QUALITY CONTROL AND REPRODUCIBILITY IN DNA MICROARRAY EXPERIMENTS

ANDRÉ FUJITA[1] andrefujita@riken.jp
JOÃO R. SATO[2] joao.sato@ufabc.edu.br
FERNANDO H.L. DA SILVA[3] lojudice@iq.usp.br
MARIA C. GALVÃO[4] mchrisgalvao@gmail.com
MARI C. SOGAYAR [3] mcsoga@iq.usp.br
SATORU MIYANO[1,5] miyano@ims.u-tokyo.ac.jp

[1] *Computational Science Research Program, RIKEN, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan*
[2] *Center of Mathematics, Computation and Cognition, Universidade Federal do ABC, Rua Satna Adélia, 166 - Santo André*
[3] *Chemistry Institute, University of São Paulo, Av. Lineu Prestes, 748 - São Paulo, 05508-000, Brazil*
[4] *Orthodontic Department, University Methodist of São Paulo, Rua Alfeu Tavares, 149 - São Bernardo do Campo, 09641-000, Brazil*
[5] *Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan*

Biological experiments are usually set up in technical replicates (duplicates or triplicates) in order to ensure reproducibility and, to assess any significant error introduced during the experimental process. The first step in biological data analysis is to check the technical replicates and to confirm that the error of measure is small enough to be of no concern. However, little attention has been paid to this part of analysis. Here, we propose a general process to estimate the error of measure and consequently, to provide an interpretable and objective way to ensure the technical replicates' quality. Particularly, we illustrate our application in a DNA microarray dataset set up in technical duplicates.

Keywords: reproducibility; quality control; technical replicate; Dahlberg's error; DNA microarrays

1. Introduction

Technical replicates are mandatory in most biosciences experiments in order to ensure the results consistency and to identify and avoid possible measurement errors derived from methodological/technical processes. However, little attention has been paid to the analysis and quality control among replicates (for convenience, along all the text, we will refer to "technical replicate" simply as "replicate"). In general, discrimination between acceptable and poor replicates is carried out in a subjective fashion, without interpretable or well-defined criteria and the replicates quality control is not judiciously described, with little attention being paid to this initial step of data analysis.

It is a common sense among biologists to use a pre-defined value to distinguish acceptable replicates from poor replicates. For example, performing a Real Time

RT-PCR experiment in duplicates, one may assume that a difference between the replicates may not be greater than a given *a priori* set Δ. Despite that, this pre-defined number Δ does not differentiate between highly and lowly expressed genes, i.e., it is not proportional to the gene expression value. To illustrate this limitation, suppose $\Delta = 0.5$ and two genes, one with expression values equal to, respectively, 10 and 30. The proportion between 0.5 in 10 and 0.5 in 30 is totally different, i.e., Δ may be more restrictive for low expression genes than high expression ones. In addition, it is difficult to statistically quantify the error of this pre-defined threshold, which varies from one operator to the other.

A second problem is the analysis of the microarray data. It has been described that some DNA microarrays such as Affymetrix, Agilent and Codelink [1, 3] provide Pearson correlation coefficients between replicates greater than 0.9. For other platforms, such as cDNA microarrays or the Mergen platform, the Pearson correlation coefficient between technical replicates varies from a low value of 0.5 and a high value of 0.95 [3, 12, 13]. For a detailed review about reproducibility in microarrays, see Draghici *et al.* [7].

As described above, microarray reproducibility is usually measured by applying Pearson's correlation [16], in which a value close to one indicates good reproducibility, otherwise, a bad reproducibility. However, Pearson's correlation assumes that the variance along the data is equal, when it is known that the variance along the spots in a microarray varies, i.e., there is heteroscedasticity [2, 26]. Moreover, Pearson's correlation is a measure of proportionality, and not a measure of how much the values of the spots from the second microarray are similar to those of the first one (reproducibility). Mean and amplitude biases (systematic errors) cannot be detected by correlation coefficient, therefore, Pearson's correlation cannot be applied to verify reproducibility. Kim *et al.* [15] have proposed the use of Spearman's correlation, nevertheless, analogously to Pearson's, the former is a measure of association and not of error.

In addition, correlation's measures do not identify the poor quality spots or the reproducibility of specific spots, but only provide the general association between the replicated microarrays.

Here, we suggest a general method to estimate the measurement error based on the concept of Dahlberg's error [5], which may be applied in most biological experiments involving technical replicates. Moreover, we present a solution which overcomes some well-known limitations of Dahlberg's error, such as homoscedasticity and incorporation of bias (systematic errors). We also apply the proposed method to actual DNA microarray data in order to illustrate usefulness of this approach. To this end, we model the heteroscedasticity, by providing a quality control criterion for each spot, based on the replicated data.

2. Materials and Methods

Primarily, we will describe an error of measure's estimator based on the Dahlberg's method and its limitations. We then introduce an estimation, based on Support Vector Regression, in order to overcome these limitations, and an algorithm to deal with the heteroscedasticity problem in DNA microarray data.

2.1. *Dahlberg's Error (D.E.)*

Consider the following model:

$$Z_{ij} = \mu_i + \varepsilon_{ij} \tag{1}$$

where Z_{ij} is the measure obtained in one biological experiment, i is the sample index $i = 1, \ldots, N$, j is the replicate number ($j = 1, 2$ in the case of duplicates), μ_i is the unknown true value of the measure and ε_{ij} is the error of measure.

For the error of measure, assume that $E(\varepsilon_{ij}) = 0$ and $Var(\varepsilon_{ij}) = \delta_\varepsilon^2$. Thus, one possible quantification of the measure's quality is the standard deviation of ε_{ij}, i.e., δ_ε. In other words, the lower is the standard deviation of the error of measure, the more reproducible the method is.

Consider

$$d_i = Z_{i2} - Z_{i1} \tag{2}$$

Therefore,

$$Var(d_i) = Var(\varepsilon_{i2} - \varepsilon_{i1}) = 2\delta_\varepsilon^2 \tag{3}$$

Assuming that there is no bias (systematic error), one intuitive estimator for $2\delta_\varepsilon^2$ is

$$2\hat{\delta}_\varepsilon^2 = \sum_{i=1}^{N} \frac{d_i^2}{N} \tag{4}$$

The quantity $\hat{\delta}_\varepsilon = \sqrt{\sum_{i=1}^{N} \frac{d_i^2}{2N}}$ is exactly the Dahlberg's formula proposed in 1940 in order to estimate the error of measure in cephalometric studies [5, 10, 11]. This estimator for the standard deviation of the error of measure is widely used in Orthodontics [5, 6, 10, 11, 14, 20, 22] and may be interpreted as root of squared error's average.

One may use this standard deviation of the error of measure in order to check whether the replicates are similar enough or not. It is known that approximately 95% of the values of a random sample generated from a normal distribution has a mean between $\mu - 1.96\delta_\varepsilon^2$ and $\mu + 1.96\delta_\varepsilon^2$. Consequently, one criterion to verify whether replicates are similar or not may be:

$$T = \frac{1}{2}\left(\frac{1.96\delta_{\varepsilon}}{|Z_{i1}|} + \frac{1.96\delta_{\varepsilon}}{|Z_{i2}|}\right) > \alpha \tag{5}$$

This quantity T indicates that in approximately 95% of the sample the ratio between the error of measure and the observed measure is lower than T, in which T is the proportion of error related to the measured value, allowing evaluation of the performed error.

If T is greater than a defined α, the replicate may be excluded and the experiment should be repeated. Notice that T is proportional to the measured data.

Unfortunately, the Dahlberg's error is extremely affected by systematic errors. Notice that any bias between the two measures is incorporated. Moreover, Dahlberg's error assumes equal means and variances between both measures. Therefore, Dahlberg's error does not discriminate between systematic error (biases in measurement which leads to measured values systematically higher or lower than the true value) and random error (unpredictable fluctuations in the measurements), rendering the interpretation of the results very difficult.

In order to overcome these limitations, we suggest an approach based on Support Vector Regression.

2.2. *Support Vector Regression (SVR)*

SVR is a robust regression developed by Vapnik and Lerner [23] and Vapnik and Chervonenkis [24] and recently applied to Bioinformatics [8, 9, 17].

Let $\{(x_1, y_1), \ldots, (x_i, y_i)\} \subset R \times R$ be the values obtained from biological experiments performed in duplicates, where y is the replicate of x.

In ε-insensitive SVR [25], it is estimated by a function $f(x)$ that has at the most ε deviation from the y_i for all the data, and is as flat as possible.

In other words, the intuitive idea is to define a tube of radius ε around the regression, where $\varepsilon > 0$, and no error is computed if y lies inside the tube. Therefore, outliers are naturally excluded from the regression computation (see Figure 1).

More technically, in the case of linear functions f:

$$f(x) = \langle w^t x \rangle + b \tag{6}$$

with $w \in R^n$, $b \in R$.

Flatness in (6) means small w, i.e., minimize $||w||^2$.

Minimize

$$\frac{1}{2}\|w\|^2 \tag{7}$$

Constrained to

$$\begin{cases} y_i - (w^t x_i) - b \leq \varepsilon \\ (w^t x_i) + b - y_i \leq \varepsilon \end{cases} \tag{8}$$

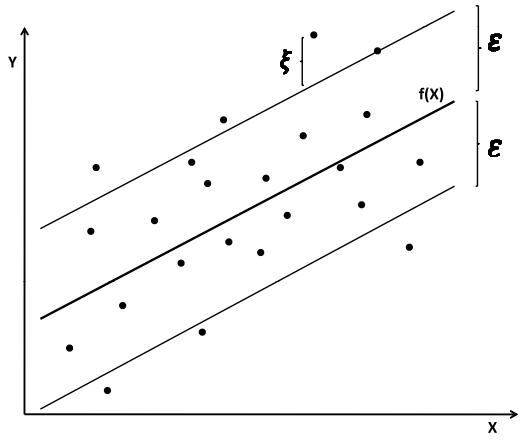

Fig. 1. An illustrative scheme of the linear SVR.

Notice that in (8) there is a function f which, with ε precision, approximates all pairs (x_i, y_i). In the cases where it is necessary to allow for some errors, the problem can be reformulated to [25]:

Minimize

$$\frac{1}{2}\|w\|^2 + C\sum(\xi_i + \xi_i^*) \tag{9}$$

Constrained to

$$\begin{cases} y_i - (w^t x_i) - b \leq \varepsilon_i \\ (w^t x_i) + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \tag{10}$$

where ξ_i, ξ_i^* are slack variables, the constant $C > 0$ is the trade-off between the amount up to which deviations larger than ε are tolerated, maintaining the flatness of f.

For more details about numerical computation and theoretical SVR information, see Smola and Schölkopf [21].

2.3. *Modeling DNA microarray data*

As described in the Introduction, for microarray data, it is known that the variance varies along the spots due to technical problems such as hybridization efficiency, probe sequence, background fluorescence, signal quantification procedures [27, 28], therefore, application of the Dahlberg's formula is not straightforward. In order to overcome this problem, we suggest the following algorithm:

Let X and Y be two DNA microarrays, with Y being the replicate of X.

(1) Perform a non-linear regression which is robust to outliers, namely Support Vector Regression [8] between $log(X)$ and $log(Y)$, i.e., $log(Y) = f(log(X)) + \varepsilon_1$. Notice that the logarithm was calculated due to the high variance observed in microarray data. This is a common practice in microarray data analysis. For other biological replicates, which do not present high variance, such as Real Time RT-PCR, it is not necessary to apply logarithm.
(2) Apply again the Support Vector Regression between ε_1^2 and $log(X)$, i.e., $\hat{\varepsilon}_1^2 = f(log(X)) + \varepsilon_2$.
(3) Calculate $\hat{\delta} = \frac{f(log(X))}{2}$, which is exactly the error of measure. Notice that with this process, we obtain one $\hat{\delta}_i$ for each spot $i = 1, \ldots, N$, where N is the number of spots in the microarray.
(4) Calculate $T_i = \frac{1.96e^{\delta_i}}{f(log(X_i))X_i} + \frac{1.96e^{\delta_i}}{f(log(X_i))Y_i}$, $i = 1, \ldots, N$, where N is the number of spots in the microarray.

It is important to note that this method is based on the normality of the residues, therefore, this condition must be checked.

This SVR-based method may be applied to any replicated data such as Real Time RT-PCR, DNA microarrays, protein quantifications etc.

2.4. *DNA Microarray*

2.4.1. *Cell Lysis and RNA Extraction*

Cell cultures were lysed and their RNA extracted using the Illustra RNAspin Mini RNA Isolation Kit (GE Healthcare), following the manufacturer's instructions. Absorbance ratio at 260/280 ηm was used to assess the RNA purity, a ratio of 1.8-2.0 indicating adequate purity.

2.4.2. *Labeling and purification of targets*

RNA samples were prepared and processed according to protocols supplied by the manufacturer (GE Healthcare). Briefly, cDNAs were synthesized from purified RNA (1μg) and control bacterial mRNAs. Samples were purified using the QIAquick Spin Kit (Qiagen) and concentrated by SpeedVac. Concentrated pellets were used in a biotinylated-UTP based cRNA synthesis using the CodeLink$^{\text{TM}}$ Expression Assay Reagent Kit (GE Healthcare). Labeled cRNAs were purified using RNeasy Kit (Qiagen) and fragmented with supplied solution at 94$^{\text{o}}$C for 20 min.

2.4.3. *Hybridization and washing of the DNA arrays*

Fragmented biotin-labeled cRNAs (10μg) were incubated with CodeLink$^{\text{TM}}$ bioarrays under agitation (300 rpm) for 20h. The bioarrays were then washed and incubated with Cy5-streptavidin (30 min). Scanning of the bioarrays was performed using a GenePix 4000 B Array Scanner (Axon Instruments) and the data were collected using the CodeLink$^{\text{TM}}$ System Software (GE Healthcare), which provides the

raw data and invalidated data from irregular spots.

3. Results and Discussions

The proposed error's estimator was applied to two sets of DNA microarrays, both set up in technical duplicates, in order to illustrate the application.

In Figure 2, the first set of microarrays experiments in duplicates is presented, constituting an example of an acceptable duplicate of microarrays, since the proportion of rejected spots, i.e., number of rejected spots/total number of spots was ∼11%. The SVR fitted curve is in blue and the rejected spots (poor quality spots) in red. Notice in Figure 2A that some bias occurs in microarrays experiments. Moreover, in Figure 2B, it is possible to verify that the variance is not constant, i.e., a modest heteroscedasticity is found in the data.

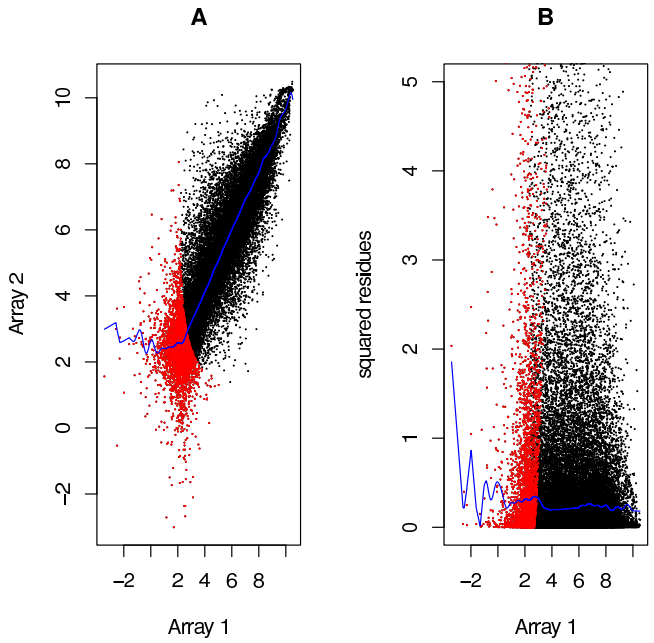

Fig. 2. An example of acceptable reproducibility in duplicated microarrays. A) Fluorescence intensities of Array 1 versus Array 2 plot of the first set of duplicates in log scale; B) Array 1 versus residues of the first set of duplicates with Array 1 in log scale. In red, the spots with $T > 10\%$; in blue, the fitted curve.

Similar characteristics may also be observed in the second set of microarrays (Figure 3). Considering an $\alpha = 10\%$, the proportion of rejected spots was ∼38%, therefore, this second set has a lower quality than the first one. It is interesting that the variance is clearly high in the low signal spots (Figure 3B).

Notice that, in both cases, most of the spots marked in red display low expression, which is to be expected, since it is known that low signal spots display higher

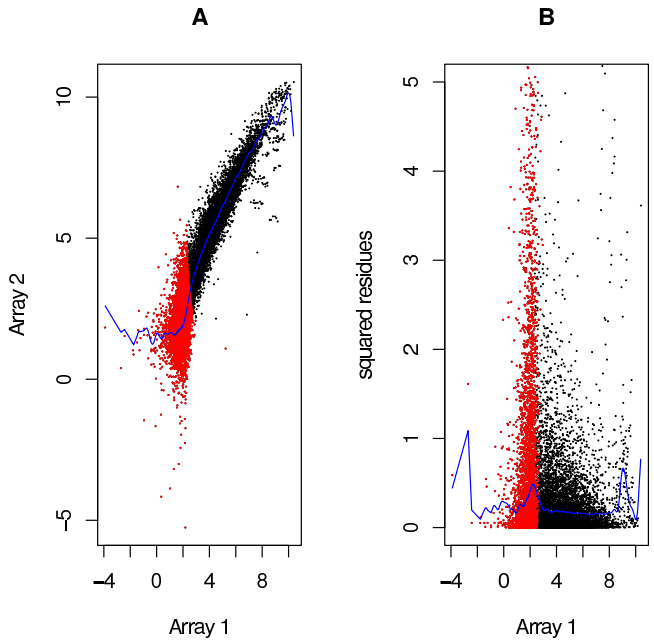

Fig. 3. An example of poor reproducibility in duplicated microarrays. A) Fluorescence intensities of Array 1 versus Array 2 plot of the second set of duplicates in log scale; B) Array 1 versus residues of the second set of duplicates, with Array 1 in log scale. In red, the spots with $T > 10\%$; in blue, the fitted curve.

variance than high signal ones [26]. Therefore, it is not possible to distinguish the true signal from the noise background signal when the spot signal is low.

By calculating the Pearson's correlation for both experiments, we have obtained $\rho_1 = 0.90$ (p-value < 0.01) and $\rho_2 = 0.93$ (p-value < 0.01), respectively. Therefore, using the Pearson's correlation results, one may conclude that the replicates are acceptable, however, the second set of replicates was previously described by our estimator of error of measure as being unsatisfactory. This illustrates the fact that high association (correlation) between duplicates is not the same as reproducibility.

In order to quantify the quality of the spots, several microarray analysis softwares, such as CodeLink^{TM} System Software [18], ArrayVision v.8.0 (Imaging Research Inc, Ontario, Canada) and TM4 [19] were developed. In most of them the analysis is based on the ratio between the background signal and the spot signal, and on the spot's shape (if its shape is well defined) to distinguish acceptable ones. It is a very important step, which should be performed, but, unfortunately, with these approaches it is not possible to identify spots with high or low error in measure derived from technical problems.

SVR does not require a high performance computer to be calculated, i.e., it may be computed in a personal computer. If millions of measurements per chip become available, by comparing a pair of technical replicates, we estimate that SVR may

take of the order of a few minutes to compute using a personal computer, i.e., it is totally feasible in practice.

Identification of high noise spots using replicates may be an additional criterion to discard spots which may influence the following steps in the microarray analysis. Therefore, sometimes, it is not necessary to discard the whole microarray due to a few genes which present poor measures. One may consider to discard the whole microarray if the number of rejected spots is higher than a certain ratio (number of rejected spots / total number of spots), for example, 20%. Otherwise, one may discard only the rejected spots.

Here, we have illustrated our estimator of error of measure for duplicates, but for more than duplicates, it may be obtained, in a straightforward manner, by calculating all d_i combinations among replicates.

In summary, we propose an interpretable and useful method in order to distinguish acceptable replicates from poor replicates. In addition, we have presented a solution to overcome some well known problems within Dahlberg's error and modeled the heteroscedasticity present in microarrays. Our illustrative examples are focused in gene expression data. However, the proposed general method may be applied to any quantification procedure, such as, protein's quantification or Real Time RT-PCR experiments.

3.1. *Acknowledgements*

This work was supported by RIKEN, Japan; and by Brazilian research agencies (FAPESP, CNPq, FINEP and CAPES).

References

[1] Bakay, M., Chen, Y.W., Borup, R., Zhao, P., Nagaraju, K. and Hoffman, E.P., Sources of variability and effect of experimental approach on expression profiling data interpretation, *BMC Bioinformatics*, 3:4, 2002.

[2] Balagurunathan, Y., Dougherty, E.R., Yidong, C., Bittner, M.L. and Trent, J.M., Simulation of cDNA microarrays via a parameterized random signal model, *Journal of Biomedical Optics*, 7:507-523, 2002.

[3] Bammler, T. et al., Standardizing global gene expression analysis between laboratories across platforms, *Nat Methods*, 2:351-356, 2005.

[4] Chirgwin, J.M., Prybyla, A.E., Macdonald, R.J. and Rutter, W.J., Isolation of biologically active ribonucleic acid from sources enriched in ribonucleases, *Biochem*, 18:5294-5299, 1979.

[5] Dahlberg, G., *Statistical methods for medical and biological students*, Interscience Publications, New York, 1940.

[6] Dellavia, C., Sforza, C., Orlando, F., Ottolina, P., Pregliasco, F. and Ferrario, V.F., Three-dimensional hard tissue palatal size and shape in Down syndrome subjects, *The European Journal of Orthodontics*, 29:417-422, 2007.

[7] Draghici, S., Khatri, P., Eklund, A.C. and Szallasi, Z., Reliability and reproducibility issues in DNA microarray measurements, *TRENDS in Genetics*, 22:101-109, 2005.

[8] Fujita, A., Sato, J.R., Rodrigues, L.O., Ferreira, C.E. and Sogayar, M.C., Evaluat-

ing different methods of microarray data normalization, *BMC Bioinformatics*, 7:469, 2006.

[9] Kiryu, H., Oshima, T. and Asai, K., Extracting relations between promoter sequences and their strengths from microarray data, *Bioinformatics*, 21:1062-1068, 2004.

[10] Houston, W.J.B., The analysis of errors in orthodontic measurements, *American Journal of Orthodontics*, 83:382-390, 1983a.

[11] Houston, W.J.B., *Walther's orthodontic notes*, 4th ed. Wright PGS, Bristol., 1983b.

[12] Jarvinen, A.K., Hautaniemi, S., Edgren, H., Auvinen, P., Saarela, J., Kallioniemi, O.P. and Monni, O., Are data from different gene expression microarray platforms comparable?, *Genomics*, 83:1164-1168, 2004.

[13] Jenssen, T.K., Langaas, M., Kuo, W.P., Smith-Sorensen, B., Myklebost, O. and Hovig, E., Analysis of repeatability in spoted cDNA microarrays, *Nucleic Acids Res.*, 30:3235-3244, 2002.

[14] Kamoen, A., Dermaut, L. and Verbeek, R., The clinical significance of error measurement in the interpretation of treatment results, *European Journal of Orthodontics*, 23:569-578, 2001.

[15] Kim, B.S., Rha, S.Y., Cho, G.B., and Chung, H.C., Spearman's footrule as a measure of cDNA microarray reproducibility, *Genomics*, 84:441-448, 2004.

[16] Naderi, A., Ahmed, A.A., Barbosa-Morais, N.L., Aparicio, S., Brenton, J.D. and Caldas, C., Expression microarray reproducibility is improved by optimising purification steps in RNA amplification and labelling, *BMC Genomics*, 5:9, 2004.

[17] Qiu, J., Sheffler, W. and Noble, W.S. Ranking predicted protein structures with support vector regression, *Proteins*, 71:1175-1182, 2008.

[18] Ramakrishnan, R., Dorris, D., Lublinsky, A., Nguyen, A., Domanus, M., Prokhorova, A., Gieser, L., Touma, E., Lockner, R., Tata, M., Zhu, X., Patterson, M., Shippy, R., Sendera, T.J. and Mazumder, A., An assessment of Motorola CodeLink^{TM} microarray performance for gene expression profiling applications, *Nucleic Acids Research*, 30:e30, 2002.

[19] Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V. and Quackenbush, J., TM4: a free, open-source system for microarray data management and analysis, *BioTechniques*, 34:374-378, 2003.

[20] Sforza, C., Peretta, R., Grandi, G., Ferronato, G. and Ferrario, V., Three-dimensional facial morphometry in skeletal Class III patients: A non-invasive study of soft-tissue changes before and after orthognathic surgery, *British Journal of Oral and Maxillofacial Surgery*, 45:138-144, 2007.

[21] Smola, A.J. and Schölkopf, B. A tutorial on support vector regression, *Statistics and Computing*, 14:199-222, 2004.

[22] Sonnesen, L. and Bakke, M., Molar bite force in relation to occlusion, craniofacial dimensions, and head posture in pre-orthodontic children, *European Journal of Orthodontics*, 27:58-63, 2005.

[23] Vapnik, V. and Lerner, A., A pattern recognition using generalized portrait method, *Automatic and Remote Control*, 24:774-780, 1963.

[24] Vapnik, V. and Chervonenkis, A., A note on one class of perceptrons, *Automation and Remote Control*, 25, 1964.

[25] Vapnik, V., *Statistical learning theory*, New York: Wiley, 1998.

[26] Wu, T.D., Large-scale analysis of gene expression profiles, *Briefings in Bioinformatics*, 3:7-17, 2002.

[27] Yang, Y.H., Buckley, M.J., Dudoit, S. and Speed, T.P., Comparison of methods for

image analysis on cDNA microarray data, *Journal of Computational and Graphical Statistics*, 11:108-136, 2002.

[28] Yuk, F.L. and Cavalieri, D., Fundamentals of cDNA microarray data analysis, *Trends in Genetics*, 19:649-659, 2003.