BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 9 2010, pages 1185–1191
doi:10.1093/bioinformatics/btq104

Gene expression

Advance Access publication March 8, 2010

Reducing the algorithmic variability in transcriptome-based inference

Salih Tuna and Mahesan Niranjan*

School of Electronics and Computer Science, University of Southampton, Southampton, UK

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: High-throughput measurements of mRNA abundances from microarrays involve several stages of preprocessing. At each stage, a user has access to a large number of algorithms with no universally agreed guidance on which of these to use. We show that binary representations of gene expressions, retaining only information on whether a gene is expressed or not, reduces the variability in results caused by algorithmic choice, while also improving the quality of inference drawn from microarray studies.

Results: Binary representation of transcriptome data has the desirable property of reducing the variability introduced at the preprocessing stages due to algorithmic choice. We compare the effect of the choice of algorithms on different problems and suggest that using binary representation of microarray data with Tanimoto kernel for support vector machine reduces the effect of the choice of algorithm and simultaneously improves the performance of classification of phenotypes.

Contact: mn@ecs.soton.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 1, 2009; revised on January 31, 2010; accepted on March 3, 2010

1 INTRODUCTION

A plethora of computational methods for the statistical analysis of high-throughput gene expression measurements is available to users interested in making inferences from transcriptomes. The preprocessing stages involved in the analysis include background correction, within and between-array normalizations, probe-specific correction and summarization. After the raw data is processed it is subject to sophisticated machine learning approaches such as classification, cluster analysis and the modelling of time-course data by means of dynamical systems. The preprocessing stages lead to quantifications of relative mRNA abundances, taking scanned images as input. The choices made here have been shown to have an important effect on the results of statistical inference approaches (Allison *et al.*, 2006; Barash *et al.*, 2004; Choe *et al.*, 2005; Cope *et al.*, 2004; Millenaar *et al.*, 2006; Ploner *et al.*, 2005; Qin *et al.*, 2006; Shedden *et al.*, 2005). We review the effect of the choice of algorithms on different problems and suggest that using binary representation of microarray data with Tanimoto kernel for support vector machine (SVM) reduces the variability due to the choice of algorithm and simultaneously improves the performance of classification on transcriptome data.

The most extensive study so far that shows significant algorithmic variability is due to P.C. Boutros[1] who analysed an impressive 19 446 different combinations of preprocessing algorithms, and quantified the sensitivity and stability of expression levels. While this and similar studies (Choe *et al.*, 2005) seek the best combination of algorithms on a small number of datasets, they do not offer a generic solution for practitioners to select a combination that leads to reliable results in downstream inference.

Our approach, starting from the premise of seeking sensible numerical precisions to represent microarray data, is similar to the ‘bar code’ method advanced by Zilliox and Irizarry (2007). In our previous work (Tuna and Niranjan, 2010), we established that gene expressions quantized to binary precision lead to minimal average loss in the quality of inference drawn from them, in a range of applications such as classification, clustering and the analysis of time-course data. Any loss of information due to binarization was shown to be easily recovered using metrics of similarity between gene expression profiles that are best suited for high-dimensional binary spaces (Tuna and Niranjan, 2009). Our results showed that the Tanimoto metric, successfully used in matching chemical fingerprints in the chemoinformatics literature (Willett, 2006), when cast in kernel SVM (Ralaivola *et al.*, 2005; Swamidass *et al.*, 2005; Trotter, 2006) and spectral clustering frameworks, is able to achieve performances often better than, and never worse than, using data to the high numerical precision with which it is often reported and archived.

1.1 Algorithmic variability during the preprocessing of raw data

Several studies have explored the effect of algorithm choice for preprocessing microarray data. The general consensus in the community is that the overall results are highly dependent on the algorithms used for preprocessing data. Most such studies focus on detecting differentially expressed genes using a single dataset (usually spike-in data) and compare the results of different algorithms in quantifying the spiked-in concentration. A review of this body of literature finds contradictory claims on the relative performance of algorithms. Part of the explanation for such contradictions arises from comparisons being made on different datasets at the level of observed gene expression.

*To whom correspondence should be addressed.

[1] Boutros P.C., Microarray Gene Expression Society Meeting (MGED), Riva del Garda, Italy (2008).

Irizarry *et al.* (2003b) introduced the robust multiarray analysis (RMA) method for analysing Affymetrix data and compared it to previously used methods (dCHIP and MAS5.0). This study mentions there is no standard way to compare the effect of the algorithm choice, and evaluates the results using three criteria on a single dataset. While in these evaluations RMA performs best amongst the three algorithms, Choe *et al*.'s (2005) study provides contradictory results on the same algorithms, but using a different dataset. Choe *et al.* (2005) use the `affy` package and consider 152 different combinations of algorithms and report their best combination, which turns out to be none of the above three. Specifically with respect to adjusting the perfect match (PM) probe intensity, the authors favour the performance of MAS5.0, and explicitly note a contradiction with Irizarry *et al*.'s (2003b) claim. A recent study by Pearson (2008) also notes contradictory claims on the results obtained from spiked-in data. Other examples of contradictory claims can be seen in Shedden *et al.* (2005), Ploner *et al.* (2005), Qin *et al.* (2006) and Millenaar *et al.* (2006). Noting that the algorithms compared lead to different levels of gene expressions, the common recommendation of those studies' for the user is to test many possible algorithms before arriving at gene expression levels. This, however, has serious computational implications and is generally infeasible.

1.2 Motivation

As reviewed above, gene expression values obtained by different preprocessing algorithms yield different results. We use a gene selection problem to illustrate the impact of this variation. We took the dataset of a breast cancer study (West *et al.*, 2001) and computed the 50 most discriminant genes that could be used as hypothetical markers for discrimination. These genes were selected using the Fisher's ratio as criterion, as is commonly done (Golub *et al.*, 1999). We notice substantial differences in the genes that were identified as carrying the most discriminant information in three different algorithmic combinations, as shown in the confusion matrix of Table 1. The three algorithmic combinations chosen here correspond to those achieving the minimum, maximum and mean class prediction performance of an SVM classifier, measured in terms of the area under the receiver operating characteristics curve (AUROC). The names of the genes selected are given in Supplementary Table S4. The selected genes between the best and the worst performing classifiers overlap by only about 50%.

This observation motivates a systematic study to explore the variability in inference quality caused by the choice of algorithms in the preprocessing stages, and possible approaches to reducing this variability. As we report in the remainder of this article, a binary representation with kernel classification in high-dimensional spaces is able to achieve a drastic reduction in the variability caused by algorithmic choice.

Table 1. Confusion matrix of the number of common most discriminant genes when different preprocessing algorithms are used

	Max (AUROC)	Min (AUROC)	Mean (AUROC)
Max (AUROC)	50 / 50	34 / 50	27 / 50
Min (AUROC)	34 / 50	50 / 50	24 / 50
Mean (AUROC)	27 / 50	24 / 50	50 / 50

Fifty most discriminant genes are selected by using the Fisher's ratio. The three preprocessing sets compared are the ones giving minimum, maximum and mean class prediction performance in terms of their AUROC.

2 APPROACH

2.1 Experiments

Our approach to demonstrating the variability due to preprocessing algorithm choice, and its reduction by a binarized data representation, was to perform training and testing of large numbers of classifiers on eight datasets taken from the archives. To this end, we restricted ourselves to datasets from Affymetrix arrays because of the very wide range of preprocessing algorithms for this class of arrays available in R packages. We were also constrained by the need for a reasonable dataset size (we used datasets with a minimum size of 20) to allow reliable cross validations by partitioning into training and test sets.

We used SVMs as classifiers. Microarray data being usually of high dimensions, SVMs tend to be the classifiers of choice in the machine learning literature (Brown *et al.*, 2000; Dettling, 2004; Statnikov *et al.*, 2005). For comparison with an alternative, we implemented K-nearest neighbour classifiers. We found the SVMs consistently outperformed the nearest neighbour approach in these problems (see results in Supplementary Material). Since the nearest neighbour approach suffers from the well-known curse of dimensionality issue (Bishop, 2006; Duda *et al.*, 2001) any difference in performance as a result of data representation will be masked by this aspect. We also considered the use of distance-to-template classifiers, in the light of Zilliox and Irizarry's (2007) bar-code work, and chose not to pursue this for a technical reason supported by our previous work (Tuna and Niranjan, 2009). It is a basic result in statistical pattern recognition that distance-to-template classifiers, where the templates stored are class-means, rely on the assumption of isotropic (each feature being independent) and equal variances (Duda *et al.*, 2001) for optimal performance. With microarray data, this assumption is unlikely to be true because genes are bound to show correlated expression arising from, for example, co-regulation by common transcription factors. See Supplementary Material for results of nearest neighbour and distance-to-template classifiers as a function of dimensionality. Kernel classifiers such as SVMs do not rely on such assumptions and are able to form complex class boundaries by capturing higher order correlations in the feature space. We report results using the AUROC as the performance measure, and for each preprocessing method, computed AUROCs by averaging over 50 random partitionings sampled with replacement of the data into training and test sets. Some authors restrict their results to single training and evaluation sets, often the partition determined by the original suppliers of the data. Given data sizes are usually small, this practice runs the risk of 'reporting good news' when an algorithm accidentally performs well on the particular test set. Averaging over 50 bootstrap partitions avoids this pitfall. In the above experimental setting, we quantified the variability in classifier performance using the continuous valued data, data quantized to binary precision using a linear kernel SVM, and binary data classified using a Tanimoto kernel SVM. The Tanimoto similarity metric is borrowed from the chemoinformatics literature where it has been used very effectively in comparing molecules using fingerprints of their chemical properties. Thus, it is seen as a good metric for high-dimensional binary spaces.

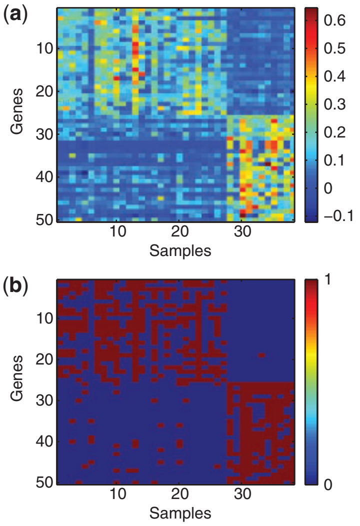

Fig. 1. Heatmap showing how binarization remove noise in gene expression measurements: (**a**) data represented to a continuous scale; and (**b**) data reduced to a binary representation by quantization. Source: Golub *et al.* (1999).

Trotter (2006) shows how such a similarity metric can be used in a kernel SVM framework, and our own previous work has shown this metric yields good results for microarray data as well (Tuna and Niranjan, 2009, 2010).

Figure 1 is an illustration, in the form of 'heatmaps', of the effect of binarization in gene expression. Noise in the continuous valued data Figure 1a is lost when expression levels are discretized as in Figure 1b. The data used here is from the widely studied acute lymphoblastic leukemia (ALL) versus acute myeloid leukemia (AML) classification problem of Golub *et al.* (1999). While such a 'heatmap' illustration shows that discrimination between the classes is retained with binarized data, the reader should be cautious that the actual classifier design is taking place in a very high-dimensional space that cannot be visualized directly.

One difficulty encountered in comparing multiple preprocessing algorithms is the fact that not all algorithms lead to valid gene expressions on all datasets. We found implementations of several algorithmic combinations failed to complete analysis of the input data. Notably there was no consistency in these failures, and hence of the 315 combinations available in the package we used only 179 algorithmic combinations that were common across all eight datasets for our main results. Such an observation is also made in Bolstad (2004). We report results using this common set, but note that the numbers of algorithms used was not a factor in the conclusions we are able to reach.

3 MATERIALS AND METHODS

3.1 Datasets

All eight datasets considered in this study used Affymetrix microarray platforms and raw data are available as CEL files. All probes in the array were used in setting up classification experiments and no feature selection is applied prior to classification. Some combinations of preprocessing algorithms returned missing values for some gene expressions (reported by the algorithms as N/A). In these cases, the gene in question was removed from the classifier design for that particular preprocessing algorithm. Below is the description of each dataset used and the GEO accession number or the authors' web page where these datasets are freely available to download:

Table 2. List of possible methods in `expresso`

Normalization	Background correction	Probe specific correction (PM)	Summary method
constant	mas	mas	avgdiff
contrasts	none	pmonly	liwong
invariantset	rma	subtractmm	mas
loess			medianpolish
qspline			playerout
quantiles			
quantiles.robust			

- Two prostate cancer datasets: the first, GSE6956, has 22 277 probe-sets with 89 samples, of which 69 are prostate and 20 are normal (Wallace *et al.*, 2008). The second has 12 625 probe-sets with 102 samples, of which 52 are prostate and 50 are normal (http://www.broad.mit.edu/; Singh *et al.*, 2002).
- Two lung cancer datasets: the first, GSE7670, has 22283 probe-sets with 66 samples, of which 30 are normal and 36 are cancer (Su *et al.*, 2007). The second, GSE10072, also has 22283 probe-sets with 107 samples, of which 58 are cancer and 49 are normal (Landi *et al.*, 2008).
- Two breast cancer datasets: the first, GSE5847, has 22283 probe-sets with 95 samples, of which 47 are normal and 48 are cancer (Boersma *et al.*, 2007). Second, obtained from http://data.genome.duke.edu/west.php, has 7 129 genes and 49 samples, (25 ER^{+} and 24 ER^{-}; West *et al.*, 2001).
- Lymph node versus tonsil problem: (GSE2665) 22283 probe-sets with 20 samples; 10 lymph node and 10 tonsils (Martens *et al.*, 2006).
- Childhood ALL: (GSE3910) 22283 probe-sets with 70 samples; 35 for each diagnosis and relapse (Bhojwani *et al.*, 2006).

Amongst these eight datasets, five used Affymetrix array `U133A`, which contains 22283 probe-sets. Singh *et al.*'s (2002) and West *et al.*'s (2001) data come from `Hum95AV2` and `HuGeneFL` arrays, respectively. Lastly, Wallace *et al.*'s (2008) dataset used `U133A2.0` array. These differences cause the total numbers of probe-sets to differ between the datasets.

3.2 Preprocessing algorithms

To select different preprocessing algorithms, we used the `expresso` module implemented in the `affy` package (Gautier *et al.*, 2004; Irizarry *et al.*, 2003a) available in Bioconductor (http://www.bioconductor.org/). In this implementation, the different stages of the preprocessing pipeline available, with the numbers of algorithms at each stage given in parentheses, are: background correction (3), normalization (7), probe-specific correction (3) and summary method (5), giving a total of 315 different combinations, as shown in Table 2.

3.3 Quantization

For quantization of the continuous data into binary, we used the mixture Gaussian approach developed by Zhou *et al.* (2003). While there are several alternative approaches to quantizing, which usually rely on setting some threshold, this particular method was preferred because it uses a probability density model of the data, making it a more principled approach. Mixture models with two components, defined as

$$P(\mathbf{x}) = \sum_{k=1}^{M} \pi_k \mathrm{N}\left(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\right), \tag{1}$$

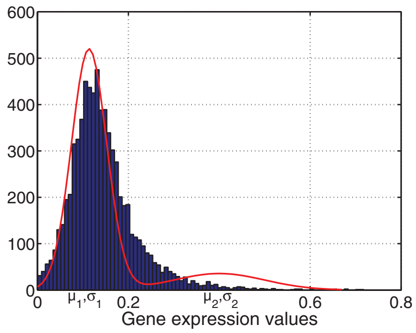

Fig. 2. Mixture Gaussian distributions and corresponding histograms of gene expression levels for an array taken from West *et al.*'s (2001) breast cancer data.

($M = 2$) were fitted to log-transformed gene expression data, and the corresponding means and SDs of the models were calculated. Using these parameters, a quantization threshold was set as

$$\theta = \frac{\mu_1 + \sigma_1 + \mu_2 - \sigma_2}{2}. \tag{2}$$

Fitting the model was done by standard maximum likelihood techniques, and we used the gmm function in NETLAB software (http://www.ncrg.aston.ac.uk) for this purpose. Quantization was applied on an array by array basis. While there are other choices (gene by gene or a global threshold across all the data), our previous experience showed there was not much difference in performance over a range of threshold settings (Tuna and Niranjan, 2010). An example of a Gaussian Mixture Model (GMM) fit to a histogram of gene expressions is shown in Figure 2.

3.4 Tanimoto kernel

Tanimoto similarity for two genes x and y is calculated as $T = c/(a+b-c)$, where a is the number of arrays in which x is expressed, b the number of arrays in which y is expressed and c the number of arrays in which both x and y are expressed. This similarity metric has been used as a kernel for SVMs (Ralaivola *et al.*, 2005; Swamidass *et al.*, 2005; Trotter, 2006). The Tanimoto kernel is defined as

$$K_{\mathrm{Tan}}(\mathbf{x}, \mathbf{z}) = \frac{\mathbf{x}^T \mathbf{z}}{\mathbf{x}^T \mathbf{x} + \mathbf{z}^T \mathbf{z} - \mathbf{x}^T \mathbf{z}}. \tag{3}$$

We implemented Tanimoto kernel in the MATLAB SVM package of Steve Gunn (http://www.isis.ecs.soton.ac.uk/isystems/kernel/) and the results of the classification were evaluated by means of AUROC. Classification experiments were performed using 50 random partitions of the data into training and test sets. Sampling with replacement was used to construct these bootstrap datasets.

3.5 Fisher's ratio

The Fisher's ratio is defined as

$$\text{Fisher's ratio} = \frac{\mathrm{abs}(\mu_1 - \mu_2)}{\sigma_1 + \sigma_2}, \tag{4}$$

where μ_1, σ_1 and μ_2, σ_2 are the means and SDs of the first and second classes, respectively. The Fisher's ratio as a measure of the ability of a feature to separate two classes is in common use in microarray class prediction literature, starting from the work (Golub *et al.*, 1999).

3.6 Statistical significance

To compare statistical significances of the difference in variances, we used the F-test as implemented in MATLAB, and similarly used the t-test to test for significance in differences between means.

4 RESULTS

Figure 3 shows the classification results obtained with different preprocessing algorithms using AUROC as criterion on eight different class prediction tasks. These box plots show that binarization consistently achieves a reduction in the variability of classifier performance. Most importantly, we note substantial improvement in a number of outlier results (Fig. 3f–h) owing to the binary data representation.

Table 3 and 4 show the variance and mean respectively achieved by the three methods across the 179 different preprocessing algorithms. Statistical significance of differences in these, where the continuous and binary linear methods are compared against the binary Tanimoto approach are shown as P-values of F-tests. We note that on all but one of the eight problems considered here [the exception being Wallace *et al.* (2008)], variability caused by algorithmic choice was significantly reduced ($P < 0.001$) by the binarization of the data. However, there is significant improvement in the average performance in this dataset (see Supplementary Table S1 for the differences in mean performances and associated levels of statistical significances).

5 DISCUSSION

Why does binarization help, in reduction in variability as we have observed here, and in prediction of tissue types as demonstrated in Zilliox and Irizarry (2007)? We suggest that a low numerical precision representation is more compatible with the environment from which microarray data are gathered, than the arbitrary length of decimal places to which they are usually reported and archived. Except in a small number of cases like cell-cycle regulation studies, where the cellular states are artificially synchronized, mRNA extraction is from a heterogeneous population of cells, each of which having a small number of copies of the mRNA species. This causes large variation in measurements across different sub-populations from the same biological sample—the so-called biological variability. A critical survey of microarray studies by Draghici *et al.* (2006) concluded: '...the existence and direction of gene expression changes can be reliably detected for the majority of genes. However, accurate measurements of absolute expression levels and the reliable detection of low abundance genes are currently beyond the reach of microarray technology'. As such, when the data representation precision matches the reality of the underlying measurement, several advantages follow, of which the algorithmic variability reduction reported here is one.

Note that the reduction in variability of performance we claim has only been demonstrated with SVM as classifiers. We did not observe this variability with the nearest neighbour classifier, overall performance of which were far worse than SVM. Hence this variability reduction should be seen as resulting from a combination of quantization and a high-dimensional representation separated by large margins (the optimization criterion of SVM). On the expression of an individual gene on a specific patient, however, binarization does not always give consistent ON/OFF outcome. This is because the quantization threshold is chosen by fitting a model through a distribution of expressions after a particular pre-processing combination is applied. Thus, we caution the reader against interpreting the results at the

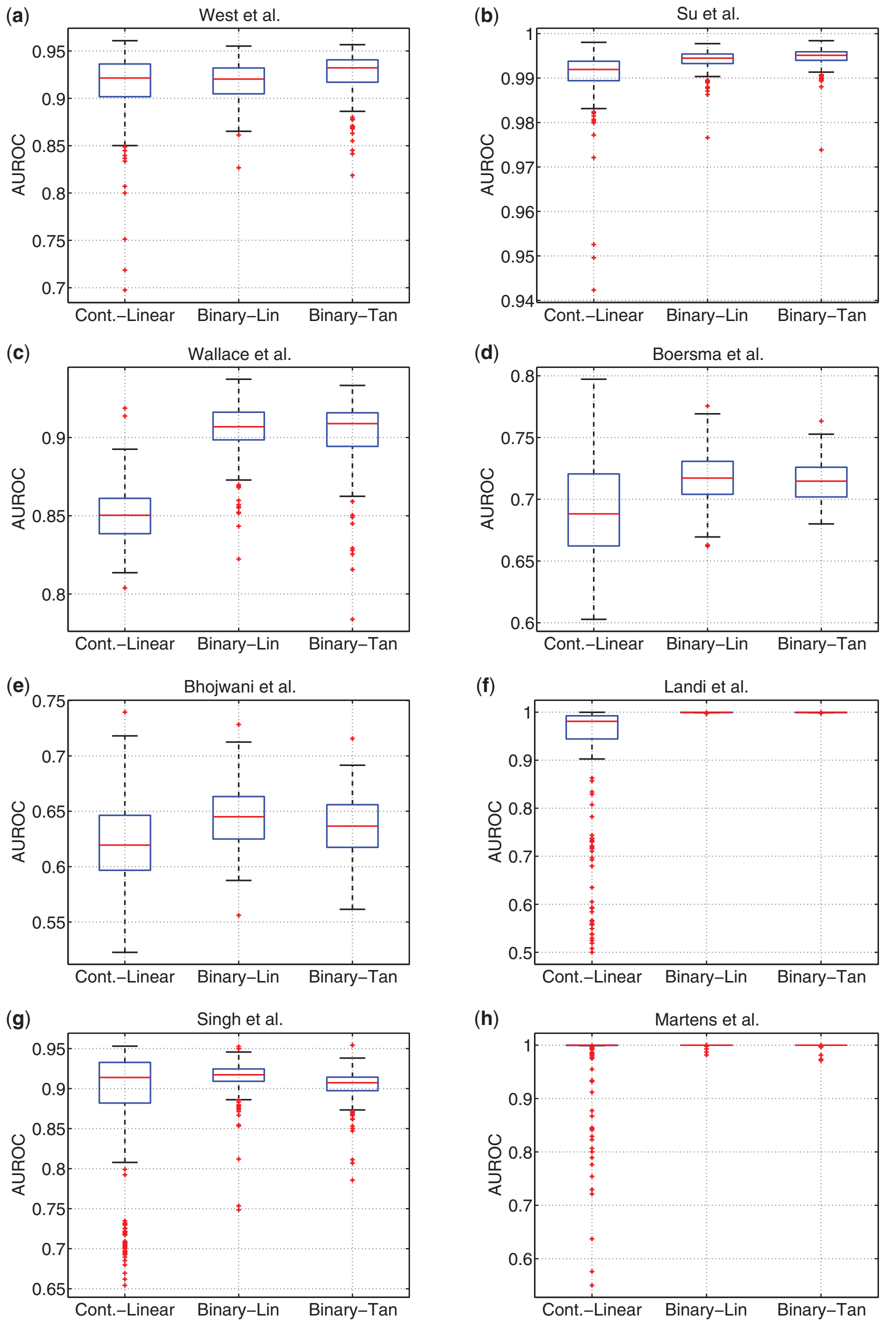

Fig. 3. Variability in classification results due to the choice of preprocessing algorithms for inference from continuous data and two forms of binarized approaches. Each point of the boxplot corresponds to the classification result of a particular combination of algorithms in the preprocessing pipeline. This result is averaged over 50 re-partitionings of the data into training and test sets. All eight datasets have been analysed with the same 179 preprocessing algorithms.

level of detecting gene expressions to binary precision in a consistent way.

6 CONCLUSION

In this article, we address the variability in the results of inferences from transcriptome studies that arises from the choice of preprocessing algorithms. This variability has previously been shown to have a significant effect on the gene expressions measured by microarrays. Our work reported here shows that the binary representation helps significantly in reducing this algorithmic choice induced variability in classification problems, when used in combination with a high-dimensional kernel method.

While previous studies have largely focused on pointing out that such a variability exists by reference to how well-measured expressions of a gene correlates with spiked-in concentration or with

Table 3. Summary of the variability due to the choice of preprocessing algorithms

Study	Continuous (SD; *P*-value)	Binary linear (SD; *P*-value)	Binary Tanimoto
Martens *et al.* (2006)	0.0743; < 0.001	0.0018; 1	0.0032
Su *et al.* (2007)	0.0064; < 0.001	0.0024; 0.6	0.0023
Wallace *et al.* (2008)	0.0169; 0.99	0.0177; 0.99	0.0222
West *et al.* (2001)	0.0392; < 0.001	0.0211; 0.92	0.0234
Singh *et al.* (2002)	0.0882; < 0.001	0.0241; 0.03	0.0212
Bhojwani *et al.* (2006)	0.0408; < 0.001	0.0264; 0.54	0.0266
Landi *et al.* (2008)	0.1391; < 0.001	0.0004; 0.001	0.0003
Boersma *et al.* (2007)	0.0457; < 0.001	0.0197; 0.05	0.0174

SDs of the AUROCs, shown as box plots in Figure 3 are shown. Statistical significances using *F*-test show levels of confidence at which our proposed method of binary Tanimoto differs from the alternate approach.

Table 4. Comparison of the mean AUROCs when different preprocessing algorithms are used

Study	Continuous (mean; *P*-value)	Binary linear (mean; *P*-value)	Binary Tanimoto
Martens *et al.* (2006)	0.98; < 0.001	0.99; 0.80	0.99
Su *et al.* (2007)	0.99; < 0.001	0.99; 0.001	0.99
Wallace *et al.* (2008)	0.85; < 0.001	0.90; 0.89	0.90
West *et al.* (2001)	0.91; < 0.001	0.92; 0.001	0.93
Singh *et al.* (2002)	0.88; < 0.001	0.91; 1	0.90
Bhojwani *et al.* (2006)	0.62; < 0.001	0.64; 0.99	0.64
Landi *et al.* (2008)	0.92; < 0.001	0.999; 0.04	0.999
Boersma *et al.* (2007)	0.70; < 0.001	0.71; 0.95	0.71

an alternate measurement such as qPCR (quantitative Polymerase Chain Reaction), we have focused on the quality of inference drawn, in the context of classification problems. To the best of our knowledge, comparisons of inference algorithms on microarray data (e.g. support vector machines versus nearest neighbour as predictors of phenotype), of which there is a very large body of literature in the statistical and machine learning communities, are based on the application of one set of preprocessing algorithms, often published by the original authors of a particular study. Given the variability we observe, we believe there is room for scepticism of conclusions drawn from such studies.

Conflict of Interest: none declared.

REFERENCES

Allison,D.B. *et al.* (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, **7**, 55.

Barash,Y. *et al.* (2004) Comparative analysis of algorithms for signal quantitation from oligonucleotide microarrays. *Bioinformatics*, **20**, 839–846.

Bhojwani,D. *et al.* (2006) Biologic pathways associated with relapse in childhood acute lymphoblastic leukemia: a Children's Oncology Group study. *Blood*, **108**, 711–717.

Bishop,C.M. (2006) *Pattern Recognition and Machine Learning*. Springer, New York.

Boersma,B.J. *et al.* (2007) A stromal gene signature associated with inflammatory breast cancer. *Int. J. Cancer*, **122**, 1324–1332.

Bolstad,B. (2004) affy: Built-in Processing Methods. Available at http://www.bioconductor.org/.

Brown,M.P.S. *et al.* (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Choe,S. *et al.* (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol.*, **6**, R16.

Cope,L.M. *et al.* (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**, 323–331.

Dettling,M. (2004) BagBoosting for tumor classification with gene expression data. *Bioinformatics*, **20**, 3583–3593.

Draghici,S. *et al.* (2006) Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.*, **22**, 101–109.

Duda,R.O. *et al*. (2001) *Pattern Classification*. John Wiley & Sons, New York, USA.

Gautier,L. *et al*. (2004) affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Irizarry,R.A. *et al*. (2003a) An R package for analyses of Affymetrix oligonucleotide arrays. In Parmigiani,G. *et al.* (eds.) *The analysis of gene expression data: methods and software*, Springer, New York, USA, pp. 102–119.

Irizarry,R.A. *et al*. (2003b) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.

Landi,M.T. *et al.* (2008) Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS ONE*, **3**, e1651.

Martens,J.H. *et al*. (2006) Differential expression of a gene signature for scavenger/lectin receptors by endothelial cells and macrophages in human lymph node sinuses, the primary sites of regional metastasis. *J. Pathol.*, **208**, 574.

Millenaar,F. *et al*. (2006) How to decide? different methods of calculating gene expression from short oligonucleotide array data will give different results. *BMC Bioinformatics*, **7**, 137.

Pearson,R.D. (2008) A comprehensive re-analysis of the Golden Spike data: towards a benchmark for differential expression methods. *BMC Bioinformatics*, **9**, 164.

Ploner,A. *et al*. (2005) Correlation test to assess low-level processing of high-density oligonucleotide microarray data. *BMC Bioinformatics*, **6**, 80.

Qin,L.X. *et al*. (2006) Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. *BMC Bioinformatics*, **7**, 23.

Ralaivola,L. *et al*. (2005) Graph kernels for chemical informatics. *Neural Netw.*, **18**, 1093–1110.

Shedden,K. *et al*. (2005) Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data. *BMC Bioinformatics*, **6**, 26.

Singh,D. *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, **1**, 203–209.

Statnikov,A. *et al*. (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, **21**, 631–643.

Su,L.J. *et al*. (2007) Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme. *BMC Genomics*, **8**, 140.

Swamidass,S.J. *et al*. (2005) Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics*, **21**(Suppl. 1), i359–368.

Trotter,M.W.B. (2006) Support vector machines for Drug Discovery. PhD Thesis, University College London, UK.

Tuna,S. and Niranjan,M. (2009) Classification with binary gene expressions. *J. biomed. sci. eng.*, **2**, 390–399.

Tuna,S. and Niranjan,M. (2010) Inference from low precision transcriptome data representation. *J. Sign. Process. syst.*, **58**, 267–279.

Wallace,T.A. *et al*. (2008) Tumor immunobiological differences in prostate cancer between African-American and European-American men. *Cancer Res*., **68**, 927–936.

West,M. *et al*. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11462–11467.

Willett,P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today*, **11**, 1046–1053.

Zhou,X. *et al*. (2003) Binarization of microarray data on the basis of a mixture model. *Mol. Cancer Ther.*, **2**, 679–684.

Zilliox,M.J. and Irizarry,R.A. (2007) A gene expression bar code for microarray data. *Nat. Methods*, **4**, 911–913.