

## Gene expression

## Identifying genes from up–down properties of microarray expression series

Karen Willbrand<sup>1</sup>, Francois Radvanyi<sup>2</sup>, Jean-Pierre Nadal<sup>1</sup>, Jean-Paul Thiery<sup>2</sup> and Thomas M. A. Fink<sup>2,3,\*</sup><sup>1</sup>Laboratoire de Physique Statistique, Ecole Normale Supérieure, 75231 Paris Cedex 05, France,<sup>2</sup>Institut Curie, CNRS UMR 144, Paris, 75248 France and <sup>3</sup>Theory of Condensed Matter, Cavendish Laboratory, Cambridge CB3 0HE, UK

Received on February 1, 2005; revised on May 2, 2005; accepted on June 16, 2005

## ABSTRACT

**Motivation:** We consider any collection of microarrays that can be ordered to form a progression; for example, as a function of time, severity of disease or dose of a stimulant. By plotting the expression level of each gene as a function of time, or severity, or dose, we form an expression series, or curve, for each gene. While most of these curves will exhibit random fluctuations, some will contain a pattern, and these are the genes that are most likely associated with the quantity used to order them.

**Results:** We introduce a method of identifying the pattern and hence genes in microarray expression curves without knowing what kind of pattern to look for. Key to our approach is the sequence of ups and downs formed by pairs of consecutive data points in each curve. As a benchmark, we blindly identified genes from yeast cell cycles without selecting for periodic or any other anticipated behaviour.

**Contact:** tmf20@cam.ac.uk

**Supplementary information:** The complete versions of Table 2 and Figure 4, as well as other material, can be found at <http://www.lps.ens.fr/~willbran/up-down/> or <http://www.tcm.phy.cam.ac.uk/~tmf20/up-down/>

## INTRODUCTION

The ability to measure thousands of gene expression levels in parallel using microarrays has provided scientists with a complex, unique fingerprint of a cell or tissue sample (The Chipping Forecast II, 2002). Understanding how this fingerprint changes during physiological or pathological processes is one of the most pressing—and promising—problems in bioinformatics. We introduce a fundamentally new approach, unrelated to clustering, to identifying genes from any collection of expression arrays that can be ordered to form a progression (e.g. time series experiments). We construct for each gene a plot of expression level as a function of progression and identify pairs of consecutive data points as increasing (+) or decreasing (−). By studying this succession of +’s and −’s, we identify the genes that are correlated with the progression without knowing the kind of pattern they might exhibit. As a benchmark, we analysed Cho’s and Spellman’s classic yeast cell cycle expression data (Cho *et al.*, 1998; Spellman *et al.*, 1998). We identified 154 (266) genes which with 95% (90%) probability are correlated with cell cycle.

Because up–down analysis can be used to study gene expression as a function of time or disease progression or any increasing dose of a stimulus (Cho *et al.*, 1998; Spellman *et al.*, 1998; Whitfield *et al.*, 2002; Alazawi *et al.*, 2002), we believe it opens a new program of microarray analysis.

We refer to whatever parameter we use to order our microarray progressions as the independent variable. The expression levels of most genes will not be correlated with the independent variable, and the corresponding curves will behave randomly. The small fraction of genes that are correlated will not exhibit random behaviour, but will display some pattern or regularity. Our goal in this paper is to identify these correlated genes by the presence or absence of pattern in their expression curves and to quantify our belief that a given gene is correlated.

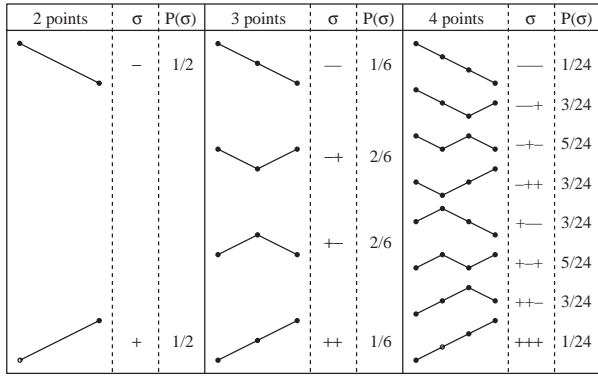
Our problem is difficult for two reasons. First, because we want to identify regulatory (or regulated) genes without knowing details of their role, we do not know what kind of pattern we are looking for. Therefore we need a test for pattern that makes no assumptions about what it might find. Second, we must be beware of false alarms in the form of random curves exhibiting pattern by chance alone. If we look at enough genes, and hence expression curves, we will find any pattern we might imagine. So whatever our test for pattern is, it must become more stringent as the number of curves examined increases.

One of the simplest ways of analysing expression curves is to connect pairs of neighbouring data points with line segments and to label these segments up or down. In this way a set of  $N + 1$  data points can be reduced to its signature  $\sigma$ , a string of +’s and −’s of length  $N$ . For example, the data 0.3, 0.8, 0.6 and 0.2 have signature  $\sigma = + - -$ , which we abbreviate by (1,2), that is, the number of +’s followed by the number of −’s, etc.

Rather than studying the up–down properties of expression curves of correlated genes explicitly, we ask a simpler question: What are the up–down properties of non-correlated, or random, curves? For small numbers of data points, we list the distribution of signatures in Figure 1 and Table 1. We say that a gene is likely to be correlated with the independent variable if the probability  $P(\sigma)$  of its up–down signature is low enough to be statistically deviant from the expected distribution for random curves.

The essence of our approach is the mapping of ordered datasets (curves) to real numbers such that random curves tend to be mapped to higher values and regular or patterned curves to lower numbers. We say a curve has more pattern than another if its algorithmic

\*To whom correspondence should be addressed.



**Fig. 1.** Probability  $P(\sigma)$  of finding a random curve with up-down signature  $\sigma$ , for 2, 3 and 4 data points. The values of  $P$  exhibit intricate mathematical behaviour. For more data points (Table 1).

**Table 1.** Numbers of permutations  $C(\sigma)$  with given up-down signatures, which we call the frequency

$N = 1$		$N = 2$		$N = 3$		$N = 4$		$N = 5$	
$\sigma$	$C$	$\sigma$	$C$	$\sigma$	$C$	$\sigma$	$C$	$\sigma$	$C$
-	1	--	1	---	1	----	1	-----	1
+	1	-+	2	--+	3	---+	4	----+	5
		+-	2	+-+	5	--++	9	---++	14
		++	1	+++	3	++++	6	+++++	10
				+- -	3	+- - -	9	+- - - -	19
				+- +	5	+- + -	16	+- + - +	35
				++ -	3	++ - -	11	++ - + -	26
				+++	1	++++	4	+++++	10
						+- - - -	4	+- - - - -	14
						+- - - +	11	+- - - + -	40
						+- - + -	16	+- - + - -	61
						+- + - +	9	+- + - + -	35
						++ - - -	6	++ - - - -	26
						++ - + -	9	++ - + - +	40
						+++ - -	4	+++ - - -	19
						+++++	1	+++++ -	5

Dividing  $C$  by  $(N + 1)!$  gives the probability  $P(\sigma)$  that a random curve has the same signature.

information content (AIC), or the length of its shortest description, is smaller than that of the other. The ideal map would be the AIC itself, but this is not, in general, computable [The AIC, or Kolmogorov complexity, of a set of data is the shortest possible description of that data given a fixed (and universal) language. Thus  $AIC(2, 4, 6, 8, 10, \dots) < AIC(2, 3, 5, 7, 11, \dots) < AIC(6, 2, 5, 9, 8, \dots)$ ; the shortest algorithm for generating the even numbers is shorter than that for the primes, which in turn is shorter than that for random numbers, the minimal algorithm for the last being ‘print(6, 2, 5, 9, 8, ...).’ Cover and Thomas (1991).

At best ordering by AIC can be approximated, and the success of our approximation using  $P(\sigma)$  results from its remarkable properties. First,  $P(\sigma)$  does not depend on the distribution from which the data is drawn. This is important because we do not in general know the kind of experimental noise to which the data has been submitted. Second,  $P(\sigma)$  is invariant over continuous one-to-one transformations of the data, such as multiplying by a constant or taking the

logarithm. Third, because it can be cast into a discrete framework,  $P(\sigma)$  can be calculated analytically, rather than by brute force.

**METHODS**

**Ordering genes by their up-down signatures**

Throughout this paper we consider  $M$  genes (or curves), each comprised of  $N + 1$  data points. In the case of yeast cell cycle  $M = 6073$  and  $N + 1 = 14, 17, 18$  or  $24$ , depending on the experiment (three time courses are available at <http://cellcycle-www.stanford.edu>). Connecting consecutive pairs of data points yields  $N$  line segments and we attach to each of these segments a plus (+) if it is increasing and a minus (-) if is decreasing. This forms a string of +’s and -’s of length  $N$ , which we call the signature  $\sigma$ ; the data points 3, 7, 9, 5 and 4, for example, have signature + + - -, abbreviated (2,2).

By assumption, most genes are not correlated with the independent variable, and hence will exhibit random fluctuations. The small number of genes that are correlated will tend to exhibit more regular behaviour, although we do not know what kind of pattern this might be. It is these genes that we would like to identify. We identify randomness (or the lack thereof, which we call pattern) by using the correlation between random data and the probability  $P(\sigma)$  of an up-down signature. Up-down signatures of random data tend to have many sign changes and higher  $P(\sigma)$ ; signatures of non-random data tend to fluctuate less and have lower  $P(\sigma)$ .

**Calculating  $P(\sigma)$**

We calculate the probability that a random curve has signature  $\sigma$  by taking advantage of the equivalence between up-down properties of random data and the up-down properties of random permutations. In particular, the probability that  $N + 1$  random data points have signature  $\sigma$  is identical to the probability that a random permutation of the integers  $1, 2, \dots, N + 1$  have the same signature  $\sigma$ . Because they are easier to work with, we study permutations instead of random curves themselves.

Consider all of the permutations of the integers  $1, 2, \dots, N + 1$ , of which there are  $(N + 1)!$ . Each permutation can be reduced to a signature  $\sigma$  of length  $N$ , of which there are  $2^N$  possibilities. Since  $(N + 1)! > 2^N$  for  $N \geq 2$ , more than one permutation will fall to some signatures, but the distribution is far from uniform. We call the number of permutations that have the same signature  $\sigma$  the frequency  $C(\sigma)$ . Table 1 lists  $C$  for various values of  $N$ . The behaviour of these numbers is not straightforward. If we let  $(i, j, \dots)$  denote a group of  $i$  pluses, followed by a group of  $j$  minuses, etc., then we can write

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

but there is no general formula for  $C(i_1, \dots, i_n)$  when  $n > 2$ . Instead we have the recursion relation

$$C(i_1, \dots, i_n) = C(i_1 - 1, \dots, i_n) + \dots + C(i_1, \dots, i_n - 1) \tag{2}$$

with boundary conditions  $C(\dots, i, 0, j, \dots) = C(\dots, i + j, \dots)$  and  $C(0, i, \dots) = C(i, \dots)$ . For example,  $C(2, 3, 2) = C(1, 3, 2) + C(2, 2, 2) + C(2, 3, 1)$  (MacMahon, 1915; Szpiro, 2001).

We place the genes in ascending order of frequency  $C(\sigma)$ . The first gene is most likely to be correlated with the independent variable and the last gene least likely. The probability  $P(\sigma)$  that  $N + 1$  random data points has a signature  $\sigma$  is

$$P(\sigma) = C(\sigma)/(N + 1)!; \tag{3}$$

for example,  $P(+ + - -) = 6/120 = 1/20$ .

The frequency  $C(\sigma)$  takes its minimum value when the data are monotonically increasing (+ + + ...) or decreasing (- - - ...) and its maximum value when the data are oscillating (+ - + - ...). In between the map is more subtle; for example, the value given by Szpiro (2001) and in algorithmic information content (AIC) [The AIC, or Kolmogorov complexity, of a set of data is the shortest possible description of that data given a fixed (and universal) language. In between the map is more subtle; (8,9), for example, is 36 times more likely to be observed than (3,14).

## Quantifying our belief by $A(\sigma, M)$

We first quantify our belief that a gene with signature  $\sigma$  is correlated with the independent variable by the probability  $A(\sigma, M)$  that  $M$  randomly generated expression curves have frequency greater than  $C(\sigma)$ . As a gene's expression curve takes on more regularity of form,  $C(\sigma)$  gets smaller and  $A(\sigma, M)$  approaches 1.

We begin by calculating the cumulative probability  $F(\sigma)$  of finding a signature of a random curve  $\mu$  such that  $C(\mu) \geq C(\sigma)$ . It is defined in terms of  $P(\mu)$  by

$$F(\sigma) = \sum_{\mu: C(\mu) > C(\sigma)} P(\mu). \quad (4)$$

For example, for  $N = 3$ , the values of  $P(\mu)$  are  $P(- - -) = 1/24$ ,  $P(- - +) = 3/24$ ,  $P(- + -) = 5/24$ , etc. Then  $F(- - -) = 22/24$ ,  $F(- - +) = 10/24$ ,  $F(- + -) = 0$ , etc. This is the total fraction of permutations with frequency greater than  $C(\sigma)$ , which is the probability that a curve has frequency greater than  $C(\sigma)$  by chance alone.

The probability that all of  $M$  random signatures have frequency greater than  $C(\sigma)$  is

$$A(\sigma, M) = F(\sigma)^M. \quad (5)$$

The quantity  $1/[1 - A(\sigma, M)]$  is the typical number of times we would have to repeat a random experiment (each comprising  $M$  curves) to find a gene with signature as unlikely as  $\sigma$ .

## Quantifying our belief by $B(\sigma, N)$

Here we estimate a second measure of our confidence that a gene is correlated with the independent variable, the Bayesian probability  $B(\sigma, N)$ .

Key to Bayesian analysis is this: Our confidence that a gene is important *after* considering some evidence (in our case the up-down signature of its expression curve) depends on our confidence before considering that evidence. The signature  $\sigma$  determines the change in our belief that a given gene is correlated.

We assume that some small fraction  $\epsilon$  of genes is correlated with the independent variable; in the case of yeast cell cycle we take  $\epsilon = 0.1$  (Payne and Garrels, 1997) (we also consider other values below). We call the set of these genes  $U$  and the set of uncorrelated genes  $V$ . The number of genes in  $U$  is  $M\epsilon$ .

Let  $H_U$  be the hypothesis that a particular gene  $g$  is in  $U$  and  $H_V$  the hypothesis that it is in  $V$ . The a priori probability that an arbitrary gene is correlated is  $\epsilon$ . We would like to know the a posteriori probability that hypothesis  $H_U$  is true after observing the signature  $\sigma$  of gene  $g$ .

Bayes theorem states that

$$P(g \in U | \sigma) = \frac{P(\sigma | g \in U)P(g \in U)}{P(\sigma | g \in U)P(g \in U) + P(\sigma | g \in V)P(g \in V)}. \quad (6)$$

The probability that  $g \in U$  is  $\epsilon$  and the probability that  $g \in V$  is  $1 - \epsilon$ . The term  $P(\sigma | g \in V)$  is the probability of finding a random curve with signature  $\sigma$ ; it is  $P(\sigma)$  from Equation (3) above.  $P(\sigma | g \in U)$  is the probability of finding a correlated gene's curve with signature  $\sigma$ ; call this  $\alpha$ . Then

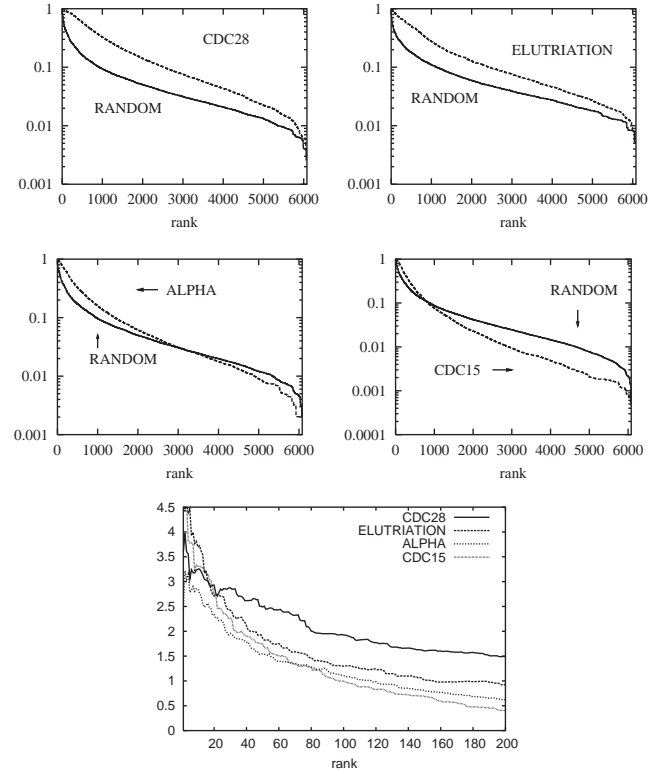
$$P(g \in U | \sigma) = \frac{\alpha\epsilon}{\alpha\epsilon + (1 - \epsilon)P(\sigma)}. \quad (7)$$

The distribution  $\alpha$  depends on the amount of noise and the kind of pattern contained in the correlated gene expression curves, neither of which are in general known. In the limit of high noise, any pattern present is lost and the data may be considered random. In this case  $\alpha$  is equal to  $P(\sigma)$  and Equation (6) reduces to  $\epsilon$ , as expected; very noisy data do not provide any evidence that a gene is correlated.

The simplest choice of distribution  $\alpha$  is to take all signatures of genes involved in cell cycle to be equally likely to occur. Then  $\alpha = 1/2^N$  and, for  $\epsilon \ll 1$ , Equation (7) reduces to

$$B(\sigma, N) = P(g \in U | \sigma) \simeq \frac{\epsilon}{\epsilon + 2^N P(\sigma)}. \quad (8)$$

In the case of  $2^N P(\sigma) \ll \epsilon$ , which is true for our top ranked genes, Equation (8) reduces to  $B(\sigma, N) \simeq 1 - 2^N P(\sigma)/\epsilon$ . This is our estimate



**Fig. 2.** Top: Log-linear plots of Bayesian probability  $B(\sigma, N)$  as a function of gene rank for both real and randomly generated data. This is equivalent to plotting column 4 of Table 2, but for each individual experiment. Bottom: Direct comparison of the four experiments by plotting  $\log[(1 - B_{\text{ran}})/(1 - B_{\text{real}})]$  as a function of gene rank. Higher values correspond to higher signal-to-noise ratios.

of the probability with which we believe that a gene is correlated with the independent variable after considering the microarray evidence. It is only an estimate because we approximated  $\alpha$  by a uniform distribution.

The two quantities  $A(\sigma, M)$  and  $B(\sigma, N)$  differ foremost in that  $B$  is the probability that a single curve is associated with the independent variable (without regard to the number or shape of other curves), whereas  $A$  is the probability that an independent random event will not fall below  $C(\sigma)$ . The quantity  $A$  depends explicitly on  $M$ : increasing  $M$  (with all else fixed) causes the value of  $A$  for a given curve to diminish, thereby becoming more stringent. Because  $B$  cannot itself vary in stringency with  $M$ , one must compare  $B_{\text{real}}$  with the random null case  $B_{\text{ran}}$ , as we do in Figure 2. But note that while only  $B$  depends on  $N$  explicitly, both  $A$  and  $B$  depend on  $N$  implicitly through  $\sigma$ .

## APPLICATION TO MICROARRAY TIME SERIES

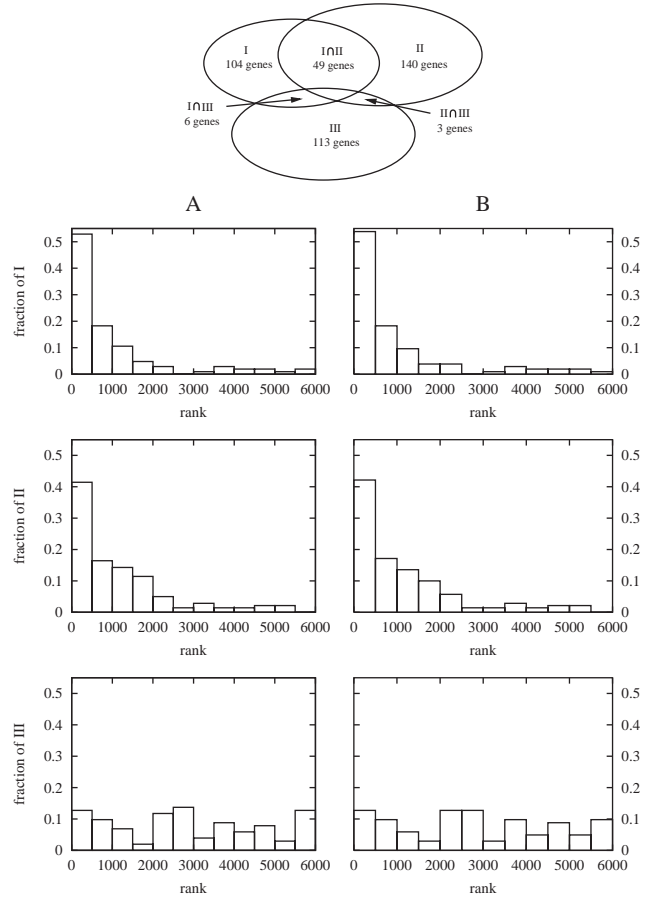
We tested our method on the yeast cell cycle transcript time courses of Cho *et al.* (1998) and Spellman *et al.* (1998), both of which have been studied at length (Shedden and Cooper, 2002; Zhao *et al.*, 2001; Wolfsberg *et al.*, 1999; Getz *et al.*, 2000; Wu *et al.*, 2002); the latter time course is available at <http://cellcycle-www.stanford.edu>. Cho generated two sets of time series data using two methods of synchronization but only one is publicly available: CDC28 (17 time points over 2 cell cycles). Spellman generated data using three methods: elutriation (14 points over 1 cycle),  $\alpha$ -factor (18 points over 2 cycles) and CDC15 (24 points over 3 cycles).

**Table 2.** Top: The 30 genes most likely correlated with yeast cell cycle, in order of Bayesian probability  $B(\sigma, N)$

Rank	Signature $\sigma$	$A(\sigma, M)$	$B(\sigma, N)$	Gene name	Experiment
1	(2,4,10)	0.9999	0.9998	YLR377C	CDC28
2	(1,1,7,5,7,2)	0.9996	0.9998	YKL177W	CDC15
3	(10,1,1,1,3)	0.9996	0.9997	YGL230C	CDC28
4	(10,1,2)	0.9998	0.9997	YOR383C	ELUT
5	(2,5,5,7,4)	0.9983	0.9995	YNL058C	CDC15
6	(1,1,1,1,5,2,1,2,9)	0.9957	0.9991	YNR044W	CDC15
7	(6,7)	0.9975	0.9984	YLR286C	ELUT
8	(7,6)	0.9978	0.9984	YDR451C	ELUT
9	(1,1,1,9,4)	0.9951	0.9983	YMR245W	CDC28
10	(1,6,4,6,4,2)	0.9858	0.9980	YPL256C	CDC15
11	(1,3,5,7,4,3)	0.9818	0.9976	YKL164C	CDC15
12	(1,2,1,1,5,1,1,3,8)	0.9808	0.9975	YGL089C	CDC15
13	(3,9,1)	0.9952	0.9974	YNL066W	ELUT
14	(3,9,1)	0.9952	0.9974	YKL181W	ELUT
15	(4,3,1,1,5,7,2)	0.9793	0.9974	YBL003C	CDC15
16	(1,1,3,1,1,9)	0.9908	0.9973	YGL117W	CDC28
17	(3,6,2,6)	0.9780	0.9955	YLL022C	ALPHA
18	(9,1,1,1,1)	0.9884	0.9954	YJL080C	ELUT
19	(9,1,1,1,1)	0.9884	0.9954	YIL123W	ELUT
20	(3,4,5,7,1,1,1,1)	0.9465	0.9948	YGR092W	CDC15
21	(1,6,2,1,1,7,3,1,1)	0.9457	0.9948	YBR088C	CDC15
22	(8,3,2,3)	0.9752	0.9947	YJL052W	CDC28
23	(5,5,1,5)	0.9711	0.9943	YOR188W	CDC28
24	(4,8,1)	0.9832	0.9942	YJL177W	ELUT
25	(4,8,1)	0.9832	0.9942	YKR042W	ELUT
26	(1,1,7,3,5)	0.9615	0.9936	YKR013W	ALPHA
27	(1,3,6,5,1,1,5,1)	0.9244	0.9934	YCL014W	CDC15
28	(1,9,1,1,1)	0.9768	0.9933	YBR296C	ELUT
29	(3,5,5,1,1,5,3)	0.9210	0.9932	YDR225W	CDC15
30	(2,9,2,1,2,1)	0.9559	0.9929	YIL111W	ALPHA
1	(8,4,1,2,1)	0.9665	0.9937	RAN2619	17
2	(1,5,3,1,6)	0.8662	0.9833	RAN2907	17
3	(1,1,2,3,1,4,7,3,1)	0.4869	0.9690	RAN4365	24
4	(2,1,6,5,2,1)	0.3868	0.9488	RAN1727	18
5	(1,1,4,7,1,1,1)	0.2887	0.9358	RAN5732	17
6	(6,1,1,1,2,5,2,1,2,1,1)	0.0424	0.9145	RAN5785	24
7	(5,1,4,2,2,1,3,2,3)	0.0398	0.9134	RAN4935	24
8	(3,2,7,2,1,1,1,1,1,3)	0.0283	0.9072	RAN4666	24
9	(1,1,3,1,6,1,3)	0.1006	0.9052	RAN3387	17
10	(1,1,4,6,1,1,1,3,3,2)	0.0131	0.8944	RAN1471	24

The complete list (6073 genes) is included in the Supplementary information. Bottom: The 10 most significant (highest  $B$  value) random genes, for comparison; note the rapid drop-off of  $B$ . The last column indicates the experiment from which each gene was identified (top) and the number of data points with which each random expression curve was generated (bottom).

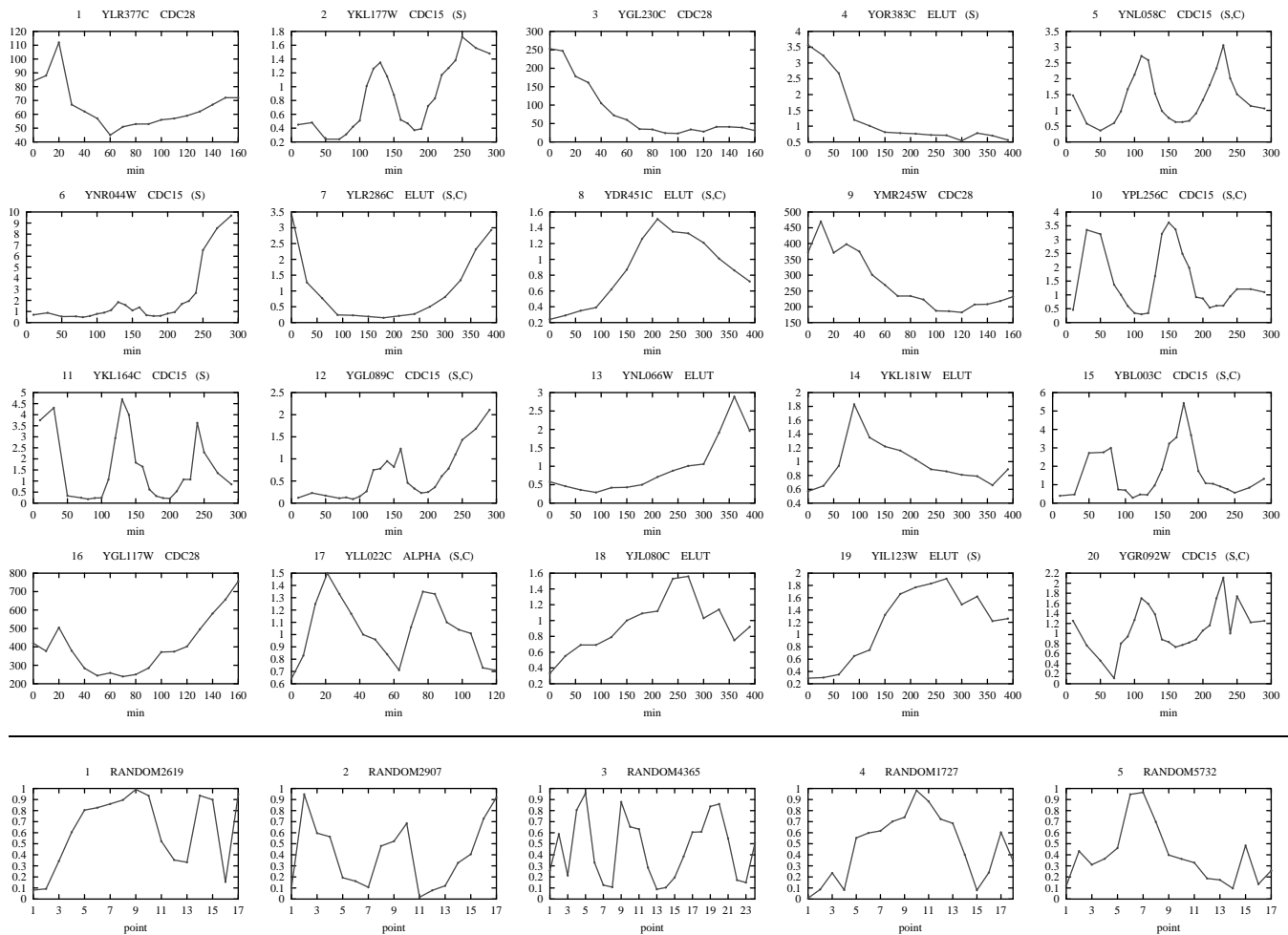
Both Cho *et al.* (1998) and Spellman *et al.* (1998) analysed their datasets by looking for expression curves that displayed periodic behaviour over two or more cell cycle periods. In the case of Cho *et al.* (1998), 78% of the expression curves were discarded on the basis of insufficient variation, with the remainder visually examined for periodicity. Spellman *et al.* (1998) removed aberrant data points before calculating the Fourier transforms of the expression curves, which were then individually scaled by a cell cycle phase peak correlation factor. The elutriation experiment was not subject to Spellman's analysis, ostensibly because, being measured over one cycle, its periodicity could not be ascertained.



**Fig. 3.** Venn diagram: The three sets of experimentally determined cell cycle genes of Spellman *et al.* (1998) (I) available at <http://genome-www.stanford.edu/cellcycle/data/rawdata/KnownGenes.doc>, Simon *et al.* (2001) (II) and (especially) Stevenson *et al.* (2002) (III) only partially overlap. This means that two or more of them have many false positives or false negatives. HISTOGRAMS: We rank genes by  $A(\sigma, M)$  and by  $B(\sigma, N)$  and compare our predictions with three sets of known cell cycle genes: (I) 104 genes collected from the literature by Spellman *et al.* (1998, <http://genome-www.stanford.edu/cellcycle/data/rawdata/KnownGenes.doc>), (II) 140 genes found to be regulated by at least one cell cycle transcription factor using genome-wide location analysis (Simon *et al.*, 2001) and (III) 113 genes found to alter cell cycle when overexpressed using flow cytometry (Stevenson *et al.*, 2002). We attributed to the genes in each set their rank numbers in the respective master list and plot the distribution of these numbers in the form of histograms.

We have interpreted Cho's and Spellman's cell cycle data by analysing the up-down signatures of the time evolution of each gene. From  $P(\sigma)$ , we computed two quantities. The first is the probability  $A(\sigma, M)$  that  $P(\sigma) \geq P(\mu_i)$  for  $M = 6073$  random curves, where  $\mu_i$  is the signature of random curve  $i$ . The second is the approximate Bayesian probability  $B(\sigma, N)$  that a gene is correlated with cell cycle given an a priori belief that 10% of the genes are cell cycle related (Payne and Garrels, 1997) and assuming that the signatures of correlated genes are uniformly distributed.

For each of the four cell cycle experiments, we ranked the genes by descending  $A(\sigma, M)$  [equivalent to ranking by ascending  $P(\sigma)$ ] and  $B(\sigma, N)$ . We do not include the eight gene lists here but instead



**Fig. 4.** Top: Expression curves of the 20 yeast genes most likely to be correlated with yeast cell cycle, ranked by  $B(\sigma, N)$ . Above each curve is listed its rank, the gene name, the experiment from which it was drawn and an indication of whether it was associated with cell cycle by the teams of Cho (C) or Spellman (S). Different curves exhibit different kinds of pattern, corresponding to experiments over one cell cycle (elutriation), two cell cycles (CDC28 and  $\alpha$ -factor) or three cell cycles (CDC15). The complete set of plots (6073 genes) is included in the Supplementary information. Bottom: The five random expression curves most likely to be correlated with the independent variable, for comparison. They form the background against which correlated curves must be distinguished.

combined the four  $B$ -ranked lists (included in the Supplementary information) to form a single master list (Table 2) in the following way. We attached to each gene the highest value of  $B(\sigma, N)$  observed for that same gene in the four experiments and ranked the genes accordingly. Because up–down analysis admits false positives with low probability, taking the maximum value for each gene should not significantly bias our results. We list the first 30 genes in Table 2. Ordering genes by  $A(\sigma, M)$  gives a largely, though not exactly, similar ranking (Fig. 3). For comparison, we include the top 10 artificial (random) genes from a computer experiment, in which the expression levels were randomly generated. The gene expression curves of the top 20 genes are plotted in Figure 4, alongside the top 5 random expression curves.

We blindly identified 154 genes which with  $B(\sigma, N) > 0.95$  are correlated with cell cycle. (This is the case for  $\epsilon = 0.1$ ; for  $\epsilon = 0.05$ , we identified 86 genes; for  $\epsilon = 0.2$ , 250 genes.) This should not be interpreted as a binary classification of genes as correlated or

uncorrelated; we observed a continuum of evidence for gene correlations and the 0.95 cutoff is an arbitrary but convenient way of quantifying the number of outliers, a 0.9 cutoff yielded 266 genes. The first 25 of the 154 are more unusual than the least likely of 6073 random genes. Unlike other studies (Cho *et al.*, 1998; Spellman *et al.*, 1998; Whitfield *et al.*, 2002), we did not bias our search towards any anticipated pattern (such as periodicity), filter the data for outliers or adjust any free parameters.

We sought independent confirmation of our predictions by comparing our master list ranked by  $A$  (not shown) and  $B$  (Table 2) to non-microarray assignments of function. We considered three sets of known cell cycle genes (<http://genome-www.stanford.edu/cellcycle/data/rawdata/KnownGenes.doc>; Simon *et al.*, 2001; Stevenson *et al.*, 2002), described in Figure 3. Sets I and II are strongly biased towards our early ranked genes. Set III appears to be uncorrelated with  $A$  or  $B$ , suggesting a poor overlap between microarray and overexpression experiments.

## DISCUSSION AND CONCLUSIONS

An unexpected outcome of our approach is that it offers a means of estimating experimental efficacy. Different experiments are more or less successful at distinguishing signal (correlated genes) from noise (uncorrelated genes), and we can compare them as follows. We plot  $B(\sigma, N)$  as a function of rank for real and randomly generated data for the four experiments in Figure 2 (top) and compare the logarithm of the ratio  $(1 - B_{\text{ran}})/(1 - B_{\text{real}})$  in Figure 2 (bottom). The greater the logarithm, the more effective an experiment is at distinguishing correlated genes from randomly fluctuating genes. This concurs (as expected) with our combined gene list: of our top 154 genes, 59 were best identified by CDC28, 34 by elutriation, 27 by  $\alpha$ -factor and 34 by CDC15. The complete individual gene lists can be found in the Supplementary information. Shedden and Cooper (2002), who analyse the same datasets, favour elutriation as the least perturbing synchronisation method and criticize Spellman *et al.* (1998) who neglected it. In our analysis elutriation clearly contains information and does a good job: it best identified nearly a quarter (22%) of our top 154 genes. Further they argue that the list of 104 confirmed cell cycle genes (set I) is biased towards non-elutriation. We see this tendency: in set I elutriation best identified is only 15%.

Throughout this paper we have distinguished between genes that are functionally associated with and genes that are correlated with the independent variable, the former being a subset of the latter. Microarray studies do not normally allow us to distinguish between the two. Moreover, there is broad spectrum of degrees of functional association, ranging from direct physical interaction to high-order knock-on effects (in the language of genetic networks, nearest-neighbour to proximate vertices). In general, however, those genes that are most correlated with the independent variable are the genes most likely to be functionally associated with the same variable.

Alongside techniques such as clustering (Getz *et al.*, 2000; Eisen *et al.*, 1998) and supervised learning (Furey *et al.*, 2000), and up-down analysis should prove to be a valuable tool in exploiting the biological fingerprint offered by DNA chips. Up-down analysis is fundamentally a technique for identifying genes associated with the independent variable of a set of observations, without regard to gene-gene interactions. The goal of clustering, on the other hand, is to identify groups of genes that exhibit similar behaviour, without regard to their association with the independent variable. Therefore up-down analysis is useful when a collection of microarray samples can be placed in a definite order. Clustering can be applied whether or not the samples form a natural sequence, but neglects the dynamical information implicit in an ordered set of points (a curve).

Up-down analysis is not fool-proof. The loss of information in reducing an  $N + 1$  dimensional data space to  $2^N$  discrete signatures leads us to occasionally overlook genes that are in fact correlated (some false negatives). On the other hand, if a gene is not correlated with the independent variable, it is unlikely to have an unusual signature (few false positives).

In this paper we applied up-down analysis to microarray times series. Our method is equally applicable to a collection of microarrays ordered by any observable behaviour. For example, by ordering a set of tumour samples by pathological severity (stage and grade),

microarray data can be used to generate curves for each gene as a function of tumour development. Moreover, we can test different hypotheses using the same set of data: genes associated with tumour aggressiveness could be identified by reordering the tumour samples by time to death; alternatively, ordering by tumour size might highlight genes associated with the disease. We are currently applying these ideas to a collection of breast tumour samples (Radvanyi, F., Thiery, J.P., Chopin, D., Graham, A., unpublished data). A fascinating extension that we do not investigate here is the ordering of microarrays according to the expression of a single gene  $x$ : the samples are permuted such that the expression of gene  $x$  is monotonically increasing. By repeating this for every gene, gene-gene correlations could be identified.

## ACKNOWLEDGEMENTS

This work was supported by the CNRS, the Institute Curie, and the Comité de Paris Ligue Nationale Contre le Cancer.

*Conflict of Interest:* none declared.

## REFERENCES

- Alazawi, W. *et al.* (2002) Changes in cervical keratinocyte gene expression associated with integration of human papillomavirus 16<sup>1</sup>. *Cancer Res.*, **62**, 6959–6965.
- Cho, R.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
- Cover, T.M. and Thomas, J.A. (1991) *Elements of Information Theory*. Wiley, New York, Ch. 7.
- Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
- Furey, T. *et al.* (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression. *Bioinformatics*, **16**, 906–914.
- Getz, G. *et al.* (2000) Super-paramagnetic clustering of yeast gene expression profiles. *Physica A*, **279**, 457–464.
- MacMahon, P.A. (1915) *Combinatorial Analysis Vol. I*, Cambridge University Press.
- Payne, W.E. and Garrels, J.I. (1997) Yeast Protein Database (YPD): a database for the complete proteome of *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **25**, 57–62.
- Shedden, K. and Cooper, S. (2002) Analysis of cell-cycle gene expression in *Saccharomyces cerevisiae* using microarrays and multiple synchronization methods. *Nucleic Acids Res.*, **30**, 2920–2929.
- Simon, I. *et al.* (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, **106**, 697–708.
- Spellman, P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
- Stevenson, L.F. *et al.* (2002) A large-scale overexpression screen in *Saccharomyces cerevisiae* identifies previously uncharacterised cell cycle genes. *Proc. Natl Acad. Sci. USA*, **98**, 3946–3951.
- Szpiro, G. (2001) The number of permutations with a given signature, and the expectations of their elements. *Disc. Math.*, **226**, 423–430.
- The Chipping Forecast II. *Nat. Genet.* **32**(suppl), 461–552 (2002).
- Warren, D.I. and Seneta, E. (1996) Peaks and eulerian numbers in a random sequence. *J. Appl. Prob.*, **33**, 101–114.
- Whitfield, M.L. *et al.* (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.
- Wolfsberg, T.G. *et al.* (1999) Candidate regulatory sequence elements for cell cycle-dependent transcription in *Saccharomyces cerevisiae*. *Genome Res.*, **9**, 775–792.
- Wu, L.F. *et al.* (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Gen.*, **31**, 255–265.
- Zhao, L.P. *et al.* (2001) Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl Acad. Sci. USA*, **98**, 5631–5636.