Dembélé and Kastner *BMC Bioinformatics* (2016) 17:462
DOI 10.1186/s12859-016-1322-0

BMC Bioinformatics

CORRESPONDENCE Open Access

Comments on: fold change rank ordering statistics: a new method for detecting differentially expressed genes

Doulaye Dembélé[1,2*] and Philippe Kastner[1,3]

Abstract

We published a new method (BMC Bioinformatics 2014, 15:14) for searching for differentially expressed genes from two biological conditions datasets. The presentation of theorem 1 in this paper was incomplete. We received an anonymous comment about our publication that motivates the present work. Here, we present a complementary result which is necessary from the theoretical point of view to demonstrate our theorem. We also show that this result has no negative impact on our conclusions obtained with synthetic and experimental microarrays datasets.

Keywords: Differentially expressed genes, Fold change, Average of ranks

Background

To search for differentially expressed (DE) genes in profiling studies, we presented a new method based on fold change rank ordering statistics (FCROS). For the derivation of this method, we considered microarrays data from two biological conditions where n probes (genes) were used with m_1 control and m_2 test samples. We performed k pairwise comparisons ($k = m_1 m_2$) of the data samples and computed fold changes (FC) for each gene. The FCs obtained for each comparison were sorted in increasing order and their corresponding ranks were associated with genes. Hence, we can form a matrix of rank values R with components r_{ij} ($i = 1, 2, \ldots, n, j = 1, 2, \ldots, k$). We noted $\mathbf{r}_i = [\, r_{i1}\ r_{i2}\ \ldots\ r_{ik}]^T$ the vector of rank values associated with gene i. We noted $\bar{r}_i$, the average of ranks (a.o.r) value for gene i. The value for $\bar{r}_i$ varies between $a = \min_i\{\bar{r}_i\}$ and $b = \max_i\{\bar{r}_i\}$. That allows to associate an unique vector of a.o.r values with the n genes: $\bar{\mathbf{r}} = [\, a,\ (a+\delta_1),\ (a+\delta_1+\delta_2),\ \ldots,\ (a+\delta_1+\ldots+\delta_{n-2}),\ b]^T$ where the scalars δ_i are the differences between consecutive ordered a.o.r. Without loss of generality, we assumed that the differences δ_i have the same value which is approximated by their mean: $\delta = \frac{b-a}{n-1}$. Using these notations, we derived a theorem showing a normal distribution for vector $\bar{\mathbf{r}}$ [1]. The content of this theorem was incomplete as shown in the following lemma we received from an anonymous reader.

Lemma 1 *Let consider the matrix of rank values R under the assumption that the rank values in each column are all distinct. Assume uniform random sampling without replacement model for the columns of R, i.e. each column of R is an independent draw from the set of all permutations of $\{1, \ldots, n\}$ with uniform probability $\frac{1}{n!}$ for each permutation. Then, the asymptotic distribution of the unordered vector average of rank (a.o.r.), $r = \left(r_i = \frac{1}{k}\sum_{j=1}^{k} R_{ij}\right), i \in 1 \ldots n$, has a mean $\frac{n+1}{2}\mathbf{1}_n$ and degenerate variance-covariance matrix $\Sigma(n,n)$, $\det \Sigma = 0$:*

$$\Sigma = \begin{pmatrix} \beta & \alpha & \ldots & \alpha & \alpha \\ \alpha & \beta & \ldots & \alpha & \alpha \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \alpha & \alpha & \ldots & \beta & \alpha \\ \alpha & \alpha & \ldots & \alpha & \beta \end{pmatrix} \quad (1)$$

with diagonal element $\beta = \frac{n^2-1}{12}$, off-diagonal element $\alpha = -\frac{\beta}{n-1}$ and $\mathbf{1}_n = [\, 1, 1, \ldots, 1]^T$.

Proof Note that for $k \to \infty$, the appearance of all elements of the set $\{1, \ldots, n\}$ in each row of R under the

*Correspondence: doulaye@igbmc.fr
[1]Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), CNRS UMR 7104, INSERM U964, Université de Strasbourg, 67404 Illkirch, France
[2]IGBMC Microarray and Sequencing Platform, 67404 Illkirch, France
Full list of author information is available at the end of the article

assumed sampling model are equally likely, hence by the weak law of large numbers ([2], page 235) the asymptotic mean is the constant vector $\left(\frac{1}{n}\sum_{i=1}^{n} i\right)\mathbf{1}_n = \frac{n+1}{2}\mathbf{1}_n$. Under the same observation, the asymptotic variance, $\forall \ell \in \{1,\ldots,n\}$, is equal to:

$$Var\,(r_\ell) \underset{k}{\longrightarrow} \beta = \frac{1}{n}\left[\sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)^2\right] = \frac{n^2-1}{12} \quad (2)$$

The asymptotic covariance is computed as a two-index summation over the set $\{1,\ldots,n\}$ with the restriction that no two indices can be the same since the columns are permutations by construction, hence $\forall \ell \neq m \in \{1,\ldots,n\}$:

$$\begin{aligned} &Cov\,(r_\ell, r_m) \underset{k}{\longrightarrow} \alpha \\ &= \frac{1}{n\,(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\left(i-\frac{n+1}{2}\right)\left(j-\frac{n+1}{2}\right) \quad (3)\\ &= \frac{1}{n\,(n-1)}\left\{\left[\sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)\right]^2 - \sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)^2\right\} \quad (4)\\ &= -\frac{1}{n\,(n-1)}\sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)^2 = -\frac{\beta}{n-1}. \quad (5)\end{aligned}$$

Thus, since $\Sigma\mathbf{1}_n = \mathbf{0}$, it follows that $\det \Sigma = 0$. □

This lemma shows that the covariance term was missed in our theorem. In the next section, we present a complete version of our theorem using the notations we adopted in [1].

Results

From our notations, we have $\bar{\mathbf{r}} = [a,\ a+\delta,\ a+2\delta,\ \ldots,\ a+(n-1)\delta]^T$ the vector with the a.o.r values. Each component of the vector $\bar{\mathbf{r}}$ can be writen as: $\mathcal{R}_\ell = (a+\ell\delta), \ell = 0,1,\ldots,n-1$. The theorem 1 in ([1], page 3) should be read as:

Theorem 1 *When the number k of the pairwise comparisons grows, the ordered average of ranks (a.o.r.)* $\bar{\mathbf{r}}$ *have a normal distribution. The mean of this distribution is* $\frac{a+b}{2}\mathbf{1}_n$*, its variance-covariance matrix has diagonal element* $\frac{n^2-1}{12}\delta^2$ *and off-diagonal element* $-\frac{n+1}{12}\delta^2$*, where a and b are the minimum and the maximum of the observed a.o.r.,* $\bar{\mathbf{r}}$*, respectively.* δ *is the average difference between consecutive ordered a.o.r.* $\bar{\mathbf{r}}$*.*

Proof From the following definitions:

$$\begin{aligned} E\{\mathcal{R}_\ell\} &= \frac{1}{n}\sum_{\ell=1}^{n}\mathcal{R}_\ell \\ Var(\mathcal{R}_\ell) &= E\{\mathcal{R}_\ell^2\} - (E\{\mathcal{R}_\ell\})^2 \\ Cov(\mathcal{R}_\ell, \mathcal{R}_m)_{m\neq\ell} &= E\{\mathcal{R}_\ell\mathcal{R}_m\} - (E\{\mathcal{R}_\ell\})^2\end{aligned}$$

and using $\delta = \frac{b-a}{n-1}$, a component of the mean of the normal distribution is:

$$E\left\{\sum_{\ell=0}^{n-1}(a+\ell\delta)\right\} = \frac{1}{n}\sum_{\ell=0}^{n-1}(a+\ell\delta) = a+\frac{n-1}{2}\delta = \frac{b+a}{2}. \quad (6)$$

A component of the variance (diagonal element) of the normal distribution matrix is:

$$\begin{aligned} Var(\mathcal{R}_\ell) &= E\left\{\sum_{\ell=0}^{n-1}(a+\ell\delta)^2\right\} - \left(a+\frac{n-1}{2}\delta\right)^2 \\ &= E\left\{\sum_{\ell=0}^{n-1}\left(a^2+2a\delta\ell+\delta^2\ell^2\right)\right\} - \left(a+\frac{n-1}{2}\delta\right)^2 \\ &= \frac{1}{n}\left(na^2+2a\delta\frac{n(n-1)}{2}+\frac{n(n-1)(2n-1)}{6}\delta^2\right) \\ &\quad -\left(a+\frac{n-1}{2}\delta\right)^2 = \frac{n^2-1}{12}\delta^2. \end{aligned} \quad (7)$$

A component of the covariance (off-diagonal element) of the normal distribution matrix is:

$$\begin{aligned} Cov(\mathcal{R}_\ell, \mathcal{R}_m)_{m\neq\ell} &= E\left\{\sum_{\ell=0}^{n-1}\sum_{\substack{m=0\\ m\neq\ell}}^{n-1}(a+\ell\delta)(a+m\delta)\right\} - \left(a+\frac{n-1}{2}\delta\right)^2 \\ &= E\left\{\sum_{\ell=0}^{n-1}\sum_{m=0}^{n-1}\left(a^2+a\delta\ell+a\delta m+\delta^2 m\ell\right)\right. \\ &\quad \left.-\sum_{\ell=0}^{n-1}(a+\ell\delta)^2\right\} - \left(a+\frac{n-1}{2}\delta\right)^2 \\ &= \frac{1}{n(n-1)}\left(n^2a^2+n^2(n-1)a\delta+\frac{n^2(n-1)^2}{4}\delta^2\right. \\ &\quad \left.-na^2-n(n-1)a\delta-\frac{n(n-1)(2n-1)}{6}\delta^2\right) \\ &\quad -\left(a+\frac{n-1}{2}\delta\right)^2 = -\frac{n+1}{12}\delta^2. \end{aligned} \quad (8)$$

□

Table 1 Values of the mean, the variance and the covariance components when *n* increases

n	10	100	1,000	10,000
$r^\star$	$\frac{1}{2}+5*10^{-2}$	$\frac{1}{2}+5*10^{-3}$	$\frac{1}{2}+5*10^{-4}$	$\frac{1}{2}+5*10^{-5}$
$\beta^\star$	$\frac{1}{12}-8.33*10^{-4}$	$\frac{1}{12}-8.33*10^{-6}$	$\frac{1}{12}-8.33*10^{-8}$	$\frac{1}{12}-8.33*10^{-10}$
$\alpha^\star$	$-9.17*10^{-3}$	$-8.4*10^{-4}$	$-8.34*10^{-5}$	$-8.33*10^{-6}$

By setting $a = \delta = 1$ and $b = n$ in the theorem 1, the mean and the variance-covariance component values are the same as in lemma 1. These setting values for a, b and δ correspond to the case we called ideal situation ([1], page 4).

For the FCROS algorithm, we used the standardized rank value, i.e., each observed rank value is divided by n. The mean and variance-covariance components should be divided by n and n^2 respectively. This leads to a mean component $r^\star = \left(\frac{1}{2} + \frac{1}{2n}\right)$, and a variance-covariance matrix with a diagonal component $\beta^\star = \left(\frac{1}{12} - \frac{1}{12n^2}\right)$ and a off-diagonal component $\alpha^\star = -\left(\frac{1}{12} - \frac{1}{12n^2}\right)\frac{1}{n-1}$. Table 1 shows the values for $r^\star, \beta^\star$ and $\alpha^\star$ when n increases. For a large value for n, the off-diagonal components of the variance-covariance matrix vanish. Hence, when n is large, a good approximation for the mean and the variance components are $\frac{1}{2}$ and $\frac{1}{12}$, respectively.

Discussion and conclusions

As shown, the theorem we previously presented was incomplete since the covariance term was missed. The present complementary result is necessary from the theoretical point of view, and we are grateful to the anonymous reader for pointing this out. This result will be useful for small values of n. However, for high throughput biological datasets, n is large, often greater than 10,000 ([1], page 2). For such values of n, the rank deficient variance-covariance matrix of the normal distribution associated with the a.o.r values is near a diagonal matrix. Hence, it is as if the a.o.r values of each gene follow a normal distribution with parameters $\frac{1}{2}$ and $\frac{1}{12}$.

Acknowledgments
We thank the anonymous reader for drawing our attention to this result.

Funding
This work was supported by funds from CNRS, INSERM and University of Strasbourg.

Availability of data and materials
Not Applicable.

Authors' contributions
DD drafted the paper and performed the analyses. Both authors developed the method and contributed to the manuscript. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not Applicable.

Ethics approval and consent to participate
Not Applicable.

Author details
[1]Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), CNRS UMR 7104, INSERM U964, Université de Strasbourg, 67404 Illkirch, France. [2]IGBMC Microarray and Sequencing Platform, 67404 Illkirch, France. [3]Faculté de Medécine, Université de Strasbourg, Strasbourg, France.

Received: 1 July 2016 Accepted: 5 November 2016
Published online: 15 November 2016

References

1. Dembélé D, Kastner P. Fold change ordering statistics: a new method for detecting differentially expressed genes. BMC Bioinforma. 2014;15(1):14.
2. Feller W. An Introduction to Probability Theory and Its Applications, vol. II, (2nd Edition). New York: John Wiley & Sons; 1971.