of a curve depends on the number of data points it contains. The most compressible curve is the one with the rarest signature —a string of N pluses or N minuses— which can be compressed by $k_{\max} = \log_2(N+1)! - N$ bits. How many data points do we need to identify the presence of pattern in a curve? This depends on the number of curves M. A minimal condition is that the probability of finding $k_{\max}$ by chance be small. In the case of M curves, $P_{k_{\max}}(M) = 1 - (1 - \frac{2}{(N+1)!})^M \simeq 2M/(N+1)!$, the approximation being true for $M \ll (N+1)!$. Enforcing $P_{k_{\max}} \ll 1$ gives $(N+1)! \gg 2M$, which is our condition on the necessary number of data points to uniquely recover a curve with the maximum pattern (increasing or decreasing). In the case of yeast, $M = 6073$ which gives $N \geqslant 8$. Beyond this, increasing N increases the sensitivity of the test to pattern.

It is also possible to order curves by the total number of runs (islands) or the longest run. Both approaches suffer from being too coarse: their range is limited to N values, as opposed to the 2^N possibile up-down signatures. Applied to the yeast cell cycle data below, they gave notably inferior results.

Application to microarray series. – Microarrays, commonly known as DNA chips, allow the simultaneous measurement of the relative concentration of thousands of genes within the cell. These numbers form a complex, unique fingerprint of the state of the cell, and will change in complex ways as a function of time, or the onset of disease, or increasing dose. We construct for each gene a plot of its expression level as a function of the independent variable (see, for example, fig. 2). Can we identify which genes are associated with the independent variable by examining their expression curves alone?

As a benchmark, we analysed the classic yeast cell cycle data of Spellman *et al.* [11], which is publicly available and has been studied at length ([12] and references therein; [13]). The data comprises RNA expression levels as a function of time for all 6073 yeast genes. Each gene was measured at 18 uniformly spaced intervals, spanning two complete cell cycles.

We converted all 6073 gene curves f into up-down signatures $\sigma(f)$, computed their frequencies $C(\sigma)$ and ordered the genes according to $k(C)$. Unlike other authors [11,12,14], we do not bias our search towards any anticipated pattern (such as periodicity) or adjust any free parameters. Expression profiles for the top 6 genes and genes from further down the list are shown in fig. 2, alongside their signatures and compressions in bits ($\langle k_M^{\max} \rangle \simeq 8.12$ for yeast). We tested our predictions by comparing our entire gene list with a validation set of 104 probable cell cycle genes generated from traditional (non-microarray) experiments; this is plotted as a histogram in fig. 3. The bulk of the experimental genes (which themselves contain false positives and negatives) are segregated towards the early part of our list. We also applied this to yeast data containing one cycle (and thus non-periodic) and three cycles with qualitatively similar results. A detailed exposition of our yeast predictions will appear elsewhere.

Discussion. – In our application to yeast cell cycle expression data, the samples were ordered by time. We do

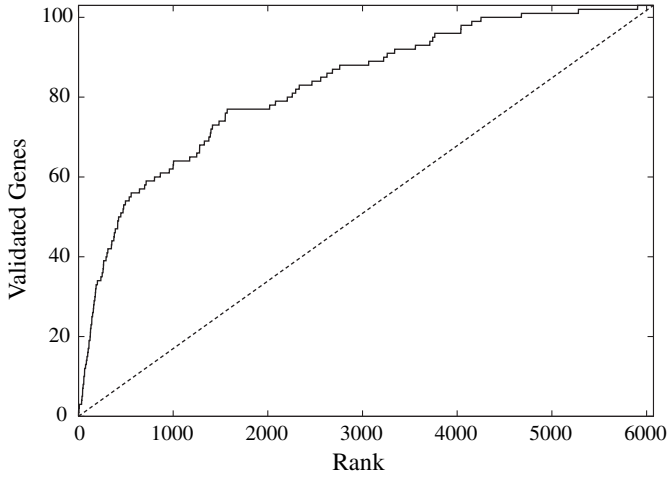


Fig. 3: We compare our ordering of Spellman's cell cycle data with a validation set of 104 probable cell cycle genes derived from experimental (non-microarray) methods. The ROC (receiver operator characteristic) curve compares our ranking (top curve) with a typical random ranking of the genes (bottom straight curve). Starting from the origin, for every gene in our rank in the validation set, the y-coordinate increases; for every gene not in the validation set, the x-coordinate increases. An ordering which perfectly segregates the genes would appear as a step function. Our ordering performs particularly well in the beginning of the list, as the sharp increase near the origin suggests.

not explicitly differentiate between correlation with cell cycle and correlation with time; the two are here synonymous because the cells were synchronised by cell cycle phase and therefore only associated genes' expressions rise and fall in tandem.

One is in general free to order samples by whatever observable one wishes. Thus, for instance, we could have ordered the same yeast cell samples by their temperature (were variation in temperature part of the experimental design). This would yield a new set of curves, based on a reordering of the samples. A compressible curve would imply that the associated gene is likely to depend on temperature.

It should be kept in mind that the presence of pattern in a curve strictly allows us to say the following: that the curve is more likely to have a simple physical dependence on the independent variable than curves with less pattern. It does not allow us to distinguish between curves (genes) explictly governed by the independent variable and genes implictly governed through an intermediate gene, though the latter is likely to exhibit less pronounced pattern. Nor does it allow us to distinguish between pattern caused by *bona fide* experimental dependence and pattern caused by systematic experimental error, because the latter is nonetheless a dependence, only not intended.

Conclusion. – Understanding the connections between the dynamical (up-down) properties of a landscape and the presence of pattern in a data series opens up a broad range of applications. As an illustration we identified known and unknown yeast cell cycle genes from microarray time series data. We are now studying data from a cervical cancer cell line [15] and bladder cancer tumours ordered by severity (Francois Radvanyi *et al.*, unpublished). A fascinating application which we have not pursued here is data derived from financial markets, *e.g.*, the change in a stock price or currency over time. A natural extension from a landscape point of view would be to study random landscapes in higher dimensions, in particular the distribution of the number of minima.

* * *

The authors thank Rosanne Kay for hospitality. KW was supported by the Fondation des Treilles.

REFERENCES

[1] Frauenfelder H. *et al.* (Editors), *Landscape Paradigms in Physics and Biology* (North Holland, Amsterdam) 1997.
[2] Mézard M., Parisi G. and Virasoro M., *Spin Glass Theory and Beyond* (North Holland, Amsterdam) 1987.
[3] Bennett C. A. and Franklin N. L., *Statistical Analysis in Chemistry and the Chemical Industry* (Wiley, New York) 1954, Chapt. 11.
[4] André D., *J. Math.*, **7** (1881) 167.
[5] Warren D. I. and Seneta E., *J. Appl. Probab.*, **33** (1996) 101.
[6] Foulkes H. O., *Discrete Math.*, **15** (1976) 235.
[7] Mallows C. L. and Shepp L. A., *Discrete Math.*, **54** (1985) 301.
[8] Derrida B. and Gardner E., *J. Phys.* (*Paris*), **47** (1986) 959.
[9] Cover Thomas M. and Thomas Joy A., *Elements of Information Theory* (Wiley, New York) 1991, Chapt. 7.
[10] Zurek W. H. (Editor), *Complexity, Entropy and the Physics of Information* (Addison-Wesley, Redwood City, CA) 1990.
[11] Spellman Paul T. *et al.*, *Mol. Biol. Cell*, **9** (1998) 3273; `http://cellcycle-www.stanford.edu`.
[12] Bar-Joseph Ziv *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, **100** (2003) 10146.
[13] Getz G., Levine E., Domany E. and Zhang M. Q., *Physica A*, **279** (2000) 457.
[14] Wolfsberg Tyra G. *et al.*, *Genome Res.*, **9** (1999) 775.
[15] Alazawi William *et al.*, *Cancer Res.*, **62** (2002) 6959.