BIOINFORMATICS

Vol. 16 no. 8 2000
Pages 685–698

Generation of patterns from gene expression data by assigning confidence to differentially expressed genes

Elisabetta Manduchi[1,*], Gregory R. Grant[1], Steven E. McKenzie[2], G. Christian Overton[1], Saul Surrey[2] and Christian J. Stoeckert Jr.[3]

[1]Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104, USA, [2]Hematology/Oncology Research, A.I. duPont Hospital for Children, Wilmington, DE 19803, and Department of Pediatrics, Jefferson Medical College, Philadelphia, PA 19107, USA and [3]Division of Hematology, The Children's Hospital of Philadelphia and Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104, USA

Received on December 15, 1999; revised on March 14, 2000; accepted on March 21, 2000

Abstract

Motivation: *A protocol is described to attach expression patterns to genes represented in a collection of hybridization array experiments. Discrete values are used to provide an easily interpretable description of differential expression. Binning cutoffs for each sample type are chosen automatically, depending on the desired false-positive rate for the predictions of differential expression. Confidence levels are derived for the statement that changes in observed levels represent true changes in expression. We have a novel method for calculating this confidence, which gives better results than the standard methods. Our method reflects the broader change of focus in the field from studying a few genes with many replicates to studying many (possibly thousands) of genes simultaneously, but with relatively few replicates. Our approach differs from standard methods in that it exploits the fact that there are many genes on the arrays. These are used to estimate for each sample type an appropriate distribution that is employed to control the false-positive rate of the predictions made. Satisfactory results can be obtained using this method with as few as two replicates.*

Results: *The method is illustrated through applications to macroarray and microarray datasets. The first is an erythroid development dataset that we have generated using nylon filter arrays. Clones for genes whose expression is known in these cells were assigned expression patterns which are in accordance with what was expected and which are not picked up by the standards methods. Moreover, genes differentially expressed between normal and leukemic cells were identified. These included genes whose expression was altered upon induction of the leukemic cells to differentiate. The second application is to the microarray data by Alizadeh* et al. *(2000). Our results are in accordance with their major findings and offer confidence measures for the predictions made. They also provide new insights for further analysis.*

Availability: *Software is available on request from the authors.*

Contact: *manduchi@pcbi.upenn.edu*

Introduction

The goal of this paper is to provide tools for the investigator to aid in the analysis of data collected from highly parallel gene expression experiments, such as hybridization array experiments. In particular we wanted to generate descriptive, yet dependable, expression patterns representing the differential expression of genes across cell types. Figure 1 illustrates how an excerpt from a typical 'raw' input might be transformed into an easily interpretable list of patterns.

Suppose we are comparing between two sample types, type A and type B (e.g. these could be two different filter arrays, or the two channels on a microarray, or the red-to-green ratios from two separate two-channel microarrays using the same reference for one of the channels). With highly parallel experiments we are presented with gene expression data from hundreds to thousands of genes simultaneously. We wish to identify those genes that are 'most likely' to be differentially expressed. Variation in

*To whom correspondence should be addressed.

(a)

gene tag	G_0 I	G_0 II	G_0 III	G_1 I	G_1 II	G_2 I	G_2 II	G_2 III	G_3 I	G_3 III
1	0.0114	0.0328	0.0151	0.0060	0.0236	0.0436	0.5640	0.8920	0.0639	0.2490
2	0.0050	0.0131	0.0061	0.0041	0.0364	0.0296	0.8830	0.7000	0.0199	0.1050
3	0.0629	0.2340	0.0431	0.2270	0.2120	0.0105	0.1400	0.0243	0.0117	0.0907
4	0.0250	0.0600	0.0264	0.1500	0.2660	0.0134	0.1860	0.0851	0.0172	0.0112

(b)

gene tag	G_1	G_2	G_3
1	0	7	2
2	0	8	1
3	2	-1	-1
4	3	1	0

Fig. 1. Excerpt from typical input data (normalized intensities) for four gene tags (a) and the patterns associated to these data by our program (b). Note there are three replicate experiments for sample types G_0 and G_2, and two replicate experiments for types G_1 and G_3. In (b) positive (resp. negative) integers represent up-regulation (resp. down-regulation) with respect to the reference group G_0.

the data (biological and/or experimental) can result in the observed level for a given gene in type B to be higher than in type A when in fact the gene is not truly up-regulated in type B. However, if C is a constant and if the intensity of a gene in type B is more than C times the intensity in type A, then if C is sufficiently large it will be unlikely that this observation is due just to variation. If C is too large, then some true cases of up-regulation will be missed unnecessarily. Therefore, the determination of an appropriate C should be based on explicit measures of confidence in order to effectively guide the predictions about differential expression. The measures of confidence should be based on the variability in the data with replicates used to gauge this variation.

Claverie in his survey paper (Claverie, 1999) points out that most published studies are 'quite elusive about measurement reproducibility and the confidence levels of the observed changes in expression are rarely assessed using standard methods.' He examines several studies that do provide replicate data and evaluates their thresholds with the standard methods, showing that rather high thresholds must be chosen when there are only two or three replicates, in order to get significance levels of 5%. Furthermore, the significance measure computed is the probability of predicting a gene is differentially regulated when it is actually not (which we call *the false-positive rate*). Because we are searching for a small set of genes from a large pool, one needs a very small false-positive rate to have a reasonable confidence in the predictions. For example, suppose that 50 of 1000 genes are up-regulated in type B, and our false-positive rate is 0.05, then there will be on average 47.5 genes falsely predicted to be up-regulated, nearly as many as are truly up-regulated. So the confidence in the prediction that a gene is up-regulated cannot be much higher than 50%. For many applications, this confidence level is much too low to be acceptable. To get the false-positive rate down and maintain reasonable cutoffs, standard methods require many replicates (see Claverie, 1999), on the order of ten or more. We have developed a method which gives reasonable cutoffs with as few as two replicates and which gives false-positive rates low enough to maintain good confidence levels. Our method exploits the fact that there are hundreds of genes to estimate appropriate gene-independent distributions in each sample type. By integrating over these distributions, false-positive rates are calculated directly. Finally, measures of confidence are then computed using Bayes theorem. Note however, this is not a Bayesian approach, we simply use Bayes theorem to reverse the conditionality clauses to turn the false-positive rate into a confidence measure.

We note that in Chen *et al.* (1997) a statistical analysis procedure is presented to determine differential expression for the special case of comparisons between the two channels of a single microarray. In contrast, our techniques are general and can be applied to many types of data. In Section **Results** we illustrate applications of the method to macroarray hematopoietic data that we have generated. We also illustrate an application to two-channel microarray data described in Alizadeh *et al.* (2000). (In the latter the comparisons are between red-to-green ratios from separate two-channel microarrays using the same reference for the green channel.) Software implementing the pattern generating algorithm has been developed and is available from the authors together with the documentation.

We end this introduction by defining some terminology that will be used throughout this paper. The gene expression data is taken from a collection of

samples. Here we use the generic term 'sample' to denote either an individual cell, or a cell type, or a tissue type, at a certain time point and under certain conditions. In hybridization arrays, data for one sample are in the form of intensities of spots on the array, where each spot corresponds to a gene. Arrays usually have *gene tags* on them, of some sort or another, for example clones or oligonucleotides, and often they have different tags for the same gene. The intensities of such tags might be merged into one datum. However, all occurrences are not always known, so to be precise we will often need to refer to expression levels and expression patterns of gene tags rather than of genes themselves.

Methods and Algorithm

In this section we describe our protocol for defining expression patterns. The input consists of normalized data, where the normalization procedure depends on the kind of experiments conducted and is therefore left to the user. In Sections **Application to an erythroid development nylon filter dataset** and **Application to a two-channel microarray lymphocyte dataset** we describe the normalization procedures used respectively for our nylon filter data and for the two-channel microarray data of Alizadeh *et al.* (2000). The input normalized intensities are subjected to preprocessing steps, which might include a shift of the intensities by a constant. These steps, together with the details on the binning procedure, are presented in Figure 2 and also discussed in Sections **Implementation** and **Results**. The reason for and the effect of the numerical shift are discussed below and in Sections **Comparison with standard statistical analysis methods** and **Discussion**.

In what follows, we will assume that the necessary preprocessing steps have taken place (including all appropriate normalizations and elimination of data with values too low to be distinguishable from background and including any numerical shift) and we describe our rationale in assigning a pattern to each gene tag under consideration. Unless otherwise stated, the expression 'intensity' refers to the processed intensity.

For each sample type, experiments should be replicated one or more times. Replicates allow the variability to be estimated. If no replicates are given for a sample type, default values are chosen, however, one can clearly make no meaningful estimate about confidence for those sample types with no replicates. We will use the expression *homotypic group of samples*, or *homotypic group* for short, to refer to a set of samples of the same type, possibly consisting of just one sample if no replicate experiments are performed on that sample type. In each gene tag's expression pattern there will be one symbol for each homotypic group. For each homotypic group and for each gene tag, we compute the average intensity of that tag over those samples in the group which have values for that tag. This will represent the intensity of that tag at that group.

The assignment of an expression pattern to each gene tag is done in two stages. In the first stage, we attach to each tag an ordered list of real numbers. In the second stage, we bin the numbers in this list, resulting in a pattern of integers. Binning levels are chosen to take into account the variability within each homotypic group and to ensure an expected false-positive rate of $s\%$ for our predictions about up- or down-regulation.

For the first stage, we start by fixing an ordering of the groups in our collection. If the collection is an ordered series (e.g. a time series), then an ordering is given to us *a priori*. There will be some reference group to which we compare our groups, i.e. with respect to which up- and down-regulation are measured. The reference group is chosen by the user according to the questions posed. Moreover, the user might want to compare expression levels to the median of the group intensities in the data, rather than to a particular group. Thus, to each tag we attach the ordered list of real numbers obtained by dividing each of its (non-reference) group intensities by its group intensity at the provided reference group or, respectively, by the median of its group intensities. We will refer to this list as the list of *ratios* attached to that tag. We denote by ℓ the length of this list (this means that we started with $\ell+1$ homotypic groups, if ratios were taken to a reference group, or with ℓ homotypic groups, if ratios were taken to the median).

For the second stage, for each (non-reference) group, we partition the range $[0,\infty)$ into disjoint subintervals, which we call *bins*. These bins depend on the group. Thus each group has its own set of bins and the number of bins used also depends on the group. Before showing exactly how to choose the bins we establish some notation and describe how the binning takes place. For each group, we number the bins from left to right using consecutive integers $-m,\ldots,0,\ldots,n$, where the bin labeled 0 is the one containing the ratio 1 (so it represents the 'no change level'). We then attach to each gene tag the (ordered) list of integers, which we will call *levels*, obtained by looking, for each group, to which bin the ratio value for the tag at that group belongs. This list will represent the *expression pattern* of that tag. More precisely, suppose that for group i we have subdivided $[0,\infty)$ into m_i+n_i+1 bins as $[0,\infty)=B_{i,-m_i}\cup B_{i,-m_i+1}\cup\cdots\cup B_{i,0}\cup\cdots\cup B_{i,n_i}$, where $B_{i,-m_i}=[0,a_{i,1})$, $B_{i,-m_i+1}=[a_{i,1},a_{i,2})$, $\ldots$, $B_{i,n_i}=[a_{i,m_i+n_i},\infty)$, for some real numbers $a_{i,1}<a_{i,2}<\cdots<a_{i,m_i+n_i}$. If the list of ratios obtained in the first stage for a certain gene tag is $(r_1,r_2,\ldots,r_\ell)$, then each r_i $(i=1,2,\ldots,\ell)$ belongs to exactly one of the bins $B_{i,j}$ $(j=-m_i,\ldots,n_i)$, say r_i belongs to B_{i,j_i}. The expression pattern associated to this tag is then $(j_1,j_2,\ldots,j_\ell)$.

1. If the user provided the list of minimum useful values, for each $i = start, \dots, \ell$ and for each $k = 1, \dots, t_i$, let $muv_{i,k}$ be the given minimum useful value for the k-th sample of the i-th group. Go to step 2.
If the user chose to provide d instead, let d be as given. If the user provided neither d nor the list of minimum useful values, let $d = 100$. For each $i = start, \dots, \ell$ and for each $k = 1, \dots, t_i$, let $muv_{i,k} = x_{h_0,i,k}$, where $x_{h_0,i,k}$ is such that $d\%$ of the $x_{h,i,k}$'s with $x_{h,i,k} \neq undef$ are greater than or equal to $x_{h_0,i,k}$.

2. For each $i = start, \dots, \ell$, let $muv_i = \frac{\sum_{k=1}^{t_i} muv_{i,k}}{t_i}$. If $start = 1$, let $muv_0 = \frac{\sum_{i=1}^{\ell} muv_i}{\ell}$.

3. Let $\mathcal{U} = \{h : 1 \leq h \leq N$ and $\forall i$ with $start \leq i \leq \ell$ there exists k with $1 \leq k \leq t_i$ and $x_{h,i,k} \neq undef\}$. For each $h \in \mathcal{U}$ and for each $i \in \{start, \dots, \ell\}$, let

$$\overline{x}_{h,i} = \frac{\sum_{x_{h,i,k} \neq undef} x_{h,i,k}}{\sum_{x_{h,i,k} \neq undef} 1}.$$

If $start = 1$, for each $h \in \mathcal{U}$ let $\overline{x}_{h,0}$ be the median of the $\overline{x}_{h,i}$'s over $i = 1, \dots, \ell$.

4. Let $\mathcal{N} = \{h \in \mathcal{U} :$ for each i with $0 \leq i \leq \ell$, $\overline{x}_{h,i} \geq muv_i\}$.

5. Let $m = \min \frac{\sqrt{t_i}\overline{x}_{h,i}}{s_{h,i}}$ and $\varsigma = \max \frac{s_{h,i}}{\sqrt{t_i}}$ where $s_{h,i}$ is the sample standard deviation of the $x_{h,i,k}$ and the min and max are taken over all $i = start, \dots, \ell$ and $h \in \mathcal{N}$ such that $x_{h,i,k} \neq undef$ for each $k = 1, \dots t_i$. For each $h \in \mathcal{N}$, $i = 0, \dots, \ell$, and $k = 1, \dots, t_i$, let $x_{h,i,k} = x_{h,i,k} + shift$ and $\overline{x}_{h,i} = \overline{x}_{h,i} + shift$, where $shift = \max\{0, (7 - m)\varsigma\}$.

6. Let $\mathcal{G} = \{h \in \mathcal{N} : \overline{x}_{h,0} \neq 0\}$. For each $h \in \mathcal{G}$ and for each $i = 1, \dots, \ell$, let $r_{h,i} = \frac{\overline{x}_{h,i}}{\overline{x}_{h,0}}$. Then the list of ratios associated to the h-th gene tag is $(r_{h,1}, r_{h,2}, \dots, r_{h,\ell})$.

7. For each $i = 1, \dots, \ell$, let min_i (resp. max_i) be the minimum (resp. maximum) of $r_{h,i}$ over all $h \in \mathcal{G}$.

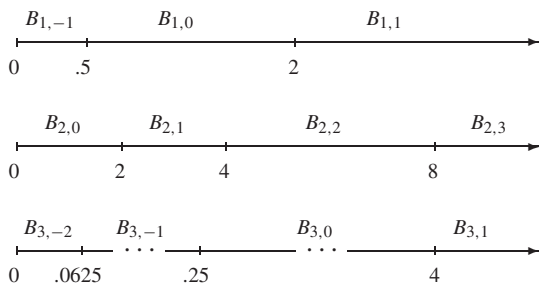
8. For each $i = 1, \dots, \ell$, compute the upper cutratio C_i and the lower cutratio c_i as explained in section Methods and Algorithm.

9. For each $i = 1, \dots, \ell$, the level cutoffs list is obtained by taking all successive powers of C_i which are strictly less then max_i and all successive powers of c_i which are strictly greater than min_i and for which there is at least one smaller non-zero $r_{h,i}$. This is the list $a_{i,1}, a_{i,2}, \dots a_{i,s_i}$, with notation as in section Methods and Algorithm. Let $B_{i,-m_i}$, $B_{i,-m_i+1}$, ..., B_{i,n_i} denote respectively the intervals $[0, a_{i,1})$, $[a_{i,1}, a_{i,2})$, ... , $[a_{i,s_i}, \infty)$, where $B_{i,0}$ is the interval containing 1.

10. For each $h \in \mathcal{G}$ and for each $i = 1, \dots, \ell$, let j_i be such that $r_{h,i} \in B_{i,j_i}$. The pattern attached to h is then $(j_1, j_2, \dots, j_\ell)$.

Fig. 2. The pattern generation algorithm in version I of the software. N is the total number of distinct gene tags, $start = 0$, if a reference group is provided (in which case this is referred as 'group 0'), $start = 1$ otherwise, and the number of non-reference groups is ℓ. For each $i = start, \dots, \ell$, the number of samples in the i-th group is denoted by t_i. For each $h = 1, \dots, N$, $i = start, \dots, \ell$, and $k = 1, \dots, t_i$, $x_{h,i,k}$ is the (normalized) intensity of the h-th gene tag at the k-th sample in the i-th group, if this is given, otherwise $x_{h,i,k} = undef$. Finally, d is such that the top $d\%$ intensity levels are the ones likely to have given hybridization signals above background. Alternatively the user can specify, for each sample, a value ('minimum useful value') such that only intensities above this value are the ones likely to have given hybridization signals above background.

EXAMPLE. $\ell = 3$;
$m_1 = 1$, $n_1 = 1$, $B_{1,-1} = [0, 0.5)$, $B_{1,0} = [0.5, 2)$, $B_{1,1} = [2, \infty)$;
$m_2 = 0$, $n_2 = 3$, $B_{2,0} = [0, 2)$, $B_{2,1} = [2, 4)$, $B_{2,2} = [4, 8)$, $B_{2,3} = [8, \infty)$;
$m_3 = 2$, $n_3 = 1$, $B_{3,-2} = [0, 0.0625)$, $B_{3,-1} = [0.0625, 0.25)$, $B_{3,0} = [0.25, 4)$, $B_{3,1} = [4, \infty)$.



If the list of ratios for the gene tag in question is $(0.2, 1.1, 4.1)$, the expression pattern attached to this tag is $(-1, 0, 1)$, since $0.2 \in B_{1,-1}$, $1.1 \in B_{2,0}$, and $4.1 \in B_{3,1}$.

We now explain how we choose the $a_{i,j}$'s (which we will refer to as *level cutoffs*). To illustrate the basic idea, suppose that we are taking ratios to a reference homotypic group, call it group 0 and let's focus on a fixed group $i, i \geq 1$. Suppose that we have replicate experiments for each of these two groups. We concentrate here on up-regulation; for down-regulation we proceed in an analogous fashion. Our goal is to achieve a certain degree of confidence in the assertion 'this gene is up-regulated at group i as compared to the reference group.' Each gene will have a certain (unknown) distribution of intensities in a group (reference or not), whose mean we will call 'the true mean intensity of the gene at that group.' Denote the random variable giving the intensity of gene g at group j by $X_{g,j}$ and denote the mean andstandard deviation of $X_{g,j}$ by $\mu_{g,j}$ and $\sigma_{g,j}$ respectively. By the statement 'gene g is up-regulated at group i as compared to the reference group' we mean that

$$\frac{\mu_{g,i}}{\mu_{g,0}} > 1.$$

We do not know these true means. We only know the observed intensities of the tags corresponding to g at the available replicates for each of the two groups. Fix such a tag h and denote the average of its observed intensities for group j by $\overline{x}_{h,j}$ and its observed intensity at the k-th replicate of this group by $x_{h,j,k}$. Let $s\%$ be the desired false-positive rate, that is the probability that we say that a gene tag h, corresponding to a gene g, shows up-regulation at group i as compared to the reference group, given that g is not up-regulated (i.e. given that $\frac{\mu_{g,i}}{\mu_{g,0}} \leq 1$). Our goal is to determine a value $C_i > 1$ (which we will call the *upper cutratio for group i*) such that, if we predict that a gene tag h is up-regulated at group i as compared to the reference group when $\frac{\overline{x}_{h,i}}{\overline{x}_{h,0}} > C_i$, then our false-positive rate is expected to be no greater than $s\%$. In order to do this, let $X_{g,j}$ be as above, and let $\overline{X}_{g,j} = (X_{g,j,1} + X_{g,j,2} + \cdots + X_{g,j,t_j})/t_j$, where t_j is the number of available replicates for group j and where the $X_{g,j,k}$'s are independent random variables each with the same distribution as $X_{g,j}$. Thus $\overline{x}_{h,j}$ is an observed value of $\overline{X}_{g,j}$ and $x_{h,j,k}$ an observed value of $X_{g,j,k}$. The false-positive rate is

$$\text{Prob}\left(\frac{\overline{X}_{g,i}}{\overline{X}_{g,0}} > C_i \,\middle|\, \frac{\mu_{g,i}}{\mu_{g,0}} \leq 1\right), \tag{1}$$

where the randomness is coming both from letting g vary over the genes and from the distributions of intensities for the genes. (1) is clearly less than or equal to

$$\text{Prob}\left(\frac{\frac{\overline{X}_{g,i}}{\mu_{g,i}}}{\frac{\overline{X}_{g,0}}{\mu_{g,0}}} > C_i \,\middle|\, \frac{\mu_{g,i}}{\mu_{g,0}} \leq 1\right). \tag{2}$$

Now, we claim that the events $\frac{\frac{\overline{X}_{g,i}}{\mu_{g,i}}}{\frac{\overline{X}_{g,0}}{\mu_{g,0}}} > C_i$ and $\frac{\mu_{g,i}}{\mu_{g,0}} \leq 1$ are independent. Note that the first inequality is the same as $\frac{\overline{X}_{g,i}}{\overline{X}_{g,0}} > C_i \frac{\mu_{g,i}}{\mu_{g,0}}$. Therefore this condition tells us that either $\mu_{g,i}$ is smaller than $\overline{X}_{g,i}$, or $\mu_{g,0}$ is larger than $\overline{X}_{g,0}$, or some combination of the two (how much smaller or larger depends on how large C_i is). In other words the condition is telling us how the sample means compare to the true means. This however should give no information on how the true means relate *to each other*. Therefore the events can reasonably be assumed to be independent. Thus the conditional probability above should equal the non-conditional probability

$$\text{Prob}\left(\frac{\frac{\overline{X}_{g,i}}{\mu_{g,i}}}{\frac{\overline{X}_{g,0}}{\mu_{g,0}}} > C_i\right). \tag{3}$$

So we are now seeking a $C_i > 1$ (as small as possible) such that

$$\text{Prob}\left(\frac{\frac{\overline{X}_{g,i}}{\mu_{g,i}}}{\frac{\overline{X}_{g,0}}{\mu_{g,0}}} > C_i\right) < s\%.$$

If we knew the distributions of $\frac{\overline{X}_{g,i}}{\mu_{g,i}}$ and of $\frac{\overline{X}_{g,0}}{\mu_{g,0}}$ (for g varying over the genes) we could compute such a C_i. But, the $\mu_{g,j}$'s are unknown. We approximate the distribution of $\frac{\overline{X}_{g,j}}{\mu_{g,j}}$ $(j = 0, i)$ for g varying over the genes by the distribution of

$$\frac{\frac{X_{g,j,k}}{\overline{X}_{g,j}} - 1}{\sqrt{t_j - 1}} + 1 \tag{4}$$

for g varying over the genes and k varying over the replicates for group j. In the Appendix we give a justification of this fact, in the case in which the intensity distribution of a gene in a group is close to normal and the distribution of ratios $(\sqrt{t_j})\mu_{g,j}/\sigma_{g,j}$, for g varying over the genes, is concentrated away from zero. The latter condition can be obtained by shifting the intensities appropriately so that the mean is increased while the standard deviation remains the same (see algorithm step 5 in Figure 2). The effect of this shift is discussed in Sections **Comparison with standard statistical analysis**

methods and **Discussion**. As we have investigated in simulations, approximation (4) holds for a much wider class of distributions than just normals (furthermore it improves quickly as t_j, the number of replicates, increases). In all cases of distributions which were not highly asymmetric, approximation (4) held well, and in most of the remaining cases it gave a conservative estimate, in that the approximation tended to have greater dispersion than the actual distribution. This failed most notably when the actual distribution was taken to be exponential, however this was not a relevant case to our applications, nor does it seem likely that it will arise in general.

Using this, we estimate the distribution of (4) from our data, i.e. from the distribution of

$$\frac{\frac{x_{h,j,k}}{\overline{x}_{h,j}} - 1}{\sqrt{t_j - 1}} + 1, \tag{5}$$

for h varying over all gene tags and k varying over the replicates for group j. Of course the more the gene tags and replicates, the better the latter approximation of (4) with an empirical distribution will be. Having a number of observations (number of gene tags $\times$ number of samples in group j) of at least 100 is suggested before adopting such an approximation. We then compute the desired C_i through integration. In particular, if f_j $(j = 0, i)$ is the density function for $\frac{\overline{X}_{g,j}}{\mu_{g,j}}$, and C is fixed, then we evaluate numerically (using the distribution of (5))

$$\int_t \int_{s>Ct} f_0(t) f_i(s) ds dt.$$

If this is above (resp. below) the desired false-positive rate, then C is raised (resp. lowered), and the integral is recalculated. C continues to be adjusted in this way (as a binary search) until the desired false-positive rate is attained. Then C_i is set to the value of C that gives this desired rate.

For down-regulation we proceed in an analogous manner and look for a $c_i < 1$ (as large as possible), which we call the *lower cutratio for group i*, such that

$$\text{Prob}\left(\frac{\frac{\overline{X}_{g,i}}{\mu_{g,i}}}{\frac{\overline{X}_{g,0}}{\mu_{g,0}}} < c_i \,\middle|\, \frac{\mu_{g,i}}{\mu_{g,0}} \geq 1 \right) < s\%.$$

To do this we use approximation (4) again.

Once the C_i's and c_i's have been determined for each non-reference group i, if the ratio r_i of the average intensity of a gene tag at group i and the average intensity of the same gene tag at the reference group (resp. the median) is between C_i and C_i^2 we say that the gene tag is up-regulated one level at this group as compared to the reference group (resp. the median). If r_i is between C_i^2 and C_i^3 then we say that the gene tag is up-regulated two levels as compared to the reference group, etc. A similar approach is taken for down-regulation. Thus we set the $a_{i,j}$'s to be

$$\ldots, c_i^2, c_i, C_i, C_i^2, \ldots.$$

Powers of C_i and c_i are used in order to respect proportions. The algorithm in Figure 2 (step 9) contains the details about how many of these powers we consider for each i.

Note that by having the false-positive rate, one can get a measure of the confidence in the statement that a gene is up-regulated at a group as compared to the reference group. Namely one can estimate the probability Prob(not up $|$predicted up) that a gene is not up-regulated given that we predict it is, by:

$$\begin{aligned} &\text{Prob(not up } \,|\, \text{predicted up)} \\ &\quad = \frac{\text{Prob(not up)}}{\text{Prob(predicted up)}} \text{Prob(predicted up } \,|\, \text{not up)} \\ &\quad \leq \frac{\text{Prob(predicted up } \,|\, \text{not up)}}{\text{Prob(predicted up)}}, \end{aligned}$$

where Prob(predicted up $|$ not up) is the false-positive rate and Prob(predicted up) can be estimated empirically from the data, simply from counting the number of gene tags to which a positive level has been assigned by our protocol for the group under consideration; similarly for down-regulation. Thus, for a fixed false-positive rate of $s\%$, the confidence attained in a positive (resp. negative) level in a pattern depends on the group, i.e. on the position in the pattern, as it depends on Prob(predicted up) for that group.

As a consequence of this approach, when we see a level different from 0, we have a certain confidence in the gene (tag) being up-regulated (if the level is positive) or down-regulated (if the level is negative) as compared to the reference group. However, when we see a 0 there is no confidence implied for the statement that the gene (tag) is not differentially regulated. We can only take 0 to mean that we do not have enough evidence to support a change in level.

Implementation

Software implementing the pattern generation protocol described in the previous section is available from the authors. A brief description of the program follows. Details on usage and input formats are described in the documentation, available from the authors (this documentation also includes a concise summary of our approach).

The program (written in Perl) takes as input a file containing a list of filenames, where each file in the list

gives the results of one experiment. In the input file, the user specifies which experiments are replicates for the same sample type and whether ratios should be taken to the median or to a reference group. Various parameters to be used can be optionally specified, otherwise default values are used. These parameters include the desired false-positive rate $s\%$, a default cutratio to be used when no replicates are available, and either a list of values (called *minimum useful values*), one per sample, or a parameter d to be used in the preprocessing step for certain experiment platforms (e.g. for filter arrays). If the list of minimum useful values is given, this means that the user deems that for each sample only those intensities above the minimum value (s)he provided for that sample are likely to have given hybridization signals above background. The user might specify a single value d instead, if from the knowledge about the experiment procedure, it is deemed that there is a d such that, for any sample in the collection under scrutiny, intensity values which are in the top $d\%$ are those which can be distinguished from background. In the latter case, for each experiment a minimum useful value is then computed using d. Finally a minimum useful value is attached to each group by averaging the minimum useful values of the replicates in that group. In the version of the algorithm given in Figure 2 (version I), only those gene tags with values above the minimum useful value in every group are assigned a pattern. There is another version (version II) of the software which considers every gene tag such that in at least one of the groups its value is above the minimum useful value (and for the groups at which the value is below the minimum useful value, the gene tag's value is raised to be equal to the latter). Thus in version II we account for any signal that might be considered as real while avoiding undue influence of background noise.

The main output of this program consists of an `html` file. The file displays information such as the list of level cutoffs for each group, various counts of interest, and the list of generated patterns. For each such pattern, the gene tags to which that pattern has been assigned are listed. Moreover, the gene tag identifiers in each cluster can be linked to appropriate databases, when specified, to give information on the gene in question (see Section **Results** for an example of this). Sample sections of an output `html` file are shown in Figure 3.

Results

Application to an erythroid development nylon filter dataset

We have used the software described above to analyze filter array experiments of different erythroid development cell samples. These experiments were performed in part to gain insight into the molecular basis for erythroleukemia.

Our erythroid development dataset contains five homotypic groups representing an erythroleukemic cell line and normal cells under different conditions. There are replicate data for each of the groups.

The groups are: CD34 positive cells (human blood progenitor cells including those for erythroid cells; we will write CD34 as a shorthand), human adult and cord erythroblasts (red blood cell precursors), HEL (human erythroleukemia) cells, and HEL cells treated with hemin (induced to express erythroid genes). The HEL group roughly represents a leukemic equivalent of the CD34 group. Similarly, the HEL+hemin group roughly represents a leukemic equivalent of the erythroblast groups (it represents a leukemic cell forced to differentiate and stop growing). The details on the preparation of the cells and on the generation of the data will be published elsewhere along with further biological interpretation of the analysis results (McKenzie *et al.*, manuscript in preparation). Briefly, mRNA was isolated from the cells and reverse-transcribed, then radioactively-labelled by random priming, and interrogated by hybridization to arrays of IMAGE clones using either GenomeSystems GDA filter v1.2 or GDA v1.3 filter 1. The hybridization signals were detected using a Molecular Dynamics Storm PhosphorImager and quantitated by GenomeSystems, or using the Genomic Solutions BioImage software package. Intensities of duplicate spots (on the array) for the same clone (gene tag) were merged into one datum by averaging. Spots which failed visual inspection for artifacts and duplicate spots whose intensities differed by more than two-fold were rejected from further analysis. The signal (clone) intensities were normalized by calculating each signal as percent of the total array signal. Since the sample types in this dataset were closely related cells and a very large number of mRNAs were assessed, it was reasonable to make the assumption that the total mRNA abundance for clones on the filter used would not change considerably from sample to sample. The identifier for each signal was the IMAGE clone ID allowing simple comparisons between the two types of filters used. The IMAGE clone ID was also used to generate names and links for further information on the genes represented on the filter array. More precisely, information on the identity of the gene represented by the clone was provided by DOTS, a database of transcribed sequences (http://www.cbil.upenn.edu/DOTS). The transcribed sequences in DOTS are consensus sequences derived from assemblies of ESTs (including all those available from IMAGE libraries). The transcribed sequences were used to search nucleic acid and protein databases for homology to known genes. The information provided (see Figure 3(b) for an example) is the name of the homologous gene or protein, the database searched, and the percent identity and length of the best high

(a)

The levels for group 1 are: 0, 1, 2
The levels for group 2 are: 0, 1, 2
The levels for group 3 are: 0, 1
The levels for group 4 are: 0, 1, 2

There are 445 gene tags

for group 1 there are 17 up symbols and 0 down symbols
for group 2 there are 7 up symbols and 0 down symbols
for group 3 there are 3 up symbols and 0 down symbols
for group 4 there are 9 up symbols and 0 down symbols

the false positive rate is 0.001

the confidence for the up symbols for group 1 is 0.973823529411765
there are no down symbols for group 1

the confidence for the up symbols for group 2 is 0.936428571428571
there are no down symbols for group 2

the confidence for the up symbols for group 3 is 0.851666666666667
there are no down symbols for group 3

the confidence for the up symbols for group 4 is 0.950555555555556
there are no down symbols for group 4

(b)

Pattern 0,2,1,0 has been attached to the following 3 gene tags:	
245543	135862 1ABY | A Brookhaven Protein Data Bank (nrdb) 66 50 Chain A, Cyanomet Rhb1.1 (Recombinant Hemoglobin) 135862 HT1701 EGAD 95 207 recombination protein A 341567 HT2857 EGAD 88 137 globin, alpha 1
296258	181602 3170176 GenPept 100 114 antigen NY-CO-3
230534	

Fig. 3. Sample sections of the `html` file output by the pattern generating software (version I) applied to the groups: CD34, adult erythroblast, cord erythroblast, and HEL+hemin with ratios to the reference group HEL (described in section 4.1) and parameters $d = 15$ and $s\% = 0.1\%$. (a) Part of the report on the data, containing the list of levels for each group, various counts, and the confidence for each group. (b) Part of the report on one of the generated patterns, displaying the gene tags with that pattern and their descriptions from DOTS, when available.

scoring pair (HSP) using BLAST. No information is provided if no significant matches were found. A link is provided through the IMAGE clone identifier for further information on the DOTS transcribed sequence containing that clone, such as chromosomal map locations for ESTs, cellular roles, source libraries, and protein motifs.

We have applied our pattern-generating software to analyze the groups described above. The available replicates were: two CD34, three adult erythroblasts, two cord blood erythroblasts, three HEL, and two HEL+hemin experiments. With notation as in Section **Implementation**, the value of d was set at 15 reflecting the consideration that only the moderate to highly abundant mRNA classes (greater than or equal to 10 copies per cell, see Zhang *et al.* (1997)) were likely to have given hybridization signals above background on the filter array.

Ratios were generated using the HEL group as reference and the software was run both by merging the adult and cord erythroblasts into one group, the erythroblasts, with five replicates and by keeping them in two separate groups. We merged them when looking for differences between normal and leukemic cells, because the developmental stage was not considered relevant in this case. In the first case, the pattern length is therefore $\ell = 3$ and the three (non-reference) groups were listed in the pattern in the order: CD34, erythroblasts, HEL+hemin. In the second case the pattern length is $\ell = 4$ and the groups were

listed in the order: CD34, adult, cord, HEL+hemin. Both version I and version II of the algorithm were used (see Section **Implementation**). For these data, the running time was always below 90 seconds, when the software was run on a Sun Ultra-10 running an UltraSPARC IIi CPU at 300MHZ with 128MB RAM.

When the adult and cord were merged, out of the 18,123 clones which were present in at least one experiment for each group, 540 were above the minimum useful value in every group and 5063 were above the minimum useful value in at least one group. The value for s was varied. The higher the s value, the richer was the expression description and the lower the confidence in our predictions, when version I was used. (As noted in Section **Methods and Algorithm** however, the confidence depends also on the probability of predicting a gene tag as up-regulated in a group. Therefore, in general, lowering the false positive rate does not necessarily cause an increase in confidence.) For $s\% = 1\%$ there are 5 levels for CD34 (from 0 to 4), 10 levels for the erythroblasts (from -1 to 8), and 6 levels for HEL+hemin (from -1 to 4). When $s\%$ is lowered to 0.1% there are three levels (from 0 to 2) for each of these three groups. When $s\%$ is lowered again to 0.01% there are 2 levels (0 and 1) for each of these groups. The following table illustrates how the confidence in predicted up-regulation changes with s and with the group. After each group the number (out of 540) of clones up-regulated in that group as compared to the reference is also displayed (denoted by '#')

$s\%$	CD34 conf.	#	ery. conf.	#	HEL+h. conf.	#
1%	85%	37	81%	28	89%	47
0.1%	97%	17	92%	7	94%	9
0.01%	98%	3	99%	5	99%	4

Clones representing the same gene were usually found to have identical or very similar patterns, as expected. Furthermore, clones representing genes whose expression is known in these cells presented patterns compatible with what was expected. For example, when version I of the software was run keeping adult and cord separate, the α- and the β-globin clones whose raw intensities are represented in Figure 4, were assigned the patterns (0, 8, 2, 0) and (0, 8, 1, 0) respectively, with moderate to high confidence. Both clones were therefore detected as up-regulated in erythroblasts and more in adult than in cord. This matches what is known about the expression of these genes (see Papayannopoulou *et al.* (1987), Dalyot *et al.* (1992), and Ni *et al.* (1999)), for example the fact that β globin is replaced with fetal γ globin in cord samples.

We can ask what genes are differentially expressed between normal (CD34, erythroblasts) and leukemic cells. From that set, we can then ask which genes are induced by hemin (HEL+hemin) to adopt a normal expression pattern, to be followed up by further experiments. The experiments were analyzed for this purpose using both version I and version II of the software. Different false-positive rates were applied to each case to achieve moderate to high confidence in the assertions of differential expression. These rates for high confidence, and the genes identified as differentially expressed for both cases, are presented in the table in Figure 5. Having more genes available to start with (5063 vs. 540) led to more genes identified as differentially expressed but at lower confidence. At similar confidence levels, starting with more genes did not necessarily lead to more genes identified as differentially expressed between normal and HEL cells as can be seen from the table. There was general agreement for both cases and several candidate genes were identified for further investigation. These include, at moderate confidence, a member of a signal transduction cascade (MAPKK2). Functional studies are required for the next step of this analysis.

Comparison with standard statistical analysis methods

Since the distributions of gene intensities for the groups in the dataset of Section **Application to an erythroid development nylon filter dataset** can be reasonably assumed to be close to normal, we have applied standard statistical analysis methods to this dataset, to compare the results with those of our protocol (we used version I). The standard methods consist in combining tests like the t-test with a 'Bonferroni-like' correction. The Bonferroni correction in itself is too strict, because it is generally used to ensure at high confidence the absence of false positives (see Claverie, 1999). However, a similar correction can be applied to ensure at high confidence that the number of false positives is no larger than $s\% \times$ (number of gene tags). (For example, this correction can be done using a Poisson distribution with parameter (number of gene tags)$\times p$, whereby the significance threshold p that one needs to use in a t-like test done on a gene tag by gene tag basis can be determined.) This puts the standard methods on a footing which allows a fair comparison with our method (further details about this can be found in the documentation available from the authors).

The shift used in our method to be able to apply approximation (4) has certain consequences that should be pointed out. Namely, it causes genes with lower expression levels to require stronger evidence for differential expression than those with higher levels. This is reasonable since, for example, a change from 1 mRNA

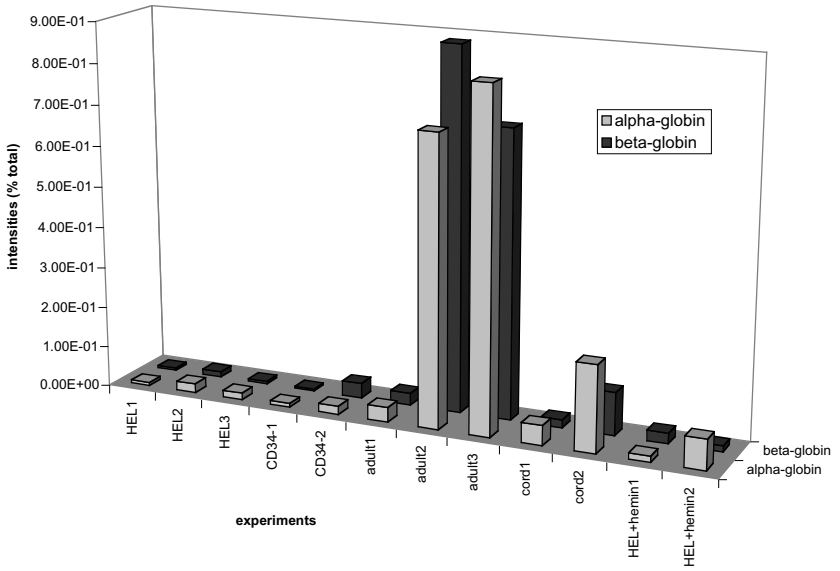


Fig. 4. Input intensity values for an α- and a β-globin clone in the erythroid development dataset. The lists of ratios of the average unshifted intensities of the non-reference groups (CD34, adult, cord, HEL+hemin) to the average unshifted intensity of the HEL (reference) group for the α- and β-globin clones were respectively: (0.965, 34.1, 8.49, 2.93) and (2.51, 66.6, 7.74, 2.5).

copy per cell in group 0 to 2 in group i is less compelling evidence for up-regulation in group i than a change from 100 to 200.

Both for the dataset where adult and cord erythroblasts were merged and for the one where they were kept separate, we compared our results for the various values of $s\%$ used with results obtained by running the standard methods. The comparison was done group by group (in other words, position by position in the patterns). For $s\% = 0.1\%$ and $s\% = 0.01\%$ (such low values are not unreasonable if one desires good confidence in the predictions) the standard methods did not detect any differentially expressed gene tags (as compared to the reference group HEL) in any position. Not even the α- and the β-globin were detected as up-regulated in erythroid cells, as they are known to be. In fact, even for $s\% = 1\%$ the latter were not detected as up-regulated by the standard methods. For $s\% = 1\%$ no gene tags were detected by the standard methods as differentially expressed in any of CD34, or the erythroblasts, whereas our method detected several at high confidence. For the HEL+hemin group the standard methods detected about 20 gene tags as upregulated with high confidence and we detected about 50 with even higher confidence. There was no overlap between these two sets. The gene tags detected by the standard methods had low intensities (as compared to our numerical shift) and we expected to miss those for the reasons explained above. (In Section **Discussion** we pick this point up again in reference to future work.)

Application to a two-channel microarray lymphocyte dataset

We have applied our protocol to analyze some of the experiments published by Alizadeh *et al.* (2000). These consist of two-channel microarray experiments, which use a specialized microarray, named 'Lymphochip.' The following nine experiments were downloaded from http://llmpp.nih.gov/lymphoma/data/rawdata/ and used in our analysis: three samples of normal blood B-cells (lc7b023, lc7b024, lc7b103), four samples of CLL (B-cell chronic lymphocytic leukemias: lc7b047, lc7b048, lc7b069, lc7b070), and two samples of FL (B-cells from follicular lymphoma: lc7b096, lc7b097). This subset was chosen because it provided replicates using the same array version (lc7b) for three closely related groups. For each experiment, only those spots on the array which were not 'flagged' as bad and which were above background (total signal $\geq 1.4 \times$ background, fraction of pixels above background ≥ 0.55 for both channels) were retained, in accordance with the authors own analysis (the resulting number of gene tags was 5595.) The median ratios (MRAT) output from the ScanAlyze software program (http://rana.stanford.edu/software/) relative to a common control sample (a pool of nine lymphoma cell lines) were used to attach a value to each clone (gene tag) by averaging over the spots representing that clone. Finally, for each experiment, the list of values attached to the clones were rescaled to have a median of 1 (this is similar to what was done by Alizadeh *et al.* (2000)). The values

IMAGE clone ID	description	version I $s\% = 0.25\%$	version II $s\% = 0.01\%$
198960	unknown, down in CD34		√
296266	unknown, down in CD34		√
241804	oncogene tre-2, down in CD34		√
144221	β-globin, up in erythroblasts		√
148425		√	√
136255		√	
241976	α-globin, up in erythroblasts	√	√
245543		√	√
296258	NY-CO-3 antigen, up in erythroblasts	√	√
306759	similar to α-2, 3-sialyltransferase, up in erythroblasts		√
230534	unknown, up in erythroblasts	√	√
201839	unknown, up in erythroblasts		√
489440	unknown, up in erythroblasts	√	
245162	ferritin, up in erythroblasts and HEL+hemin	√	
194153	unknown, up in CD34	√	√
116431	unknown, up in CD34	√	√
267479	ribosomal protein S3, up in CD34	√	
203939	splicing factor, up in CD34	√	
233938	kinetochore motor CENP-E, up in CD34	√	
233993	similar to adipogenesis-inhibitory factor, up in CD34	√	
346624	similar to myosin, heavy polypeptide embryonic, up in CD34	√	
116427	similar to RNA helicase, up in CD34	√	
197281	unknown, up in CD34	√	
36332	unknown, up in CD34	√	
36118	unknown, up in CD34	√	
179849	unknown, up in CD34	√	
191910	unknown, up in CD34	√	
193802	unknown, up in CD34	√	
491451	ribosomal protein L27a, up in CD34	√	
207962	similar to H326, up in CD34	√	
210513	KIAA0381, up in CD34	√	
258673	unknown, up in CD34	√	
36821	unknown, up in CD34	√	
347522	unknown, up in CD34	√	
341125	histone H2A.2, up in CD34 and HEL+hemin	√	
171864	similar to neuroD, up in CD34 and HEL+hemin	√	
222633	unknown, up in CD34 and HEL+hemin	√	√

Fig. 5. Differentially-expressed genes between normal (CD34, erythroblasts) and HEL cells. Genes altered upon hemin induction (HEL+hemin) to a normal pattern are indicated. The checkmarks denote those clones which had the differential expression specified in the description column. With version I of the software and a false-positive rate of 0.25%, the confidence for up-regulation in the CD34 group was 94% (no down-regulation), the confidence for up-regulation in the erythroblast group is 83% (no down-regulation), and the confidence for up-regulation of the HEL+hemin group is 78% (no down-regulated genes with acceptable confidence). With version II of the software and a false-positive rate of 0.01%, the confidence for the CD34 group was 83% for up-regulation and 83% for down-regulation, the confidence for the erythroblast group was 93% for up-regulation (no down-regulation), and the confidence for the HEL+hemin group was 87% for up-regulation (no down-regulation).

obtained in this way were the normalized intensities input into our program.

Our program found at high confidence differentially expressed genes between the lymphoma (CLL, FL) and normal blood B-cells despite their high overall similarity in expression profiles (Alizadeh *et al.*, 2000). It also found germinal B-cell associated genes (BCL-6, A-myb) to be up-regulated only in FL and it did not detect cell proliferation genes as differentially expressed in either CLL or FL. These results are in accordance with the major findings for these sample types in Alizadeh *et al.* (2000). In addition, we found (but they did not report) that fos and jun transcription factors (which form the AP-1 complex) were down-regulated at high confidence (93%)

in CLL alone. The inability of tumor necrosis factor (TNF) to induce c-fos and c-jun in CLL cells has been linked to the refractory nature of CLL to stimulation of cell proliferation (Jabbar *et al.*, 1994).

Our approach therefore provides results which are consistent overall with results obtained from clustering and visual inspection methods and at the same time allows us to attach confidence to the predictions made about up- or down-regulation. Moreover it provides new insights into differentially expressed genes.

Finally, we note that standard statistical analysis methods cannot in general be applied to situations when the comparisons are between red-to-green ratios from two separate microarrays (like the situation of these data), as the assumption that the intensities so normalized have a close to normal distribution is no longer valid.

REMARK. For both the dataset of Section **Application to an erythroid development nylon filter dataset** and that of Section **Application to a two-channel microarray lymphocyte dataset** we deemed that the normalized intensities fit in the framework of applicability of approximation (4). For the former dataset, they could be reasonably assumed to have distributions close to normal. For the second, they could be assumed to have distributions which were not highly asymmetric, because of the way the gene tags were selected (denominators in the MRATs having distributions concentrated away from 0).

Discussion

We wanted to formally address the issue of generating descriptive, yet reliable gene expression patterns for datasets containing a few replicate (hybridization array) experiments per sample type and this paper presents a novel protocol in this direction.

Some investigators choose to take the position that there is so much noise in this kind of data that the best that can be done is to divide genes (for each group) into two categories (on or off); i.e. they choose a binary representation of gene expression. This approach seems overly pessimistic and will likely cause one to miss real and interesting changes in expression levels that are represented reliably in the data. Also this approach does not allow for comparisons between different sample types if a gene is turned on in both types. When possible, we give a representation of expression levels which is richer than a binary one. The levels are discrete (as we bin according to the variability of the data). Our protocol will in fact assign binary representations if the data in question are sufficiently noisy to merit that. In the most extreme case, our protocol might even assign only a *single* level to one or more sample types, in case the data are so variable that they cannot give any reliable indications of differential expression.

It is desired to generate patterns in such a way as to have a certain degree of confidence in the predictions made for up- and down-regulation. When many replicates are available, standard (or quasi-standard) statistical methods can be used to detect differentially expressed genes for certain kinds of data (see for example Claverie (1999) and Golub *et al.* (1999)). These methods involve measures of variability that depend on the sample type as well as on the gene. However, for various reasons, chief among which is the current high cost of the experiments, many laboratories generate data which do not have many replicates. Thus one needs to get as much as possible out of just a few replicates. Moreover the ultimate goal is not just to have a false-positive rate on the order of 5% or even 1%, but to have a high confidence in the predictions made. In other words we desire a low probability that a gene is not up- (resp. down-) regulated, given that we predict it is. Since the percentage of differentially expressed genes in two sample types is usually low (in closely related cells this is expected to be 1.5% to 2.5%, see (Sagerstrom *et al.*, 1997)), false-positive rates much lower than these are needed to obtain reasonable confidences. Our method was developed with these two aspects in mind. While few replicates are needed, it should be stressed that, without any replicates (as in the case of a number of recently published two-color microarray studies), one cannot determine the false-positive rate as a function of the criteria by which differential expression is predicted, making a judicious choice of cutoffs difficult to impossible.

It is important to note that the issue at hand is one of gaining confidence in differential expression. Therefore, in the patterns generated with our protocol, when one sees an expression level at a group of 1 or more (resp. -1 or less), there is a certain degree of confidence that the gene is up-regulated (resp. down-regulated) in this group as compared to the reference group. However, if there is an expression level of 0, there is not the same degree of confidence that it is *not* up- or down-regulated as compared to the reference. It is not possible to have confidence about both phenomena at the same time without losing some information. One could, however, develop methods to extract confidence measures of non-differential expression. We have not done so here, but will investigate this in future work.

In the applications discussed in Sections **Application to an erythroid development nylon filter dataset** and **Comparison with standard statistical analysis methods**, our method picked up differentially expressed genes which were not picked up by the standard methods at comparable confidence and which were known to be differentially expressed. In Section **Comparison with**

standard statistical analysis methods we noted that, as an effect of the numerical shift used in our approach to apply approximation (4), genes whose expression is very low as compared to the shift require stronger evidence before we detect differential expression. In some of the runs performed there were a few genes that the standard methods declared as differentially expressed and that we did not. This only occurred though when $s\% = 1\%$ and never occurred for lower values of s. These tended to be genes in the low intensity category (as compared to our shift), which we were not expecting to detect. Plans for future work include the development of protocols to apply our method iteratively to different intensity level subsets of the gene tags under consideration to pick up those genes in the low intensity category that might be true positives and that are missed on the first run (we have some promising preliminary results in this respect). For the cases where the standard methods are applicable, we are also considering ways of combining our method with the standard methods so to lower the false-negative rate of our predictions.

As a final remark, we note that the input to our program consists of normalized intensities and it is up to the user to choose an appropriate normalization based on the way the data have been generated. Moreover, it is also important that the intensity distributions (after normalization) are appropriate for approximation (4) (see Section **Methods and Algorithm** and Remark in Section **Results**).

In summary, we have developed a novel protocol to assign measures of confidence to differentially expressed genes. The protocol takes advantage of the large number of genes typical of hybridization array experiments to control for the variability in gene expression values, using few replicates in an approach that provides low false-positive rates and high confidence. The protocol goes beyond a simple binary description of expression, when justified, and provides discrete descriptions of gene expression patterns based on the desired level of confidence.

Note: Software documentation and further discussion of issues relative to this paper can be found at http://www.cbil.upenn.edu/PaGE.

Appendix

To justify the approximation of the distribution of $\frac{\overline{X}_{g,j}}{\mu_{g,j}}$ $(j = 0, i)$, for g varying over the genes, by the distribution of (4), in the case in which the gene intensities are normally distributed and the distribution of $\rho_{g,j} = (\sqrt{t_j})\mu_{g,j}/\sigma_{g,j}$ (for g varying over the genes) is concentrated over sufficiently large values, we use the following lemma.

LEMMA 1. *Let $X_1, X_2, \ldots X_t$ be a random sample of a normal random variable X with positive mean μ_X and standard deviation σ_X. Let $\overline{X} = (X_1 + X_2 + \cdots + X_t)/t$. Let $k \in \{1, 2, \ldots, t\}$. Then,*

$$\frac{\frac{X_k}{\overline{X}} - 1}{\sqrt{t-1}} = \frac{Z_1}{\rho_X + Z_2} \tag{6}$$

and

$$\frac{\overline{X}}{\mu_X} - 1 = \frac{Z_2}{\rho_X}, \tag{7}$$

where Z_1 and Z_2 are two independent standard normal random variables and $\rho_X = (\sqrt{t})\mu_X/\sigma_X$.

PROOF. From the properties of sums of random variables it is easy to check that the random variables $X_k - \overline{X}$ and $\overline{X}$ are independent and normally distributed with respective means 0 and μ_X and respective standard deviations $\left(\sqrt{\frac{t-1}{t}}\right)\sigma_X$ and $\frac{1}{\sqrt{t}}\sigma_X$ (to check independence it is sufficient to check that the correlation is 0 since the two random variables are normal). So

$$X_k - \overline{X} = \left(\sqrt{\frac{t-1}{t}}\right)\sigma_X Z_1 \tag{8}$$

and

$$\overline{X} = \frac{1}{\sqrt{t}}\sigma_X Z_2 + \mu_X = \frac{\sigma_X Z_2 + (\sqrt{t})\mu_X}{\sqrt{t}}. \tag{9}$$

Dividing both sides of (8) by $\sqrt{t-1}$, then taking the ratio of (8) and (9) and dividing top and bottom of the resulting right hand side by σ_X, we get (6). Since $\overline{X}$ is normal with positive mean μ_X and standard deviation $\frac{1}{\sqrt{t}}\sigma_X$, $\frac{\overline{X}}{\mu_X} - 1$ is normal with mean 0 and standard deviation $1/\rho_X$, thus (7) follows. □

With notation as in Section **Methods and Algorithm**, if we fix a homotypic group, say group j and we apply this lemma to $X = X_{g,j}$, we get that

$$\frac{\frac{X_{g,j,k}}{\overline{X}_{g,j}} - 1}{\sqrt{t_j - 1}} = \frac{Z_1}{\rho_{g,j} + Z_2} \tag{10}$$

and

$$\frac{\overline{X}_{g,j}}{\mu_{g,j}} - 1 = \frac{Z_2}{\rho_{g,j}}. \tag{11}$$

Therefore, if the distribution of $\rho_{g,j}$ (for g varying over the genes) is concentrated away from zero, the distribution of (10) will be a very good approximation to that of (11). When $\rho_{g,j}$ is around 8 the approximation is extremely close.

Acknowledgments

We thank Warren Ewens for many useful conversations and Eric Slud for suggesting the proof given in the **Appendix**. We also thank Brian Brunk for the use of DOTS; Hong Ni, Christopher Orr and Linda Schmidt for their collaboration in generating the erythroid development biological data; and Ash Alizadeh for providing the mapping of clones to array elements for their data. Finally, we thank the referees for their comments.

This work has been supported in part by the NSF training grant BIR 9413215, the NIH grants RO1-RR-04026 and N01 CN 95037, and by the Nemours foundation.

References

Alizadeh,A.A. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Claverie,J.M. (1999) Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet. Sci.*, **8**, 1821–1832.

Chen,Y., Dougherty,E.R. and Bittner,M.L. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics*, **2**, 364–374.

Dalyot,N., Fibach,E., Rachmilewitz,E.A. and Oppenheim,A. (1992) Adult and neonatal patterns of human globin gene expression are recapitulated in liquid cultures. *Exp. Hematol.*, **20**, 1141–1145.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**, 14863–14868.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Jabbar,S.A., Hoffbrand,A.V. and Gitendra Wickremasinghe,R. (1994) Regulation of transcription factors NF kappa B and AP-1 following tumour necrosis factor-alpha treatment of cells from chronic leukaemia patients. *Br. J. Haematol.*, **86**, 496–504.

Ni,H., Yang,X.D. and Stoeckert,C.J.,Jr (1999) Maturation and developmental stage-related changes in fetal globin gene expression are reproduced in transiently transfected primary adult human erythroblasts. *Exp. Hematol.*, **27**, 46–53.

Papayannopoulou,T., Nakamoto,B., Kurachi,S. and Nelson,R. (1987) Analysis of the erythroid phenotype of HEL cells: clonal variation and the effect of inducers. *Blood*, **70**, 1764–1772.

Sagerstrom,C.G., Sun,B.I. and Sive,H.L. (1997) Subtractive cloning: past, present, and future. *Ann. Rev. Biochem.*, **66**, 751–783.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Glolub,T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, **96**, 2907–2912.

Wen,X., Fuhrman,S., Michaels,G.S., Carr,D.B., Smith,S., Barker,J.L. and Somogyi,R. (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA*, **95**, 334–339.

Zhang,L., Zhou,W., Velculescu,V.E., Kern,S.E., Hruban,R.H., Hamilton,S.R., Volgestein,B. and Kinzler,K.W. (1997) Gene expression profiles in normal and cancer cells. *Science*, **276**, 1268–1272.