

A nitty-gritty aspect of correlation and network inference from gene expression data

Lev Klebanov*, Andrei Yakovlev**

*Department of Probability and Statistics, Charles University, Sokolovska 83, Praha-8, CZ-18675, Czech Republic, **Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Avenue, Box 630, Rochester, New York 14642, U.S.A.

Abstract

Background: All currently available methods of network/association inference from microarray gene expression measurements implicitly assume that such measurements represent the actual expression levels of different genes within each cell included in the biological sample under study. Contrary to this common belief, modern microarray technology produces signals aggregated over a random number of individual cells, thereby causing a random effect that distorts the correlation structure of intra-cellular gene expression levels.

Results: This paper provides a theoretical consideration of the random effect of signal aggregation and its implications for correlation analysis and network inference. An attempt is made to quantitatively assess the magnitude of this effect from real data. Some preliminary ideas are offered to mitigate the consequences of random signal aggregation in the analysis of gene expression data.

Conclusions: Resulting from the summation of expression intensities over a random number of individual cells, the observed signals may not adequately reflect the true dependence structure of intra-cellular gene expression levels needed as a source of information for network reconstruction. While our preliminary analysis suggests that in reality the reported effect may not be as extreme as theoretical considerations allow, the main concern still remains and the problem calls for further systematic exploration. The usefulness of inference on genetic regulatory structures from microarray data depends critically on the ability of investigators to overcome this obstacle in a scientifically sound way.

1. Introduction

Inferring gene regulatory networks from microarray data has become a popular activity in recent years, resulting in an ever increasing volume of publications. There are many pitfalls in network analysis that remain either unnoticed or scantily understood. A critical discussion of such pitfalls is long overdue. In the present paper, we discuss one feature of microarray data the investigators need to be aware of when embarking on a study of putative associations between elements of networks and pathways. We believe that the present discussion pinpoints the crux of the difficulty in correlation analysis of microarray data and network inference based correlation measures. The same caveat is of even greater concern in reference to more sophisticated methodologies that are designed to extract more information from the joint distributions of expression signals, Bayesian network inference being a relevant example. Methods of network reconstruction from designed gene perturbation experiments are beyond the scope of this paper. The fact that the latter strategy can be limited to mean expression levels makes it fundamentally different from the inference based on genome-wide expression measurements. Some limitations of the network inference from gene perturbation experiments have been discussed by other authors (see, e.g., [1]). The multiple testing aspect of the problem will not be touched upon either despite its direct bearing on this type of data analysis.

2. Aggregated expression intensities

In a paper published in 2003, Chu et al. [1] pointed out the important fact that the measurements of mRNA abundance produced by microarray technology represent aggregated expression signals and, as such, may not adequately reflect the molecular events occurring within individual cells. To illustrate this conjecture, the authors proceeded from the observation that each gene expression measurement produced by a microarray is of the sum of the expression levels over many cells. Let ν be the number of cells contributing to the observed expression signal U (see Remark 1 below) and denote by X_i the expression level of a given gene in the i th cell. The notation Y_i is used for the second gene in a given pair of genes. A simplistic model of the observed expression signals in this pair is given by

$$U = \sum_{i=1}^{\nu} X_i, \quad V = \sum_{i=1}^{\nu} Y_i, \quad (1)$$

where X_i and Y_i are two sequences of independent and identically distributed (i.i.d.) random variables (r.v.s), while X_i and Y_i in each pair (X_i, Y_i) may be dependent with joint distribution function $F(x, y)$. Limiting themselves to the case where ν is non-random, Chu et al. [1] showed that, except for some very special and biologically irrelevant cases, the Markov factorization admitted by the expression levels within individual cells does not survive the summation (aggregation) in formula (1), thereby stymieing any network inference based on the joint distribution. The importance of this observation cannot be emphasized enough. However, as apparent from the relevant literature, it went entirely unnoticed.

In their concluding remarks, Chu et al. [1] note that the mean vector and covariance matrix remain “invariant under aggregation up to a simple linear transformation”. The same is obviously true for the correlation matrix. They saw some hope in that fact as reflected in the following quote from their paper: “Thus, while waiting for the technologies capable of measuring efficiently the expression levels in single cells, in experimental studies, we can still make valid – although probably more limited – inferences about the regulatory networks based only on the first two moments of the joint distribution and the independence relations.”

Unfortunately, this hope is deflated when considering the case of random ν . Indeed, let each X_i have the same distribution as X , while each Y_i is distributed as Y . Then the following formula holds for the correlation coefficient $\rho(U, V)$ between U and V :

$$\rho(U, V) = \frac{\mu_\nu \text{Cov}(X, Y) + \sigma_\nu^2 \mu_x \mu_y}{\sqrt{\mu_\nu \sigma_x^2 + \sigma_\nu^2 \mu_x^2} \sqrt{\mu_\nu \sigma_y^2 + \sigma_\nu^2 \mu_y^2}}, \quad (2)$$

where $\mu_\nu = \mathbb{E}(\nu)$, $\mu_x = \mathbb{E}(X)$, $\mu_y = \mathbb{E}(Y)$, $\sigma_\nu^2 = \text{Var}(\nu)$, $\sigma_x^2 = \text{Var}(X)$, $\sigma_y^2 = \text{Var}(Y)$, and $\text{Cov}(X, Y)$ is the covariance between X and Y . Formula (2) can be represented as

$$\rho(U, V) = \frac{\rho(X, Y)}{\sqrt{1 + a^2 \tau}} \frac{1}{\sqrt{1 + b^2 \tau}} + \frac{\tau ab}{\sqrt{1 + a^2 \tau} \sqrt{1 + b^2 \tau}}, \quad (3)$$

where $\tau = \sigma_\nu^2 / \mu_\nu$, $a = \mu_x / \sigma_x$, $b = \mu_y / \sigma_y$, and $r = \rho(X, Y)$ is the coefficient of correlation between X and Y . Therefore, $\rho(U, V) = \rho(X, Y)$ if and only if $\sigma_\nu = 0$.

Remark 1. If the hybridization reaction reaches equilibrium, an assumption widely adopted in the physical chemistry of microarrays [2], the random variable (r.v.) ν can be interpreted as the total number, N , of cells from which the total RNA is extracted. In the practical use of

microarray technology, however, the reaction is typically stopped before equilibrium has been reached. In the latter case, the r.v. ν represents the number of cells that collectively yield the ultimate number of bound target-probe duplexes. Therefore, the random parameter ν is unobservable and should be thought of as a virtual number of cells associated with each batch of target RNA produced by them. This notion provides a constructive way of bridging the processes of gene expression at the genomic and tissue levels, which is the main thrust of our discussion. The conventional protocol of a microarray experiment implies that it is the total amount of RNA that is controlled (kept constant) across the arrays (subjects) rather than the number of cells ending up on each array. Therefore, the random fluctuations of ν cannot be controlled directly. Even if a tight control of N could be provided in experiments, it is unclear whether this would have had a diminishing effect on the variance of ν .

An upper bound for the deviation between $\rho(U, V)$ and $\rho(X, Y)$ is given by

$$|\rho(U, V) - \rho(X, Y)| \leq \frac{1}{2}\tau \left((a+b)^2 + a^2b^2\tau \right). \quad (4)$$

This result follows from formula (3) and the following chain of inequalities:

$$\begin{aligned} |\rho(U, V) - \rho(X, Y)| &\leq \left| \rho(X, Y) \left(1 - \frac{1}{\sqrt{1+a^2\tau}\sqrt{1+b^2\tau}} \right) - \frac{\tau ab}{\sqrt{1+a^2\tau}\sqrt{1+b^2\tau}} \right| \\ &\leq 1 - \frac{1}{\sqrt{1+a^2\tau}\sqrt{1+b^2\tau}} + \tau ab \leq \frac{(a^2+b^2)\tau + a^2b^2\tau^2}{\sqrt{1+a^2\tau}\sqrt{1+b^2\tau}(\sqrt{1+a^2\tau}\sqrt{1+b^2\tau} + 1)} + \tau ab \\ &\leq \frac{1}{2} \left((a^2+b^2)\tau + a^2b^2\tau^2 \right) + \tau ab = \frac{1}{2}\tau \left((a+b)^2 + a^2b^2\tau \right). \end{aligned}$$

Recall that the equality $\rho(U, V) = \rho(X, Y)$ holds when $\tau = 0$. Considering $R = \rho(U, V)$ as a function of τ , one can verify that $R(\tau)$ either increases monotonically or attains a minimum before starting to increase with increasing τ . In both cases, $R \rightarrow 1$ when $\tau \rightarrow \infty$. The function $R(\tau)$ is smooth at $\tau = 0$, but its initial slope may be quite high as our sample computations show. An additional quantitative insight into the potential impact of this unobservable variation on the correlation structure of microarray data is possible as described in Section 4.

Remark 2. Like Chu et al. [1], our paper considers the usual notion of correlation as a characteristic of the joint distribution of two r.v.s. Whenever the r.v.s are directly observable, a

consistent estimator of the population correlation coefficient is given by its empirical counterpart known as the Pearson correlation coefficient (PCC). We would like to warn against the intentional use of highly heterogeneous data sets in the analysis of regulatory relationships among genes, even if such relationships are perceived as merely statistical associations and not causal effects. The most widely-used approach to inferring genetic regulatory structures is to collect microarray data from different sources of tissues (sometimes even from different species) and identify co-expressed genes from this mixed set of data treating it as a sample in the statistical sense, i.e., as a collection of i.i.d. random vectors. For example, one specific data set of this type includes 101 samples from 43 different human tissues and three cell lines [3]. Some of such observations may be replicated, i.e., represented by arrays obtained from different subjects (usually in small numbers), and some may be represented by only a single array. The well-known Novartis Gene Atlas represents one of the most extreme examples with only one array per each tissue type. The strength of association of gene expression levels is frequently measured by simply calculating the PCC from the pooled data set as recommended by the originators of the relevance network concept [4]. Even if every group (tissue type) includes many arrays, such heterogeneous data sets are not amenable to correlation analysis. This follows from the fact that the compounded correlation coefficient, i.e., the population characteristic ρ_c to which the PCC converges in large samples, is a function of many parameters such as the within-group first and second moments of the marginal distributions, covariances, and expected proportions of each group in the mixture. These parameters are not uniquely determined by ρ_c and the PCC can no longer be thought of as an estimator for a clearly defined measure of dependence. In particular, the hypothesis $H_0 : \rho_c = 0$ becomes meaningless from the statistical standpoint.

3. An alternative representation of $\rho(X, Y)$ and its implications

Recalling the model given by (1), we derive a formula that allows us to better understand the principal difficulty brought about by the random nature of the parameter ν . Let us find the covariance between the unobservable r.v.s $\frac{1}{\nu}U$ and $\frac{1}{\nu}V$. We have

$$\text{Cov}\left(\frac{1}{\nu}U, \frac{1}{\nu}V\right) = \mathbb{E}\left(\frac{1}{\nu}\sum_{s=1}^{\nu}X_s\frac{1}{\nu}\sum_{k=1}^{\nu}Y_k\right) - \mathbb{E}\frac{1}{\nu}\sum_{s=1}^{\nu}X_s\mathbb{E}\frac{1}{\nu}\sum_{k=1}^{\nu}Y_k$$

$$\begin{aligned}
&= \sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{s=1}^n \sum_{k=1}^n \mathbb{E}(X_s Y_k) \mathbb{P}\{\nu = n\} \\
&\quad - \sum_{n=1}^{\infty} \frac{1}{n} \sum_{s=1}^n \mathbb{E}X_s \mathbb{P}\{\nu = n\} \sum_{n=1}^{\infty} \frac{1}{n} \sum_{k=1}^n \mathbb{E}Y_k \mathbb{P}\{\nu = n\}, \tag{5}
\end{aligned}$$

where \mathbb{P} and \mathbb{E} are the symbols of probability and expectation, respectively. Consider the first term

$$\begin{aligned}
&\sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{s=1}^n \sum_{k=1}^n \mathbb{E}(X_s Y_k) \mathbb{P}\{\nu = n\} = \sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{s=1}^n \sum_{k=1}^n \mathbb{E}X_s \mathbb{E}Y_k \mathbb{P}\{\nu = n\} \\
&+ \sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{s=1}^n \text{Cov}(X_s, Y_s) \mathbb{P}\{\nu = n\} = \mathbb{E}X \mathbb{E}Y + \text{Cov}(X, Y) \sum_{n=1}^{\infty} \frac{1}{n} \mathbb{P}\{\nu = n\}.
\end{aligned}$$

For the second term in (5), we have

$$\sum_{n=1}^{\infty} \frac{1}{n} \sum_{s=1}^n \mathbb{E}X_s \mathbb{P}\{\nu = n\} \sum_{n=1}^{\infty} \frac{1}{n} \sum_{k=1}^n \mathbb{E}Y_k \mathbb{P}\{\nu = n\} = \mathbb{E}X \mathbb{E}Y.$$

It follows from formula (5) that

$$\text{Cov}\left(\frac{1}{\nu}U, \frac{1}{\nu}V\right) = C \cdot \text{Cov}(X, Y),$$

where

$$C = \sum_{n=1}^{\infty} \frac{1}{n} \mathbb{P}\{\nu = n\} = \mathbb{E}\left(\frac{1}{\nu}\right)$$

is an unknown constant. Considering the variances of U/ν and V/ν in a similar way, we arrive at the following important assertion:

$$\rho\left(\frac{1}{\nu} \sum_{i=1}^{\nu} X_i, \frac{1}{\nu} \sum_{i=1}^{\nu} Y_i\right) = \rho(X, Y). \tag{6}$$

This formula implies that estimating the correlation between the unobservable variables X and Y in each gene pair amounts to estimating the correlation between their averages over a random number of cells, thereby showing the earlier-mentioned nonidentifiability aspect of the problem in terms of the basic random variables. Note that the model given by (1) can be represented as

$$U = \nu \left(\frac{1}{\nu} \sum_{i=1}^{\nu} X_i\right) = \nu \bar{X}, \quad V = \nu \left(\frac{1}{\nu} \sum_{i=1}^{\nu} Y_i\right) = \nu \bar{Y},$$

where the correlation between \bar{X} and \bar{Y} is the same as that between X and Y , albeit the distributions of the corresponding vectors can be arbitrarily dissimilar. The above representation shows that the r.v. ν can be interpreted as a multiplicative random noise as long as the main focus is on pairwise correlations. However, this interpretation is to no avail. The noise ν and the signals \bar{X} and \bar{Y} are inherently dependent under this model. Therefore, the popular model of independent random effect is unlikely to serve a good approximation to the aggregated signals. In Section 6, we will invoke formula (6) in our discussion of the utility of the Law of Large Numbers within the framework of model (1).

Formula (6) also illustrates one restrictive assumption behind the model that may have gone unnoticed in its construction. Specifically, the assumption that (X_i, Y_i) are i.i.d. random vectors implies exchangeability of these vectors across cells and subjects so that the joint distribution of (X, Y) exhaustively describes both types of variability in formula (1). Put another way, the baseline joint distribution of expression levels of all genes introduced at the cellular level is implicitly compounded with respect to a latent random parameter describing the inter-subject variability. In this case, the correlation between expression signals within each cell appears to be the same as the correlation between their random averages (as formula (6) shows), both correlations being computed across subjects. If one wants to separate the two types of biological variability in a mechanistic model, e.g., by incorporating a random effect into the expression signals associated with single cells and thus making them dependent within each subject, the resultant formulas will become quite cumbersome and contain additional unobservable parameters.

4. Assessing the effect of signal aggregation

While our discussion at the end of the previous section suggests that model (1) is quite simplistic, we presently have no better vehicle to assess the potential deviation of the correlation between X and Y from that between U and V . To gain an idea of how strong the effect of the parameter ν variability can be, let us first compute the coefficient $R = \rho(U, V)$ for some parameter values, assuming that gene expressions within single cells are stochastically independent ($\rho(X, Y) = 0$). By way of example, suppose $\sigma_\nu/\mu_\nu = 0.23$ and $\mu_\nu = 2 \times 10^5$ cells. From formula (3), we obtain $R = 0.999942$ for $a = 1, b = 2$ and $R = 0.999952$ for

$a = 1, b = 5$. When setting $\rho = 0.5$ or $\rho = 0.9$, the values of R change only in the fifth digit. The same magnitude of R still stands even when $\rho = -0.9$. Notwithstanding arbitrariness of the chosen parameters, this indicates an extremely serious problem arising in studies of dependence structures in general and regulatory networks in particular.

Do our calculations imply that the true correlations between gene expressions are absent or weak? The answer is definitely “No” for the following three reasons. First, the assumption of gene independence is biologically implausible and in conflict with a large body of independent experimental evidence, including the known effects of noncoding RNAs and involvement of genes in biochemical pathways. Second, the situation observed in real data is not as severe as in our sample computations: positive correlations tend to be lower and even a small proportion of negative correlations has been documented. It would appear reasonable that many strong negative correlations are hidden in the overwhelmingly positive correlation structure of microarray data. Third, the unobservable parameters chosen in our computations may be very far from reality. Therefore, we have to base our assessment on real gene expression data rather than imaginary parameters of the model. One possible approach to real data analysis is presented below.

Remark 3. It should be noted that negative correlations are typically much more prevalent in normalized versus not normalized data. This does not mean, however, that the commonly used normalization procedures can restore the true correlations. A profound effect of such procedures on the correlation structure of microarray data is well-documented [5, 6]. This effect is hardly beneficial as normalization procedures distort the aggregated signals in an unpredictable way [7] and interfere in the true correlation structure [5]. There are also other theoretical reasons for the fact that data normalization does not provide a satisfactory solution to the problem; these reasons will be discussed at length in another paper.

From formula (2) it follows that

$$\rho(X, Y) = \frac{\rho(U, V)\sigma_u\sigma_v - z_\nu^2\mu_u\mu_v}{\sqrt{(\sigma_u^2 - z_\nu^2\mu_u^2)(\sigma_v^2 - z_\nu^2\mu_v^2)}}, \quad (7)$$

where $z_\nu = \sigma_\nu/\mu_\nu$. As a function of z , the coefficient $\rho(X, Y)$ either decreases monotonically

or attains a maximum at the point

$$z^* = \frac{\sqrt{2ab - a^2R - b^2R}}{\sqrt{a^3b + ab^3 - 2a^2b^2R}}, \quad (8)$$

where

$$R = \rho(U, V), \quad a = \frac{\mu_u}{\sigma_u}, \quad b = \frac{\mu_v}{\sigma_v}.$$

Therefore, the effect of signal aggregation is not unidirectional - the correlation coefficient $\rho(X, Y)$ may be smaller or higher than the observed coefficient $\rho(U, V)$. Formula (7) can be represented in a more concise form

$$\rho(X, Y) = \frac{\rho(U, V)\xi_u\xi_v - z_\nu^2}{\sqrt{(\xi_u^2 - z_\nu^2)(\xi_v^2 - z_\nu^2)}}, \quad (9)$$

where $\xi_u = \sigma_u/\mu_u, \xi_v = \sigma_v/\mu_v$ are the corresponding variation coefficients.

All the parameters entering formulas (7) or (9) can be estimated from microarray data except for z_ν , which is unobservable. However, there are natural mathematical constraints that must be imposed on z_ν . First of all, we have to require that $z_\nu < \xi_u$ for any gene, i.e.,

$$z_\nu < \min_{1 \leq j \leq m} \xi_{u_j}, \quad (10)$$

where $\xi_{u_j}, j = 1, \dots, m$, is the variation coefficient for the j th gene and m is the total number of genes. However, condition (10) does not ensure that $|\rho(X, Y)| \leq 1$. To meet the second condition, we derive from (7) the following requirement:

$$z_\nu^2 \leq \frac{\sigma_u^2\sigma_v^2[1 - \rho^2(U, V)]}{\text{Var}(\mu_u V - \mu_v U)}, \quad (11)$$

for all pairs of genes simultaneously.

The above conditions allow us to deduce a realistic range of possible values of the unobservable variation coefficient z from a specific set of microarray data. If $\rho(X, Y)$ appears to be a monotonically decreasing function of z_ν , which property can be verified with real data, then we can use formula (7) to estimate its maximal deviation from $\rho(U, V)$ by evaluating $\rho(X, Y)$ at the right extreme of z_ν yielded by conditions (10) and (11). In this case, we obtain a reasonably realistic upper estimate of the actual effect of signal aggregation in accordance with model (1). If $\rho(X, Y)$ passes through a maximum as a function of z_ν , this estimate will become

conservative to shifts towards lower values of the true correlation coefficients. Such estimates need to be produced for all gene pairs, of course. More accurate estimates of the effect in both directions (up and down) can be obtained by evaluating the behavior of $\rho(X, Y)$ over the whole range of admissible values of z_ν in each gene pair, but this approach is computationally extremely expensive and requires parallel computations.

The mean and minimal (across genes) variation coefficients of gene expression were estimated from the following five sets of microarray data:

BCC: Breast cancer cells cultured *in vitro* (represented solely by "vehicle" control samples that were treated with the medium used to solubilize the inhibitor) with HG_U133A Affymetrix Chip used to produce microarray measurements [8];

TELL and HYPERDIP: two types of childhood leukemia, U95A Affymetrix Chip [9];

PCTUM: prostate cancer, U95Av2 Affymetrix Chip [10];

PCNORM: normal prostate tissue obtained from prostate cancer patients, U95Av2 Affymetrix Chip [10].

The results are shown in Table 1. These estimates are consistent with the earlier reported observation that the variation coefficient of gene expression is virtually constant across genes [16]. Using the above-described approach, we analyzed all gene pairs in the HYPERDYP data set reporting expression levels of $m=7084$ genes for $n = 88$ patients with a specific type of childhood leukemia. In this case, Table 1 offered $\min_{1 \leq j \leq m} \xi_{u_j}^2 = 0.044$ as an upper bound for z_ν^2 . A more accurate estimate of 0.041 was given by inequality (11). Therefore, we used the latter value as the conservative estimate of z_ν^2 when computing the correlation coefficient $\rho(X, Y)$ by formula (7). Testing for monotonicity was performed by partitioning the admissible range of z_ν^2 (given by condition (11)) into four intervals and using formula (7) to compute the corresponding increments of $\rho(X, Y)$ for each interval. If at least one increment happened to be positive in a given pair, this event was recorded as a "monotonicity violation". There were less than 0.2% of all gene pairs that could be suspected for such violations in the HYPERDYP data. While this frequency of monotonicity violation may be reckoned as quite small, it should be kept in mind that possible shifts in $\rho(X, Y)$ towards values higher than the observed $\rho(U, V)$ were entirely ignored in this analysis.

Let us now evaluate the numerical results of this study. For the HYPERDIP data set, the

mean (over all gene pairs) value of $\rho(U, V)$ is 0.904 and the corresponding standard deviation equals 2.34×10^{-5} . For the unobservable coefficient $\rho(X, Y)$ these parameters are 0.797 and 4.16×10^{-5} , respectively. The total number of gene pairs with negative values of $\rho(U, V)$ is only 9442. The number of negative values of $\rho(X, Y)$ is much larger; it equals 223,826 in the data set under study. To gain a better idea of how dissimilar $\rho(U, V)$ and $\rho(X, Y)$ may be, it is worth estimating the mean and standard deviation (across all gene pairs) of the relative deviation

$$\Delta_\rho = \left| \frac{\rho(X, Y) - \rho(U, V)}{\rho(U, V)} \right|. \quad (12)$$

The resultant estimates are 0.154 and 8×10^{-4} , respectively. This does not strike us as a formidable relative difference. However, two caveats are in order here. First, the above estimates are not very stable. If we replace $z_\nu^2 = 0.041$ with $z_\nu^2 = 0.035$, the mean value of Δ_ρ falls to 0.112, while the number of gene pairs with negative values of $\rho(X, Y)$ goes down to 109,574. Second, the model (1) used for assessing the deviation Δ_ρ may still be overly simplistic as discussed in the previous section. Much more research needs to be done, both theoretically and experimentally, to shed more light on this methodological difficulty.

5. Signal aggregation and technical noise

Our estimates in Table 1 and those resulted from condition (11) give only a rough idea of the magnitude of σ_ν/μ_ν , and making them more accurate is highly desirable. We discuss one possibility to attain these ends in the present section. Consider an experimental design that supposedly eliminates the biological variation, thereby yielding information on measurement errors only. Suppose that a sample of n arrays is available that consists solely of technical replicates representing gene expression measurements taken from one and the same subject. Proceeding from the traditional multiplicative noise model,

$$\tilde{X}_j = \epsilon_j X_j, \quad j = 1, \dots, m$$

where m is the total number of genes (probe sets), \tilde{X}_j is the observed random signal, and ϵ_j is an independent random technical (both gene- and array-specific) noise, one would model this situation as

$$\tilde{X}_j = \epsilon_j C_j, \quad j = 1, \dots, m \quad (13)$$

where C_j are nonrandom constants. If the expression levels are log-transformed, we have

$$\log \tilde{X}_j = \log \epsilon_j + \log C_j.$$

Therefore,

$$\text{Var}(\log \tilde{X}_j) = \text{Var}(\log \epsilon_j),$$

so that, relying on model (13), one can measure the variance, $\text{Var}(\log \epsilon_j)$, of the log-transformed technical noise directly from technical replicates. In particular, one can estimate the variance of the mean noise across all probe-sets, i.e.,

$$\sigma_{\bar{\epsilon}}^2 = \text{Var} \left\{ \frac{1}{m} \sum_{j=1}^m \log \epsilon_j \right\}.$$

We resorted to the above line of reasoning in [11] when reanalyzing the Microarray Quality Control Study (MAQC) [12]. For this data set, the estimated $\sigma_{\bar{\epsilon}}$ is equal to 0.09, which is slightly smaller than the mean (across probe-sets) of estimated standard deviations reported in [11]. Since the overwhelming majority of genes have typically much larger (> 0.3) standard deviations of their log-expression signals in biological replicates (different subjects), this level of technical noise can be deemed negligibly small. This estimate also leads us to conclude that the true correlation between the unobservable signals $\log X_j$ is really strong. Indeed, the contribution of $\text{Var}\{\frac{1}{m} \sum_{j=1}^m \log X_j\}$ to the variance of log-expressions observed in biological data is much larger than the contribution of $\text{Var}\{\frac{1}{m} \sum_{j=1}^m \log \epsilon_j\}$ estimated independently from the MAQC data, while a strong correlation between true biological signals (i.e., their values in the absence of measurement errors) is the only explanation for such a discrepancy when the number m of genes is very large. This also explains why the Law of Large Numbers (LLN) is not met in microarray data when applied to log-expression levels across genes [13, 14].

The situation is no longer the same when we proceed from model (1) in an effort to measure the technical noise stemming from the random nature of the parameter ν . For any gene j , formula (1) gives

$$U_j = \sum_{i=1}^{\nu} X_{ij}, \quad j = 1, \dots, m, \quad (14)$$

and it is the parameter ν that plays the role of the technical noise here. It is clear from (14) that the biological variability cannot be entirely eliminated from gene expression signals

even when they are produced by purely technical replicates. Designed to assess the technical variability, the experiment described above may only reduce the variance of the r.v.s X_{ij} by eliminating the inter-subject variability, but there will always be some residual biological variability associated with different cells. Under such experimental conditions, we have

$$\tilde{U}_j = \sum_{i=1}^{\nu} \tilde{X}_{ij}, \quad j = 1, \dots, m \quad (15)$$

where \tilde{X}_{ij} are i.i.d. r.v.s representing the expression levels of the j th gene in different cells obtained from the same subject and their common (conditional) variance is expected to be smaller than that of X_{1j} in (14). Formula (15) also suggests that the MAQC data are far from ideal for the purposes of noise assessment because the technical replicates in this study were produced from a mix of many dissimilar tissue sources; this heterogeneity of samples may inflate the variance of \tilde{X}_{ij} while it should be kept as low as possible.

To remove the scaling factor C_j from model (13), when deriving the variance of its noise component, we log-transformed the observed expression signals \tilde{X}_j . This trick does not work for model (15) and this significantly complicates the noise assessment. More complications arise when extending the model represented by formula (15) to include an additive term that describes sources of technical variation other than ν . Under the extended model, the original expression level of the j th gene in technical replicates is given by

$$\tilde{U}_j = \sum_{i=1}^{\nu} \tilde{X}_{ij} + \eta.$$

In the presence of the noise component attributable to ν , the error term η does not need to be array-specific as it essentially reflects the equipment-related optical noise.

Since the overall variance of \tilde{U}_j is expected to be much lower than that of U_j [11], one can use technical replicates to make the range of admissible values of z_ν (see Section 4) much narrower, thereby providing more accurate estimates of $\rho(X, Y)$ and Δ_ρ in accordance with formulas (9) and (12), respectively. This idea of combining information from biological and technical replicates deserves careful consideration and even a generous investment in specially-designed experiments because it offers an improved experimental protocol that may make the usual correlation analysis, as well as the network inference based on correlation measures, more meaningful. If the idea works in general, the new protocol will require producing a separate

set of technical replicates in each biological experiment in order to estimate the range of z_ν and then using this estimate to reconstruct $\rho(X, Y)$ in each gene pair. This suggestion is based on a plausible assumption that the variation coefficient z_ν is the same for the biological and technical replicates produced by a given biological experiment. To make the estimate of the range of z_ν as accurate as possible, it is imperative that the technical replicates be produced from a homogeneous biological material derived from the same tissue (initially collected from several subjects) that is used to produce the corresponding biological replicates. While more laborious and expensive, the experiments thus designed may provide a practically workable solution to the problem discussed in the present paper. Some additional thoughts of this kind are offered in Section 7.

6. The law of large numbers and random summation

The following claims seem to be natural in the context of the model given by formulas (1):

1. The observed expression signal U is a result of summation of the inter-cellular signals X_i over a random number of cells ν , thereby defining the basic model structure represented by formulas (1). The random summands X_i are i.i.d. positive r.v.s independent of ν .
2. While the r.v. ν is nondegenerate, it tends to take on large values with high probability because the number of cells is expected to be large.

In what follows, we examine some indirect corroborative evidence for the above claims.

Suppose for a moment that the number of summands $\nu = k$ is nonrandom. Then the distribution of the corresponding sum in (1) is M -divisible, i.e., it can be represented as the convolution of M distribution functions. In this particular case, the fourth central moment $\mu_4(U)$ satisfies the inequality [18]:

$$\mu_4(U) \geq \left(3 - \frac{2}{M}\right)\sigma_u^4. \quad (16)$$

For infinitely divisible distributions, condition (16) assumes the form

$$\mu_4(U) \geq 3\sigma_u^4. \quad (17)$$

Under mild conditions, these inequalities hold in the case of random ν as well [18]. If inequality (16) is met in real biological data, this fact will lend additional support to the presence of

signal summation in microarray technology. When testing the corresponding inequalities for empirical counterparts of the moments $\mu_4(U)$ and σ_u in (16) and (17), we observed the event of their violation to be of relatively rare occurrence. For example, inequality (17) was violated for 18.6% of the 7084 genes in the HYPERDIP data. As expected, this proportion was lower for any finite M in (16). Although there is no objective criterion for declaring this frequency consistent with the property of infinite divisibility, we deem it quite low in view of the fact that $\mu_4(U)$ and σ_u in (17) were replaced with their sample counterparts. To corroborate our perception, we generated 7000 independent samples of size $n = 88$ from a log-normal distribution with parameters $\mathbb{E}(\log U) = 0.7$ and $\text{Var}(\log U) = 0.09$. The experiment was repeated 1000 times. The mean proportion of “inconsistent” cases was equal to 23.3%, suggesting that the random chance of the event under observation may be high even when the underlying distribution is known to be infinitely divisible.

Yet another underpinning for the presence of signal summation is provided by considering the accompanying distributions of random sums. In the classical summation scheme, the notion of accompanying infinitely divisible distributions was introduced by Gnedenko [19]. This idea was later extended to the random summation by Klebanov and Rachev [20]. Consider the random sum

$$U_p = \sum_{i=1}^{\nu_p} X_i, \quad (18)$$

where $\{\nu_p, p \in \Theta\}, \Theta \subset (0, 1)$ is a family of positive integer-valued r.v.s independent of $X_i, i \geq 1$. The r.v. ν_p is assumed to have finite expectation equaling $1/p$ for all p . It is known that the random sum U_p can be approximated by its accompanying ν -infinitely divisible random variable S_p under the condition of non-negativity of X_i only. In this case, it can be shown [21] that the Laplace transform of S_p converges to the Laplace transform of U_p in the uniform metric as $p \rightarrow 0$.

Since the number of cells ν is expected to be large, it is tempting to apply the Law of Large Numbers (LLN) to the normalized random sum

$$Z_k = \frac{1}{\nu_k} \sum_{i=1}^{\nu_k} X_i, \quad (19)$$

where $k \in \mathcal{N}$, and make some predictions based on its behavior as $\nu_k \rightarrow \infty$ ($k \rightarrow \infty$) in probability. As before, we will assume that the sequence of nonnegative integer-valued r.v.s

ν_k is independent of $X_i, i \geq 1$, and $\nu_k \rightarrow \infty$ (in probability) as $k \rightarrow \infty$. The continuous r.v.s X_i are i.i.d. and positive. If μ_x is finite, it is known [22] that $Z_k \rightarrow \mu_x$ as $\nu_k \rightarrow \infty$, with both limit relations holding in probability as $k \rightarrow \infty$. This is the LLN for random sums.

There is no way of ascertaining whether the LLN is met in real microarray data because the r.v. ν is unobservable. However, we intend to use this powerful tool to predict certain properties of expression signals and then verify them with real data. In doing so, we rely on the following simple result.

Assertion. *Under the above conditions, the random vector $\mathbf{Z}_k = Z_{1k}, \dots, Z_{mk}$, with its components defined by*

$$Z_{jk} = \frac{1}{\nu_k} \sum_{i=1}^{\nu_k} X_i^{(j)} = \frac{U_k^{(j)}}{\nu_k}, \quad j = 1, \dots, m, \quad (20)$$

converges in distribution (\xrightarrow{d}) to a degenerate random vector as $k \rightarrow \infty$.

Proof. Denote by $f(\mathbf{t}) = f(t_1, \dots, t_m)$ the multivariate characteristic function (c.f.) of $(X_1^{(1)}, \dots, X_1^{(m)})$. Then the c.f. of $\mathbf{Z}_k = \mathbf{U}_k/\nu_k$, where $\mathbf{U}_k = U_k^{(1)}, \dots, U_k^{(m)}$, is given by

$$\mathbb{E}e^{i(\mathbf{U}/\nu_k, \mathbf{t})} = \sum_{n=1}^{\infty} f^n(t_1/n, \dots, t_m/n) \mathbb{P}\{\nu_k = n\}, \quad (21)$$

where $\mathbf{t} = (t_1, \dots, t_m)$. Let $\mathbf{a} = (a_1, \dots, a_m)$ be the vector of mean values for $(X_1^{(1)}, \dots, X_1^{(m)})$.

We have

$$f^n(\mathbf{t}/n) = \left(1 + i(\mathbf{a}, \mathbf{t}) \frac{1}{n} + o(\|\mathbf{t}\|/n)\right)^n \rightarrow e^{i(\mathbf{a}, \mathbf{t})} = \prod_{j=1}^m e^{ia_j t_j}, \quad (22)$$

as $n \rightarrow \infty$. The convergence in (21) is uniform with respect to \mathbf{t} taking on values from a compact set. It follows from (21) and (22) that

$$\begin{aligned} \left| \mathbb{E}e^{i(\mathbf{U}/\nu_k, \mathbf{t})} - e^{i(\mathbf{a}, \mathbf{t})} \right| &= \left| \sum_{n=1}^{\infty} (f^n(t_1/n, \dots, t_m/n) - e^{i(\mathbf{a}, \mathbf{t})}) \mathbb{P}\{\nu_k = n\} \right| \\ &\leq \sum_{n=1}^{\infty} |f^n(t_1/n, \dots, t_m/n) - e^{i(\mathbf{a}, \mathbf{t})}| \mathbb{P}\{\nu_k = n\}. \end{aligned} \quad (23)$$

From (22), we can claim that for any $\varepsilon > 0$ there exists such $N_\varepsilon > 1$ independent of k that

$$|f^n(\mathbf{t}/n) - e^{i(\mathbf{a}, \mathbf{t})}| < \varepsilon \quad (24)$$

for all $n > N_\varepsilon$ and all \mathbf{t} from any fixed compact set. Using (23) and (24) one can write

$$|\mathbb{E}e^{i(\mathbf{Z}_k, \mathbf{t})} - e^{i(\mathbf{a}, \mathbf{t})}| \leq 2 \sum_{n=1}^{N_\varepsilon} \mathbb{P}\{\nu_k = n\} + \varepsilon. \quad (25)$$

Since $\nu_k \rightarrow \infty$ as $k \rightarrow \infty$, we can assert that $\mathbb{P}\{\nu_k = n\} \rightarrow 0$ as $k \rightarrow \infty$ for any $n \leq N_\varepsilon$. It finally follows from inequality (25) that

$$|\mathbb{E}e^{i(\mathbf{Z}_k, \mathbf{t})} - e^{i(\mathbf{a}, \mathbf{t})}| \rightarrow 0 \quad (26)$$

as $k \rightarrow \infty$. This completes the proof of the convergence

$$\mathbf{Z}_k \xrightarrow{d} \mathbf{a}. \quad (27)$$

The fact that the multivariate limit distribution of \mathbf{Z}_k is a degenerate one is consistent with the asymptotic behavior of $\text{Cov}(U/\nu, V/\nu)$ (and consequently of $\text{Var}(U/\nu), \text{Var}(V/\nu)$) considered in Section 3. Indeed, we have

$$\text{Cov}\left(\frac{U_k}{\nu_k}, \frac{V_k}{\nu_k}\right) = \mathbb{E}\left\{\frac{1}{\nu_k}\right\} \cdot \text{Cov}(X, Y). \quad (28)$$

By the same argument as that in the proof of (26), it is easy to show that $\mathbb{E}\left\{\frac{1}{\nu_k}\right\} \rightarrow 0$ as $k \rightarrow \infty$. Therefore, the covariance in the left-hand side of (28) tends to zero when ν_k is large in probability. The same is true for the variances of U_k/ν_k and V_k/ν_k , of course. While this behavior of the two central moments is consistent with the convergence established in (27), the correlation coefficient ρ is not well-defined for degenerate random vectors. At the same time, the fact that all components of \mathbf{Z}_k are asymptotically independent is not in conflict with formula (6) because the latter is valid for any value of ν . Nor does it come into conflict with the observation that the intra-cellular gene expression levels are strongly correlated (see Section 4). When deriving formula (6), we divide $\text{Cov}(U_k/\nu_k, V_k/\nu_k)$ by the product of the corresponding standard deviations of U_k/ν_k and V_k/ν_k , which is why the proportionality coefficient $\mathbb{E}\left\{\frac{1}{\nu_k}\right\}$ cancels out and the uncertainty manifesting itself in the limit distribution becomes resolved.

Taken together, the results given by (27) and (6) imply that, while the r.v.s U/ν and V/ν are asymptotically independent, the correlation between the components of each pair (X_i, Y_i) may be arbitrarily strong even when ν takes on large values with high probability.

Such “paradoxical” situations are not uncommon in the theory of probability. It should be emphasized that the above assertion is valid for the random sum of ν i.i.d. random summands normalized by the same random variable ν , and not for other possible ways of normalization. For limit theorems related to the random sums normalized by sequences of nonrandom numbers we refer the reader to [23, 24].

Now we are in a position to make and test the following two predictions:

Prediction 1. The ratios of the observed expression levels U_j and U_r , $j \neq r$, where $j, r = 1, \dots, m$ and m is the total number of genes, tend to have small variances. The covariance between different ratios U_j/U_r is expected to be small as well.

Indeed, proceeding from the LLN, we expect the asymptotic relation

$$\frac{U_j}{U_r} = \frac{\frac{1}{\nu} \sum_{i=1}^{\nu} X_{ij}}{\frac{1}{\nu} \sum_{i=1}^{\nu} X_{ir}} \sim \frac{\mathbb{E}X_{1j}}{\mathbb{E}X_{1r}}, \quad j \neq r, \quad (29)$$

to hold true as $\nu \rightarrow \infty$ in probability. This suggests that every ratio U_j/U_r is virtually constant (across arrays). The above-proven assertion also suggests that every two ratios of the form U_j/U_r and U_l/U_q (with different indices) have small covariances.

To verify *Prediction 1*, we formed all pairs from 1000 randomly selected genes. The mean (over the gene pairs) standard deviation of the ratios U_j/U_r ($j \neq r$) in the HYPERDIP data was equal to 0.102, which is a small value compared to the corresponding mean of the estimated expectations $\mathbb{E}(U_j/U_r)$, the latter value being equal to 1.044. The histogram of the standard deviations in Figure 1 illustrates this point further. Shown in Figure 2 is the histogram of the estimated covariances between U_j/U_r and U_l/U_q in all quadruples formed from 100 randomly selected genes in the HYPERDIP data set. It is clear that they tend to be small as well. This observation explains the most salient properties of the so-called δ – *sequence* [13], as well as the remarkable success of significance testing for differential expression of genes when the relevant methods are applied to the elements of this sequence rather than to the original expression levels [13, 14].

Prediction 2. The average (expectation) of the ratio U_j/U_r is approximately equal to the ratio of the averages of U_j and U_r , $j \neq r$.

It is obvious that

$$\mathbb{E}U_j = \mathbb{E}\nu \mathbb{E}X_{1j}, \quad \mathbb{E}U_r = \mathbb{E}\nu \mathbb{E}X_{1r}.$$

By applying the LLN to the representation

$$\frac{U_j}{U_r} = \frac{\frac{1}{\nu} \sum_{i=1}^{\nu} X_{ij}}{\frac{1}{\nu} \sum_{i=1}^{\nu} X_{ir}},$$

we arrive at the approximate equality

$$\mathbb{E} \left(\frac{U_j}{U_r} \right) \approx \frac{\mathbb{E}U_j}{\mathbb{E}U_r}, \tag{30}$$

whenever ν is large with high probability.

Replacing the expected values with their sample counterparts, we computed the absolute difference between the left- and right-hand sides of equality (30) for all gene pairs formed from 1000 randomly selected genes in the HYPERDIP data set. The resultant histogram (Figure 3) clearly indicates that such differences are very small with their mean (across the gene pairs) being equal to 0.006.

Hence, both predictions appear to be consistent with real data. Similar analyses of the other data sets referred to in Section 3 have confirmed this conjecture.

7. Discussion and concluding remarks

The world of stochastic phenomena is complex and uncanny. Intuition is not the best guide in that world. Some stochastic effects may seem to defy common sense but nevertheless they may be very real from theoretical and practical perspectives. In the present paper, we describe and explore to the best of our ability the impact of random signal aggregation on the correlation structure of microarray gene expression data. While our analysis of real data suggests that this impact may be deemed reasonably moderate in some situations, the main concern still remains because the estimates employed are not sufficiently stable and the underlying model may still need further refinements. A similar concern arises in regard to other standard measures of dependence such as mutual information. The latter characteristic is used extensively for the purposes of relevance network inference (see, e.g., [15]) and all caveats (including Remark 2 of Section 2) pertain equally to such applications.

We overlooked the phenomenon of signal aggregation when discussing the correlation structure of microarray data in our earlier publications [13, 16]. In [16], we tried to make the case that the observed strong and long-ranged correlation between gene expression levels are of a

biological nature rather than a technical flaw of the microarray technology. Our belief was based on the premise that the effects of the technical noise [11] and multiple targeting [16] on the correlation structure of microarray data appeared to be negligible. There is no reason to revise this premise. However, the effect of random summation of expression signals reported in the present paper is a drastically different story. While technical in nature, this effect represents a serious obstacle standing in the way of correlation analysis and network inference. At the same time, the estimates reported in the present paper still indicate the presence of strong correlations between the expression signals produced by different genes at the level of individual cells.

There are statistical questions, other than the estimation of correlation coefficients, that may be relatively insusceptible to the effect of signal aggregation. For example, we hypothesize that it may still be sensible to compare correlation vectors associated with each gene in two different phenotypes in order to extract more information on pathogenesis of some diseases or responses to drug therapies. However, this conjecture invites a special investigation. It is clear that the crux of the difficulty has to do with a natural desire to make inferences about “microscopic” processes of transcription within individual cells from “macroscopic” observations yielded by gene expression measurements. From this perspective, the mixing effect caused by signal summation should be considered as confounding [17] and, as such, is undesirable. Needless to say, one can employ the correlation coefficients between observed expression signals as more global characteristics of the cell system under study rather than associations between gene activities within each cell. Such characteristics may still represent a source of useful information. From this viewpoint, the results of correlation analysis of gene expression data can be interpreted in terms of aggregated (over the cells) genes, an obvious departure from the interpretation that has been in wide use among molecular biologists and bioinformaticians. Since tissue-specific mechanisms regulating cell functions are not well-understood, it is premature to judge whether or not this cautious interpretation is of biological interest.

The most critical question still remains: How can the true correlation be extracted from observed expression levels despite the masking effect of signal aggregation? At present, we have no satisfactory answer to this question. However, some practical expedients mitigating the adverse consequences of signal aggregation can be envisioned. As discussed in Section 5, one

approach is to combine the information provided by technical and biological replicates using the mathematical treatment of the problem presented in this paper. Yet another possibility is to modify the experimental protocol so that the total DNA rather than the total RNA be kept constant across the arrays (see Remark 1). The rationale for this suggestion is that the correlation between the parameter ν and the total number of cells (gauged by the DNA content) in a given biological sample may well be stronger than that between ν and the total RNA. The main problem with hybridization-based technologies is that the latent parameter ν is not accessible to direct control. The situation is not the same with the sequencing technology that produces counts of all transcripts present in the biological sample. It seems likely that the sequence-based technology offered by Illumina (Solexa) may make it much easier to keep the parameter ν constant across biological samples. All the above-mentioned possibilities have yet to be verified in biological experiments, of course.

Finally, a search for measures of dependence or relations between gene expression signals that are preserved under signal aggregation is warranted. For example, introduce the following characteristic

$$\rho(X, Y)\xi_x\xi_y = \mu_\nu \left[\rho(U, V)\xi_u\xi_v - \xi^2(\nu) \right]. \quad (31)$$

It is easy to see that the inequality

$$\rho(U_1, V_1)\xi_{u_1}\xi_{v_1} > \rho(U_2, V_2)\xi_{u_2}\xi_{v_2} \quad (32)$$

implies

$$\rho(X_1, Y_1)\xi_{x_1}\xi_{y_1} > \rho(X_2, Y_2)\xi_{x_2}\xi_{y_2}, \quad (33)$$

because μ_ν and $\xi^2(\nu)$ are the same for all genes. Therefore, the coefficient $\rho(U, V)\xi_u\xi_v$ introduced in (31) preserves inequalities between the corresponding coefficients for (X, Y) . In this connection, it is important to recall that the variation coefficient of observed expression levels is almost constant across genes, a fact mentioned in Section 4. Under such conditions, inequalities (32) and (33) imply the same inequalities for the corresponding correlation coefficients. This suggests that the ranking of gene pairs by the correlation coefficient may still be possible and such inference can probably be improved by stratifying the population of genes by the value of the variation coefficient. Whether this observation is of real utility in studying relationships between genes within the network paradigm has yet to be explored.

The future of the whole research area dealing with regulatory networks hinges on our ability to surmount the obstacle described in the present paper either by means of mathematics (including recourse to parametric methods) or through radical technological improvements.

Acknowledgements

This research is supported by grant MSM 002160839 from the Ministry of Education, Czech Republic, NIH/NIGMS grants RO1 GM075299, R21 GM079259, and NIH/NIEHS grant T32 ES 007271. Ms. Linlin Chen conducted computations for Section 4 and we are grateful for her assistance.

References

- [1] Chu T, Glymour C, Scheines R, Spirtes P: **A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays**. *Bioinformatics* 2003, **19**: 1147-1152.
- [2] Held GA, Grinstein G, Tu Y: **Modeling of DNA microarray data by using physical properties of hybridization**. *Proc. Natl. Acad. Sci. USA* 2003, **100**(13): 7575-7580.
- [3] Pujana MA, Han J-DJ, Starita LM, Stevens KN, Tewari M, Ahn JS, Rennert G, Moreno C, Kirchhoff T, Gold B, Assmann V, ElShamy WM, Rual, J-F, Levine D, Rozek LS, Gelman RS, Gunsalus KC, Greenberg RA, Sobhian B, Bertin N, Venkatesan K, Ayivi-Guedehoussou N, Solé X, Hernández P, Lázaro C, Nathanson KL, Weber BL, Cusick ME, Hill DE, Offit K, Livingston DM, Gruber SB, Parvin JD, Vidal M: **Network modeling links breast cancer susceptibility and centrosome dysfunction**. *Nat Genet* 2007, **39**: 1338-1349.
- [4] Butte AJ, Tamayo P, Slonim D, Golub TR, Kohar IS: **Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks**. *Proc Natl Acad Sci USA* 2000, **97**: 12182-12186.

- [5] Qiu X, Brooks AI, Klebanov L, Yakovlev A: **The effects of normalization on the correlation structure of microarray data.** BMC Bioinformatics 2005, **6**: Article 120.
- [6] Lim WK, Wang K, Lefebvre C, Califano A: **Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks.** Bioinformatics 2007, **23**: i282-i288.
- [7] Chen L, Klebanov L, Yakovlev AY: **Normality of gene expression revisited.** J Biol Syst 2007, **15**(1): 39-48.
- [8] Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet J-P, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Ru Wei R, Carr SA, Lander, ES, Golub TR: **The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease.** Science 2006, **313**: 1929-1935.
- [9] Yeoh, EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, Cheng C, Campana D, Wilkins D, Zhou X, Li J, Liu H, Pui CH, Evans WE, Naeve C, Wong L, Downing JR: **Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling.** Cancer Cell 2002, **1**(2): 133-143.
- [10] Singh D, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers WR: **Gene expression correlates of clinical prostate cancer behavior.** Cancer Cell 2002, **1**: 203-209.
- [11] Klebanov L, Yakovlev A: **How high is the level of technical noise in microarray data?** Biology Direct 2007, **2**: Article 9.
- [12] Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, Lou Y, Sun YA, Willey JC, Setterquist RA, Fischer GM, Tong W, Dragan YP, Dix DJ, Frueh FW, Goodsaid FM, Herman D, Jensen RV, Johnson CD, Lobenhofer EK, Puri RK, Sherf U., Thierry-Mieg J, Wang C, Wilson M, Wolber PK: **The microarray quality control (MAQC) project shows inter- and**

- intraplatform reproducibility of gene expression measurements.** Nat Biotechnol 2006, **24**(9): 1151-1161.
- [13] Klebanov L, Yakovlev A: **Diverse correlation structures in microarray gene expression data and their utility in improving statistical inference.** Annals of Applied Statistics 2007, **1**(2): 538-559.
- [14] Klebanov L, Qiu X, Yakovlev AY: **Testing differential expression in non-overlapping gene pairs: A new perspective for the empirical Bayes method.** J Bioinformatics Comput Biol, in press.
- [15] Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: **Reverse engineering of regulatory networks in human B cells.** Nat Genet 2005, **37**: 382-390.
- [16] Klebanov L, Chen L, Yakovlev A: **Revisiting adverse effects of cross-hybridization in Affymetrix gene expression data: do they matter for correlation analysis?** Biology Direct 2007, **2**: Article 28.
- [17] Leek JT, Storey JD: **Capturing heterogeneity in gene expression studies by surrogate variable analysis.** PLoS Genet 2007, **3**(9): e161.
- [18] Melamed JA(1989). **Limit theorems in the set-up of summation of a random number of independent and identically distributed random variables.** In: Stability Problems for Stochastic Models, Lecture Notes in Mathematics 1989, **1412**: 194-228, Springer, Berlin.
- [19] Gnedenko BV: **On convergence of laws of a distribution of sums of independent summands.** Doklady Akad. Nauk USSR 1938, **18**(4-5): 231-234.
- [20] Klebanov LB, Rachev ST: **Sums of a random number of random variables and their approximations with ν -accompanying infinitely divisible laws.** Serdica Math J 1996, **22**: 471-496.
- [21] Klebanov L, Kozubowski TJ, Rachev ST: **Ill-posed problems in probability and stability of random sums.** Nova Science Publishers, NY, 2006.

- [22] Révész P: **The laws of large numbers**. Academic Press, NY, 1968.
- [23] Robbins H: **The asymptotic distribution of the sum of a random number of independent random variables**. Bull Am Math Soc 1948, **54**: 1151-1161.
- [24] Gnedenko BV, Korolev VY: **Random summation. Limit theorems and applications**. CRC Press, Boca Raton, 2000.

Table 1: Variation coefficients of gene expression levels estimated from different data sets. The average and minimal (across genes) values are presented.

Dataset	Average ξ_u	min ξ_u
TELL	0.233	0.188
HYPERDIP	0.267	0.211
BCC	0.299	0.173
PCNORM	0.299	0.213
PCTUM	0.251	0.152

Figure legends

Figure 1: Histogram of standard deviations for the ratios of expression levels in all gene pairs formed from 1000 genes. The HYPERDIP data set.

Figure 2: Histogram of covariances between the ratios of gene expressions in all quadruples from a subset of 100 genes. The HYPERDIP data set.

Figure 3: Histogram of the differences between $\mathbb{E}(U_j/U_r)$ and $\mathbb{E}(U_j)/\mathbb{E}(U_r)$ estimated by replacing the expected values with the corresponding sample means. The HYPERDIP data set.

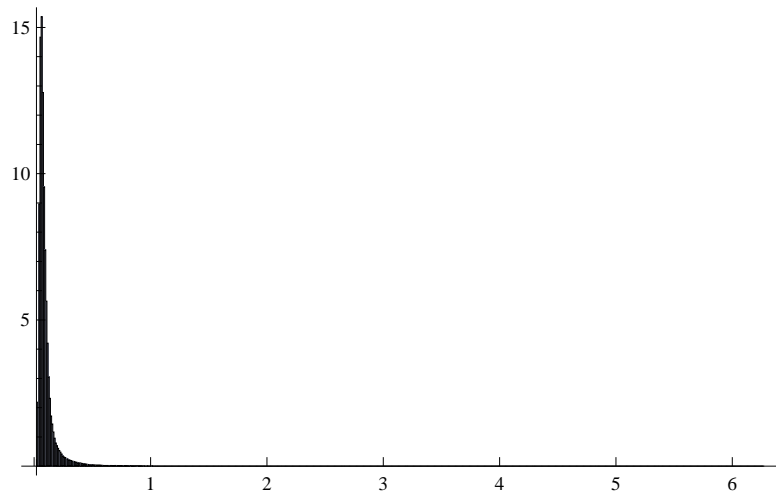


Figure 1: Histogram of standard deviations for the ratios of expression levels in all gene pairs formed from 1000 genes. The HYPERDIP data set.

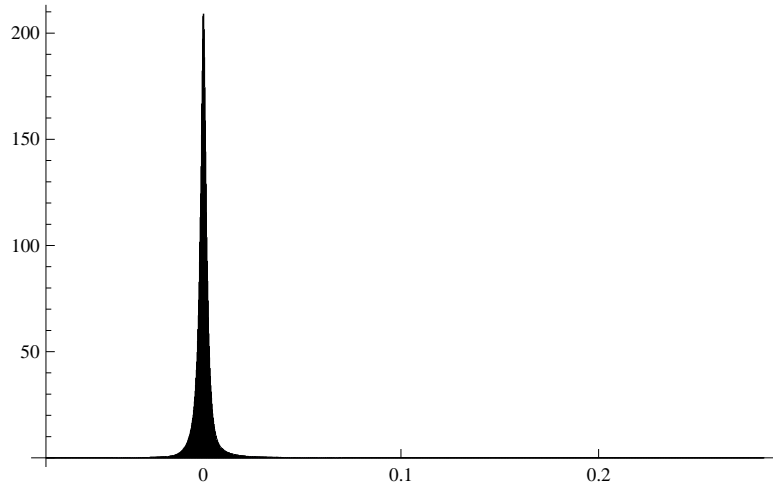


Figure 2: Histogram of covariances between the ratios of gene expressions in all quadruples from a subset of 100 genes. The HYPERDIP data set.

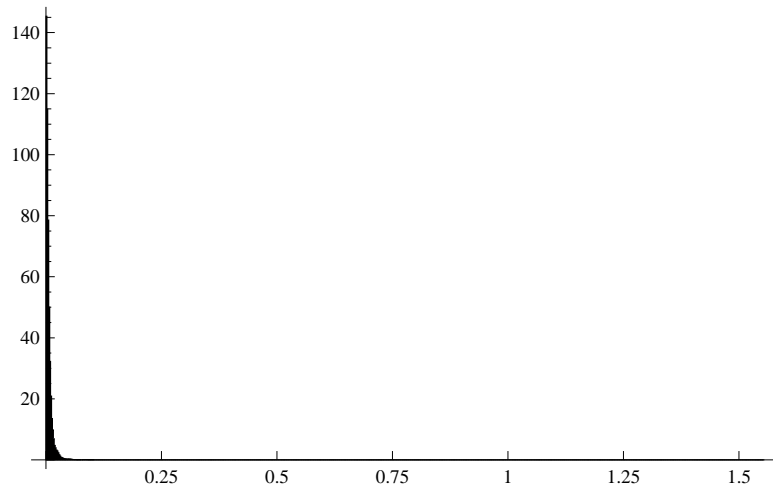


Figure 3: Histogram of the differences between $\mathbb{E}(U_j/U_r)$ and $\mathbb{E}(U_j)/\mathbb{E}(U_r)$ estimated by replacing the expected values with the corresponding sample means. The HYPERDIP data set.