doi:10.1016/j.jmb.2004.11.018 J. Mol. Biol. (2005) 345, 645–649

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

COMMUNICATION

ArrayVigil: A Methodology for Statistical Comparison of Gene Signatures Using Segregated-one-tailed (SOT) Wilcoxon's Signed-rank Test

Haseeb Ahmad Khan

Department of Biochemistry Building 5, Room 2B9, College of Science, King Saud University P.O. Box 2455, Riyadh 11451 Saudi Arabia

Due to versatile diagnostic and prognostic fidelity molecular signatures or fingerprints are anticipated as the most powerful tools for cancer management in the near future. Notwithstanding the experimental advancements in microarray technology, methods for analyzing either whole arrays or gene signatures have not been firmly established. Recently, an algorithm, ArraySolver has been reported by Khan for two-group comparison of microarray gene expression data using two-tailed Wilcoxon signed-rank test. Most of the molecular signatures are composed of two sets of genes (hybrid signatures) wherein up-regulation of one set and down-regulation of the other set collectively define the purpose of a gene signature. Since the direction of a selected gene's expression (positive or negative) with respect to a particular disease condition is known, application of one-tailed statistics could be a more relevant choice. A novel method, ArrayVigil, is described for comparing hybrid signatures using segregated-one-tailed (SOT) Wilcoxon signed-rank test and the results compared with integrated-two-tailed (ITT) procedures (SPSS and ArraySolver). ArrayVigil resulted in lower *P* values than those obtained from ITT statistics while comparing real data from four signatures.

Keywords: gene signatures; microarray; non-parametric comparisons; gene expression; algorithm

Microarray gene clustering is a commonly used computation tool for molecular classification of disease states, functional grouping of genes, and biological description of gene regulation.[1–5] The pattern of expressed genes in a microarray demonstrates a typical profile in relation to cancer type or disease severity. These unique sets of genes defining specific pathophysiology are regarded as molecular "signatures" or "fingerprints". Tumors with closely related genetic lesions will have similar signatures and also will be expected to have similar clinical behaviors.[6] The information encoded in gene signatures can provide valuable insights in cancer diagnosis and prognosis.[2,7–12]

A prominent role for the ideas of statistical inference in microarray studies has been suggested by Kerr & Churchill.[13] Selection of an appropriate statistical method for two-group comparisons of expression data is highly desirable for effective application of gene signatures. Although gene clustering is an important tool for the identification of like-groups in a microarray, this methodology cannot be used for conventional two-group comparisons. However, numerous statistical procedures including *t*-test,[14,15] analysis of variance,[16] Pearson correlation,[17] Welch test[18] and Mann-Whitney U test[19,20] have been used for comparison of microarray data. Wolfinger *et al*.[21] proposed a mixed model for simultaneous assessment of significant differences between multiple types of biological samples. Recently, an algorithm based on Wilcoxon's matched-pair signed-rank test has been reported for comparing microarray gene expression data.[22] However, prior to any statistical comparison, microarray data need to be normalized (adjusted) for potential confounding effects usually arising from variation in technology.[23,24]

Since the configuration of a gene signature is selectively comprised of fixed numbers of either up

Abbreviations used: SOT, segregated-one-tailed; ITT, integrated-two-tailed.

E-mail address of the corresponding author: khan_haseeb@yahoo.com

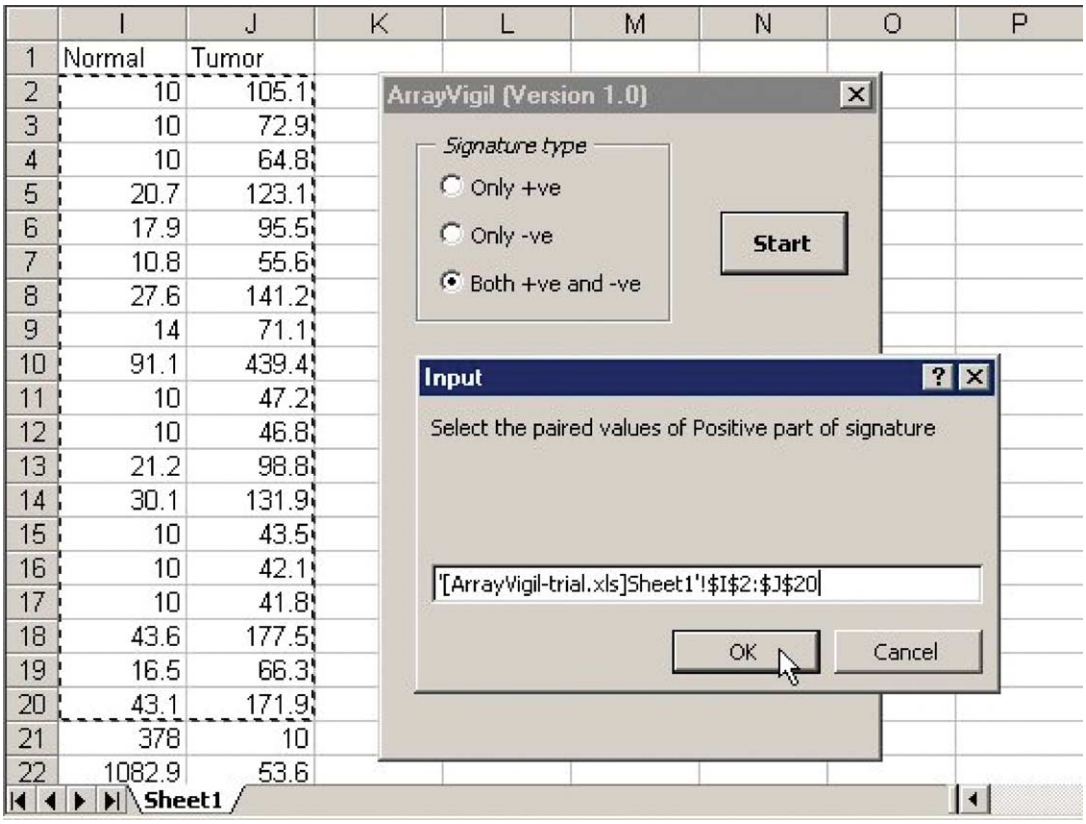

Figure 1. Excel worksheet displaying the gene expression data of signature D (Table 1) as well as the functioning of ArrayVigil software. After selecting the signature type (hybrid signature in this example) using the option button, the Start button is clicked to activate input boxes. The Figure shows the selection of paired gene expression values for the positive part of signature D. Clicking the OK button at this stage will reset the input box for selecting the paired values of the negative part of the gene signature. The next click on the OK button will execute the program and the results will be displayed.

or down-regulated specific gene probes or quite often a combination of both (hybrid signature) it was hypothesized that one-step statistical analysis of hybrid signatures may not be as realistic as the segregated analysis of both up and down-regulated genes within a hybrid signature. Based on these assumptions, an attempt has been made to utilize segregated-one-tailed (SOT) analysis of gene signatures using Wilcoxon signed-rank test and the results from the real data have been compared with the integrated-two-tailed (ITT) statistics.

For SOT analysis, the statistical significance of both positive and negative portions of the hybrid signature is determined individually using one-tailed Wilcoxon signed-rank test and then the overall significance is computed on the basis of quantitative sharing of representative gene sets using the formula:

$$P = [P_{(+)}N_{(+)} + P_{(-)}N_{(-)}]/N \quad (1)$$

P is the probability of whole signature and N is the total number of genes in the signature. The subscripts $(+)$ and $(-)$ suffixed to P and N denote probability (significance) and number of genes corresponding to the positive (up-regulated) and negative (down-regulated) parts of gene signature, respectively.

Microsoft Excel (Version 2000) platform was used to develop ArrayVigil. The program has been designed for both, individual (positive or negative) as well as hybrid (positive and negative) signatures and these modes of signatures can be selected by

Table 1. Characteristics of molecular signatures used in this study

Signature identity	Application	Number of genes			Reference
		Over expressed (+ ve)	Under expressed (− ve)	Total	
Signature A	Ovarian carcinoma	15	15	30	1
Signature B	Ulcerative colitis	11	12	23	25
Signature C	Leukemia	23	25	48	2
Signature D	Adenocarcinoma	19	47	66	14

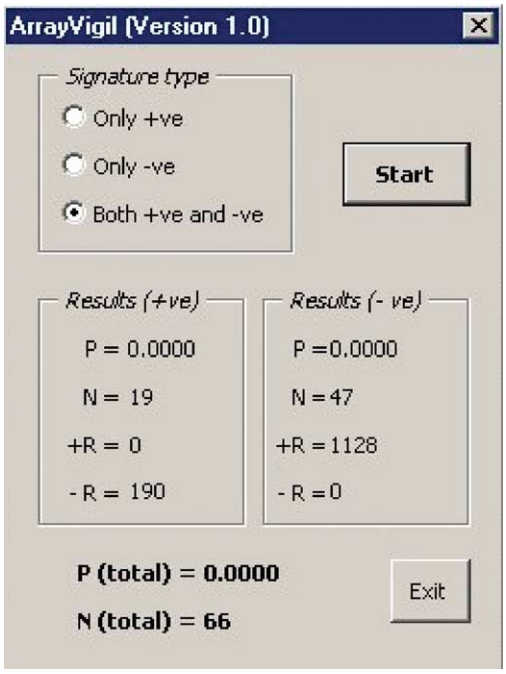

Figure 2. Report window of ArrayVigil software showing the results of two-group comparison of gene expression data of signature D. Clicking the Start button will initiate a new analysis whereas clicking the Exit button will unload the program. Abbreviations: *P*, probability; *N*, number of genes; *R*, sum of ranks. The prefixes + and − to *R* denote positive and negative ranks, respectively. *P* (total) is computed using formula (1) as described in the text.

using the respective option button (Figure 1). The selection of gene expression data is interactive and controlled by the input boxes (Figure 1). The "Start" button on the main window is used to activate input boxes, whereas clicking the "OK" button executes the program and the results are displayed (Figure 2).

In order to validate the program four gene signatures comprised of different numbers of positive and negative genes were selected (Table 1). The published data corresponding to above signatures[1,2,14,25] were subjected to two-group (control *versus* disease) comparisons using ITT Wilcoxon signed-rank test with the aid of two different programs, SPSS[26] and ArraySolver,[22] and the results were compared with those obtained from ArrayVigil (Table 2). The *P* values for all the four signatures resulting from SPSS and ArraySolver were identical but quite different from the *P* values resulting from ArrayVigil. The two-group comparisons for signatures B and C appeared to be insignificant using SPSS or ArraySolver as opposed to statistically significant outcome using ArrayVigil (Table 2). This variation in the *P* values is attributed to the differences in the methodologies used in ArraySolver or SPSS (ITT) and ArrayVigil (SOT) that produced different sums of positive and negative ranks (Table 2).

The Wilcoxon signed-rank test is a non-parametric test that examines differences between dependent groups,[27] hence supposed to be more useful for analyzing microarray expression data. Wilcoxon signed-rank test has been applied for pairwise comparison of gene expression data obtained from reverse-transcription PCR,[28] real-time PCR,[29] *in situ* hybridization,[30] immunohistochemistry[31] and microarrays.[22] Despite their conservativeness or having a lower statistical power with normalized data[32] non-parametric tests have been suggested to be more advantageous when the computationally identified genes need to be tested biologically.[33] Furthermore, a careful selection of tail type (SOT or ITT) could also have a significant impact on the observed probability. Since the gene sets corresponding to favorable/unfavorable direction of gene expression for a particular gene signature are known to the investigator, application of SOT statistics would result in more appropriate interpretation as demonstrated here.

Table 2. Comparative evaluation of various programs for the analysis of molecular signatures using ITT or SOT based Wilcoxon's signed-rank statistics

Signature	SPSS[a]	ArraySolver[a]	ArrayVigil[b]
Signature A	$P=0.021$ $Z=-2.314$ $+R=345, -R=120$	$P=0.0208$ $Z=-2.3139$	$P=0.0050$ $P_{(+)}=0.005, +R_{(+)}=0, -R_{(+)}=120$ $P_{(-)}=0.005, +R_{(-)}=120, -R_{(-)}=0$
Signature B	$P=0.067$ $Z=-1.829$ $+R=198, -R=78$	$P>0.05$ $T=78$	$P=0.0050$ $P_{(+)}=0.005, +R_{(+)}=0, -R_{(+)}=66$ $P_{(-)}=0.005, +R_{(-)}=78, -R_{(-)}=0$
Signature C	$P=0.182$ $Z=-1.333$ $+R=458, -R=718$	$P=0.1836$ $Z=-1.3333$	$P=0.0050$ $P_{(+)}=0.005, +R_{(+)}=0, -R_{(+)}=276$ $P_{(-)}=0.005, +R_{(-)}=325, -R_{(-)}=0$
Signature D	$P=0.000$ $Z=-3.565$ $+R=547, -R=1663$	$P=0.00046$ $Z=-3.5677$	$P=0.0000$ $P_{(+)}=0.000, +R_{(+)}=0, -R_{(+)}=190$ $P_{(-)}=0.000, +R_{(-)}=1128, -R_{(-)}=0$

[a] ITT.
[b] SOT. The suffixes (+) and (−), respectively, define up and down-regulated portions of a gene signature. The prefixes + and − to *R* define the sum of positive and negative ranks, respectively.

Availability of software

Contact the author to get the ArrayVigil software.

Acknowledgements

The author is highly thankful to the research groups of Dr Kai Wang (Chiroscience R&D, Inc., Bothell, WA, USA); Dr Thomas P. Dooley (Integri-Derm Inc., Birmingham AL, USA); Dr Todd R. Golub (Massachusetts Institute of Technology, Cambridge, MA, USA) and Dr Daniel A. Notterman (Princeton University, Princeton, NJ, USA) for using their published data to validate the ArrayVigil methodology.

References

1. Wang, K., Gan, L., Jeffery, E., Gayle, M., Gown, A. M., Skelly, M. *et al*. (1999). Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. *Gene*, **229**, 101–108.
2. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P. *et al*. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
3. Gaasterland, T. & Bekiranov, S. (2000). Making the most of microarray data. *Nature Genet.* **24**, 204–206.
4. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E. *et al*. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci.* **96**, 2907–2912.
5. Munagala, K., Tibshirani, R. & Brown, P. O. (2004). Cancer characterization and feature set extraction by discriminative margin clustering. *BMC Bioinformatics*, **5**, 21.
6. Ladanyi, M., Chan, W. C., Triche, T. J. & Gerald, W. L. (2001). Expression profiling of human tumors: the end of surgical pathology. *J. Mol. Diagnos.* **3**, 92–97.
7. Martin, K. J., Kritzman, B. M., Price, L. M., Koh, B., Kwan, C. P., Zhang, X. *et al*. (2000). Linking gene expression patterns to therapeutic groups in breast cancer. *Cancer Res.* **60**, 2232–2238.
8. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A. *et al*. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.
9. Chuma, M., Sakamoto, M., Yamazaki, K., Ohta, T., Ohki, M., Asaka, M. *et al*. (2003). Expression profiling in multistage hepatocarcinogenesis: identification of HSP70 as a molecular marker of early hepatocellular carcinoma. *Hepatology*, **37**, 198–207.
10. Tan, Z. J., Hu, X. G., Cao, G. S. & Tang, Y. (2003). Analysis of gene expression profile of pancreatic carcinoma using cDNA microarray. *World J. Gastroenterol.* **9**, 818–823.
11. Ramaswamy, S., Ross, K. N., Lander, E. S. & Golub, T. R. (2003). A molecular signature of metastasis in primary solid tumors. *Nature Genet.* **33**, 49–54.
12. Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H. *et al*. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA*, **98**, 10869–10874.
13. Kerr, M. K. & Churchill, G. A. (2001). Statistical design and the analysis of gene expression microarray data. *Genet. Res.* **77**, 123–128.
14. Notterman, D. A., Alon, U., Sierk, A. J. & Levine, A. J. (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res.* **61**, 3124–3130.
15. Tanaka, T. S., Jaradat, S. A., Lim, M. K., Kargul, G. J., Wang, X., Grahovac, M. J. *et al*. (2000). Genome-wide expression profiling of mid-gestation placenta and embryo using a 15000 mouse developmental Cdna microarray. *Proc. Natl Acad. Sci. USA*, **97**, 9127–9132.
16. Bushel, P. R., Hamadeh, H. K., Bennett, L., Green, J., Ableson, A., Misener, S. *et al*. (2002). Computational selection of distinct class- and subclass-specific gene expression signatures. *J. Biomed. Inform.* **35**, 160–170.
17. Bouras, T., Southey, M. C., Chang, A. C., Reddel, R. R., Willhite, D., Glynne, R. *et al*. (2002). Stanniocalcin 2 is an estrogen-responsive gene coexpressed with the estrogen receptor in human breast cancer. *Cancer Res.* **62**, 1289–1295.
18. Han, G. M., Chen, S. L., Shen, N., Ye, S., Bao, C. D. & Gu, Y. Y. (2003). Analysis of gene expression profiles in human systemic lupus erythematosus using oligonucleotide microarray. *Genes Immun.* **4**, 177–186.
19. Kihara, C., Tsunoda, T., Tanaka, T., Yamana, H., Furukawa, Y., One, K. *et al*. (2001). Prediction of sensitivity of esophageal tumors to adjuvant chemotherapy by cDNA microarray analysis of gene-expression profiles. *Cancer Res.* **61**, 6474–6479.
20. Rus, V., Atamas, S. P., Shustova, V., Luzina, I. G., Selaru, F., Magder, L. S. *et al*. (2002). Expression of cytokine- and chemokine-related genes in peripheral blood mononuclear cells from lupus patients by cDNA array. *Clin. Immunol.* **102**, 283–290.
21. Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P. *et al*. (2001). Assessing gene significance from cDNA microarray expression data *via* mixed models. *J. Comput. Biol.* **8**, 625–637.
22. Khan, H. A. (2004). ArraySolver: an algorithm for colour-coded graphical display and Wilcoxon signed-rank statistics for comparing microarray gene expression data. *Comp. Func. Genom.* **5**, 39–47.
23. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
24. Kerr, M. K., Martin, M. & Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. *J. Comput. Biol.* **7**, 819–837.
25. Dooley, T. P., Curto, E. V., Reddy, S. P., Davis, R. L., Lambert, G. W., Wilborn, T. W. *et al*. (2004). Regulation of gene expression in inflammatory bowel disease and correlation with IBD drugs screening by DNA microarrays. *Inflamm. Bowel Dis.* **10**, 1–14.
26. SPSS for Windows, SPSS Inc., Chicago, Illinois, USA.
27. Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometric Bull.* **1**, 80–83.
28. Beenken, S. W., Hockett, R., Jr, Grizzle, W., Weiss, H. L., Pickens, A., Perloff, M. *et al*. (2002). Transforming growth factor-alpha: a surrogate endpoint biomarker. *J. Am. Coll. Surg.* **195**, 149–158.
29. Pfaffl, M. W., Wittmann, S. L., Meyer, H. H. D. &

Bruckmaier, R. M. (2003). Gene expression of immunologically important factors in blood cells, milk cells and mammary tissue of cows. *J. Dairy Sci.* **86**, 538–545.
30. Robinson, P., White, A. C., Lewis, D. E., Thornby, J., David, E. & Weinstock, J. (2002). Sequential expression of the neuropeptides substance P and somatostatin in granulomas associated with murine cysticercosis. *Infect. Immun.* **70**, 4534–4538.
31. Johnston, S. R., Saccani-Jotti, G., Smith, I. E., Salter, J., Newby, J., Coppen, M. *et al*. (1995). Changes in estrogen receptor, progesterone receptor, and pS2 expression in tamoxifen-resistant human breast cancer. *Cancer Res.* **55**, 3331–3338.
32. Thomas, J. G., Olson, J. M., Tapscott, S. J. & Zhao, L. P. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.* **11**, 1227–1236.
33. Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D. & Altman, R. B. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, **18**, 1454–1461.

Edited by J. Karn

(Received 21 September 2004; received in revised form 3 November 2004; accepted 9 November 2004)