BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 23 2006, pages 2883–2889
doi:10.1093/bioinformatics/btl339

Gene expression

Effects of replacing the unreliable cDNA microarray measurements on the disease classification based on gene expression profiles and functional modules

Dong Wang[1], Yingli Lv[1], Zheng Guo[1,2,*], Xia Li[1], Yanhui Li[1], Jing Zhu[1], Da Yang[1], Jianzhen Xu[1], Chenguang Wang[1], Shaoqi Rao[1,3,4] and Baofeng Yang[2,*]

[1]Department of Bioinformatics, [2]Department of Pharmacology and Bio-pharmaceutical Key Laboratory of Heilongjiang Province and State, Harbin Medical University, Harbin 150086, China, [3]Department of Molecular Cardiology and [4]Department of Cardiovascular Medicine, The Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, Ohio 44195, USA

Received on May 25, 2006; revised and accepted on June 16, 2006

Advance Access publication June 29, 2006

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Microarrays datasets frequently contain a large number of missing values (MVs), which need to be estimated and replaced for subsequent data mining. The focus of the paper is to study the effects of different MV treatments for cDNA microarray data on disease classification analysis.

Results: By analyzing five datasets, we demonstrate that among three kinds of classifiers evaluated in this study, support vector machine (SVM) classifiers are robust to varied MV imputation methods [e.g. replacing MVs by zero, *K* nearest-neighbor (KNN) imputation algorithm, local least square imputation and Bayesian principal component analysis], while the classification and regression tree classifiers are sensitive in terms of classification accuracy. The KNNclassifiers built on differentially expressed genes (DEGs) are robust to the varied MV treatments, but the performances of the KNN classifiers based on all measured genes can be significantly deteriorated when imputing MVs for genes with larger missing rate (MR) (e.g. MR > 5%). Generally, while replacing MVs by zero performs relatively poor, the other imputation algorithms have little difference in affecting classification performances of the SVM or KNN classifiers. We further demonstrate the power and feasibility of our recently proposed functional expression profile (FEP) approach as means to handle microarray data with MVs. The FEPs, which are derived from the functional modules that are enriched with sets of DEGs and thus can be consistently identified under varied MV treatments, achieve precise disease classification with better biological interpretation. We conclude that the choice of MV treatments should be determined in context of the later approaches used for disease classification. The suggested exclusion criterion of ignoring the genes with larger MR (e.g. >5%), while justifiable for some classifiers such as KNN classifiers, might not be considered as a general rule for all classifiers.

Contact: guoz@ems.hrbmu.edu.cn; yangbf@ems.hrbmu.edu.cn

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

*To whom correspondence should be addressed.

1 INTRODUCTION

Precise classification of disease phenotypes based on gene expression profiles has been one of the most successful applications of micrarrays (Furey *et al*., 2000; Dudoit *et al*., 2002; Zhang *et al*., 2003) and continues to be the focus of many recent studies (Asyali *et al*., 2006). However, some uncertainties remain in deciphering high-throughput expression experiments because of a large amount of data errors introduced by diverse factors such as technical failure and low signal-to-noise ratio. When these unreliable microarray measurements are discarded during the image analysis and data normalization, missing values (MVs) in up to 95% of the monitored genes can appear in the resulting dataset (de Brevern *et al*., 2004). Since many algorithms for microarray analysis require a complete data matrix as the input, MVs must be imputed before the subsequent analyses.

The MVs are usually replaced by their estimated values based on information available in the dataset. Several methods for data imputation for high-dimension microarray data have been proposed, including *K* nearest-neighbor (KNN) imputation (Troyanskaya *et al*., 2001), local least squares imputation (Kim *et al*., 2005) and Bayesian principal component analysis (Oba *et al*., 2003). It has been shown that different imputation strategies for MVs can seriously bias a subsequent analysis, e.g. significant expression analysis (Jornsten *et al*., 2005; Scheel *et al*., 2005) and gene clustering analysis (de Brevern *et al*., 2004). For disease classification analysis, many studies (Zhang *et al*., 2003; Norsett *et al*., 2004) follow the preprocessing protocol suggested by Dudoit *et al*. (2002), i.e. screening out the genes with missing data in more than a certain number of arrays, and then replacing the remaining MVs by using a data imputation approach. However, to the best of our knowledge, no study has been done to address systematically the sensitivity of the alternative classification algorithms to various data imputation methods, which can be problematic for some uncertainties in interpreting and comparing the disease classification results based on cDNA microarray data. In this study, by using five real cDNA datasets, we explore the impacts of several commonly used data imputation methods on the performances of three different classifiers [support vector machine (SVM), KNN and classification and

Table 1. Missing value occurrences in five datasets: the GMV, the OMR and the number of genes with MR in different ranges

MR	Sample	Tissue	GMV[a] (%)	OMR[b] (%)	Unigene	0%	5%	10%	20%	30%	40%
Breast cancer	59	21 ILC 38 IDC	65	14	20 849	7211	10 154	13 119	17 688	19 001	19 112
Prostate cancer	112	71 tumor 41 normal	75	13	20 815	5257	12 355	14 503	16 351	18 509	18 968
Lymphoma	20	9 FL 11 CLL	48	2	2188	1157	1520	1765	2038	2147	2179
						0%	1%	3%	5%	10%	20%
Gastric cancer[c]	132	103 tumor 29 normal	47	6	20 152	10 479	14 666	17 304	18 688	19 523	19 719
Liver cancer	156	82 tumor 74 normal	78	6	9904	2179	3246	4585	5514	7598	9784

[a]GMV: the percentage of genes with at least one MV in each dataset.
[b]The Overall Missing Rates (OMR): the percentage of MVs with respect to the whole data in each dataset.
[c]For the gastric and liver cancer data, there are >97% genes with MR $\leq$20%, respectively, we set a more detailed spectrum of MR for subsequent analysis.

regression trees (CART)]. The classifiers are built on either all genes measured in the microarrays or the feature genes selected as 'differentially expressed genes (DEGs)' (Tusher *et al*., 2001). Selecting a subset of feature genes, including the identification of the DEGs characterizing the varied expression patterns between disease states, can often enhance the classification performance on high dimensional microarray data (Bo and Jonassen, 2002; Li *et al*., 2005). However, a major problem for using the feature genes is that some factors such as the thresholds of expression significance (Pan *et al*., 2005) and the MV imputation methods (Jornsten *et al*., 2005; Scheel *et al*., 2005) can lead to very different DEGs (or feature genes) from the same experiment.

Despite the uncertainty in selecting feature genes, it has been shown (Hosack *et al*., 2003) that using Gene Ontology (GO) (Ashburner *et al*., 2000), the functional modules enriched with DEGs identified by different methods are relatively robust. For human disease analysis, although thousands of genes can be measured simultaneously in microarray experiments, many important disease relevant genes may actually be absent (or missing) on the microarrays. In such cases, the underlying mechanism of the disease(s) under study is more probably being captured within the functional context of the biologically related genes measured on the microarrays. Therefore, it is of great interest, especially in the scenarios of microarray experiments with a large number of missing genes, that we can still classify disease samples based on the robust functional modules with the capability of explicitly introducing biological knowledge into data analysis and greatly reducing the high dimensional microarray data. In our previous study (Guo *et al*., 2005), we developed such a modular approach based on some summary measures of the expression data of the DEGs contained in the functional modules. In this study, we will demonstrate that the modular approach can achieve precise disease classification with better biological interpretations based on the functional modules that are very robust to different MV treatments.

2 METHODS

2.1 cDNA microarray data and preprocessing

In this paper, we focus on two-class classification. In each of the five publicly available cDNA microarray datasets, two disease subtypes with the largest sample sizes are thus selected. The breast cancer dataset contains 20 849 genes (UniGene clusters) measured on 21 invasive lobular carcinoma (ILC) and 38 invasive ductal carcinoma (IDC) samples (Zhao *et al*., 2004). The prostate cancer dataset contains 20 815 genes measured on 71 prostate tumors and 41 normal prostate specimens (Lapointe *et al*., 2004). The lymphoma dataset contains 2188 genes measured on 9 follicular lymphoma (FL) and 11 chronic lymphocytic leukemia (CLL) samples (Alizadeh *et al*., 2000). The gastric cancer dataset contains 20 152 genes measured on 103 gastric tumors and 29 normal gastric samples (Chen *et al*., 2003). The liver cancer dataset contains 9904 genes measured on 82 liver tumors and 74 normal liver samples (Chen *et al*., 2004).

The characteristics for the five datasets are shown in Table 1. The Genes with MV (GMV) column, recording the percentage of genes with at least one MV, is 75, 65, 48 and 47%, and 78% for the 5 datasets respectively. The Overall Missing Rates (OMR) column, denoting the percentage of MVs in the whole data, is 14, 13, 2, 6 and 6% for the five datasets respectively. We define a gene's missing rate (MR) as the percentage of its missing data points in all samples in a dataset. The distributions of the genes with different amount of MVs are also shown (Table 1). When the MR threshold is set at 5%, there are 49, 59, 69, 93 and 56% genes remained (i.e. with MR $\leq$ 5%) in the breast, prostate, lymphoma, gastric and liver cancer datasets respectively, and the rate increases to 70, 63, 81, 97 and 77% when the MR threshold is 10%. Henceforth, we can retain most information for lymphoma and gastric cancer datasets if we set MR threshold at 5 or 10%, while 37 and 30% genes are lost for breast and prostate datasets when the MR threshold is set at 10%. Prior to data imputation, we perform two preprocessing procedures: (1) carrying out base two logarithmic transformations; and (2) using median normalization to subtract the median from each gene so that the observations have median 0.

2.2 Imputation methods

Before classifying samples between biological subtypes based on microarray data, MVs have to be imputed on the data matrix, say G with the element g_{ij} denoting the expression level of the i-th gene in the j-th sample. In this study, we investigate four different methods for data imputation. Replacing MVs by zero (Alizadeh *et al*., 2000) is the simplest approach to dealing with MVs, and we refer to it as Zimpute. The most frequently used Nearest-Neighbor imputation algorithm (KNNimpute) for data imputation (Troyanskaya *et al*., 2001) estimates a MV of the gene i in sample j by the weighted average of expression values in sample j of the k closest genes, based on the Euclidean distance measure for estimating the similarity of neighboring genes. For the weighted average, the contribution of each gene is weighted by its inverse distance to gene i. As the method estimates MVs well within the range

of 10–20 neighbors and is relatively insensitive to the exact value of k (Troyanskaya *et al*., 2001), we set $k = 15$ in this study. Local Least Square imputation (LLSimpute) (Kim *et al*., 2005) is a regression-based method to estimate the MVs of a gene using its most similarly co-expressed k genes, based on the absolute measure of Pearson correlation coefficient. The Bayesian Principal Component Analysis (BPCA) (Oba *et al*., 2003) estimates the MVs within a Bayes inference framework consisting of three components, which are principal component regression, Bayesian estimation and iterations based on expectation maximization. To impute the MVs, BPCA assumes a global covariance structure of the gene expression data, while KNNimpute and LLSimpute use local gene co-expression structure. Details for these imputation techniques were described already in the original papers.

2.3 Differentially expressed genes and functional modules

We further investigate the power and feasibility of using the concepts of DEGs and functional modules for analyzing microarray data with MVs, in particular, the effects of the MV imputation on the classifiers trained on the DEGs and the functional modules. We select DEGs using SAM (Tusher *et al*., 2001) with false discovery rate ≤ 0.1.

Functionally related genes tend to express and perform their highly integrated roles in modular fashions (Hartwell *et al*., 1999), often reflected by a high degree of concert of the gene reactions to stimuli such as disease conditions (Segal *et al*., 2004). Based on the most widely used gene functional annotation system GO (Ashburner *et al*., 2000), we apply a hypergeometric distribution (Draghici *et al*., 2003) to calculate the probability p of a GO 'biological process' category having the number of the annotated DEGs by random chance. When the p-values are used as statistical significance levels for selecting modules, generally, correction for multiple tests should be addressed (Osier *et al*., 2004). However, it has been suggested that in disease subtypes, many genes disturbed by disease condition(s) change systematically (Alizadeh *et al*., 2000), indicating that the modular expressions of the genes in the biological pathways appear in a relatively active fashion rather than in a clearly separated way. Therefore, we use the p-value as a heuristic measure for roughly ranking the relative enrichment of DEGs in the GO classes, which directs the analysis to the functionally most active classes. We select the GO categories enriched with DEGs, with p-value ≤ 0.05, and refer to them as candidate 'differentially expressed functional modules' or 'modules' for short. We restrict to analysis of those modules annotated with at least five genes. When two modules are of a general-specific relationship, only the module with more specific description is retained. Because a module is determined by the joint statistical behaviors of a set of genes, it is insensitive to a few outliers and as such robust to MV treatments.

According to the SAM statistics, we classify the DEGs in a module as up- and down-regulated across the disease subtypes. For a disease sample, we use a summary measure (the arithmetic mean or the median) of the gene expression values in a module, separately for the up- and down-regulated DEGs, to reflect the modular expression of the genes in the sample. The modular summary measures for all samples in a dataset produce the functional expression profiles (FEPs) of the disease subtypes (Guo *et al*., 2005). We classify disease subtypes based on the FEPs, which consider the gene expressions within a functional module as an integrated data point and thus reduce the high dimensional gene expression profiles of thousands of genes to a small number of modules that are robust to MV treatments. These biological modules can also facilitate the interpretation of the classification results at the modular level.

2.4 Classifiers

The three classification algorithms, as described by Dudoit *et al*. (2002), are evaluated in this study: hard-margin SVM with a first-degree dot product kernel function, KNN classifier with $k = 11$ and CART. A classifier is evaluated using leave-one-out cross-validation (LOOCV), in which each sample in the training set is left out in turn, and the accuracy rate of the classifier is computed as the percentage of the number of times that the classifier is correct in its predictions. The LOOCV procedure provides an unbiased estimate of the true accuracy rate of the classification procedure and has been widely used to evaluate classifiers (Furey *et al*., 2000; Zhang *et al*., 2003).

If the phenotypic information of the test samples is used in training the classifier, although it has been suggested that the accuracy rate estimate may still be proper for the purpose of ranking true accuracy rate of classifiers (Braga-Neto *et al*., 2004), the estimate may be biased because of some resubstitution effects. To attempt to provide an unbiased estimate of the accuracy rate of the classifier, we perform selection of DEGs and GO categories 'within the cross-validation loop' for each leave-one-out training set (Simon *et al*., 2003). As the gene expression values of the left out test sample may influence the imputation result, it might be proper to perform the very time-consuming MVs imputation for samples in each of the cycles in LOOCV. On the other hand, most imputation methods estimate MVs in some unsupervised ways by exploiting the co-expression relationships among genes across different samples, which are akin to the gene cluster analysis. In such unsupervised imputation methods, samples are treated as data points and their class labels are ignored, which is quite different from the supervised analysis such as the DEG selection targeted at the best separation between phenotypic classes of samples. By considering the unsupervised imputation procedure as an independent data preprocessing step, we thus estimate MVs only once for a set of experiments, as performed in most of similar data analyses (Furey *et al*., 2000; Dudoit *et al*., 2002; Zhang *et al*., 2003).

3 RESULTS

3.1 Classification based on gene expression profiles

In breast, prostate and lymphoma datasets, given the MR threshold at $5n\%$ ($n = 0, 1, 2, 4, 6, 8$), the MVs of the genes with MR smaller than the threshold are replaced by zero (Zimpute) or their estimated values derived from KNNimpute, LLSimpute and BPCA respectively, while the data of the genes with higher MR are ignored. For the gastric cancer data, there are 93% genes with MR $\leq$ 5%, and for the liver cancer data, there are 99% genes with MR $\leq$ 20%. Therefore, we set a more detailed spectrum of MR (1, 3, 5, 10 and 20%) for these two datasets for subsequent analysis.

The accuracies for the SVM, KNN and CART classifiers trained on breast and prostate cancer data with different MV treatments are shown in Figure 1. The results for three other datasets are provided in a Supplementary file (Supplementary Data 1). The SVM classifiers perform well and are very robust to varied MV treatments. Compared with SVM classifiers, the performances of KNN and CART classifiers are significantly affected by MV treatments. In the breast dataset, when no MV is allowed, the accuracy of SVM reaches its highest value (90%), and drops slightly when varying amount of MVs are replaced by different MV imputation methods (Fig. 1a). However, in the other four datasets, the SVM classifiers are not affected by data imputation and methods (see Fig. 1d and Supplementary Data 1). The accuracy of KNN classifier reaches its highest value (88%) for the liver cancer when the MR threshold is set at a small value (5 or 10%), or when no MV is allowed (MR = 0) for the other four datasets (Fig. 1b for breast cancer and Fig. 1e for prostate cancer, and the Supplementary Data for other cancers). For genes with larger MR (e.g. MR > 5%), using their estimated MVs actually deteriorates the performances of the KNN classifiers. As demonstrated in Figure 1c and f, CART classifiers are very sensitive to varied MV treatments.

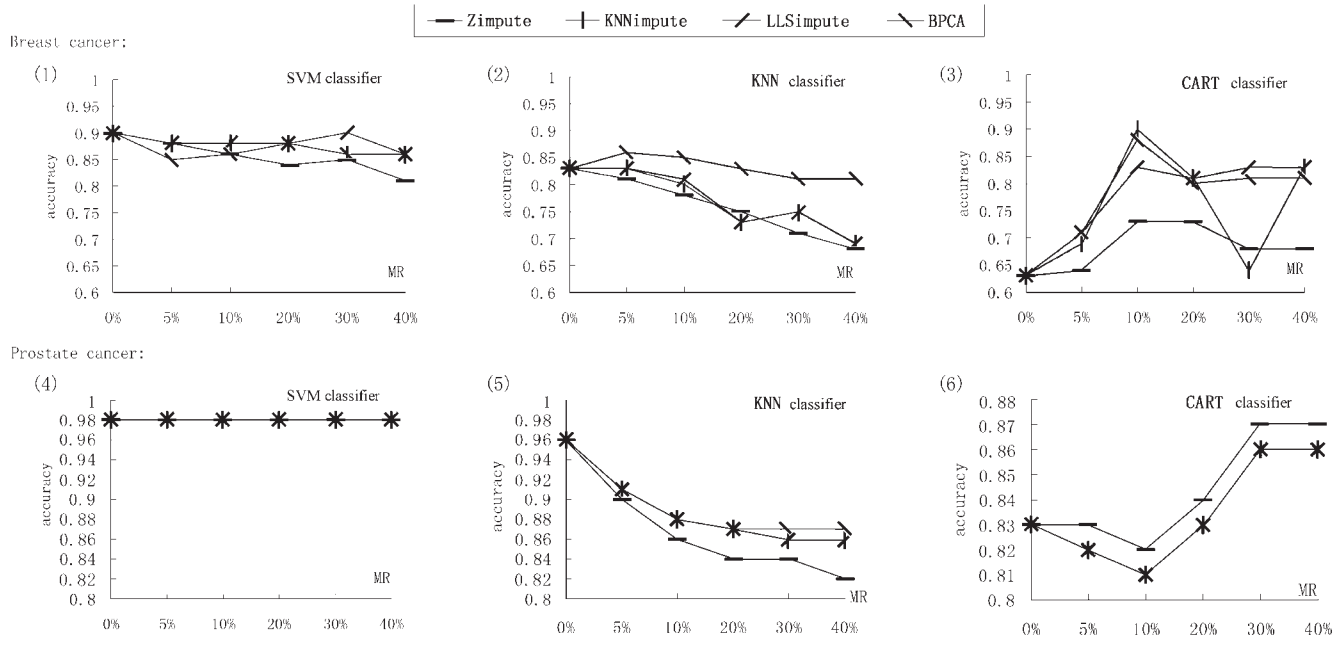

Fig. 1. Accuracy rates of three different classifiers, with varied MV treatments in breast and prostate cancer datasets.

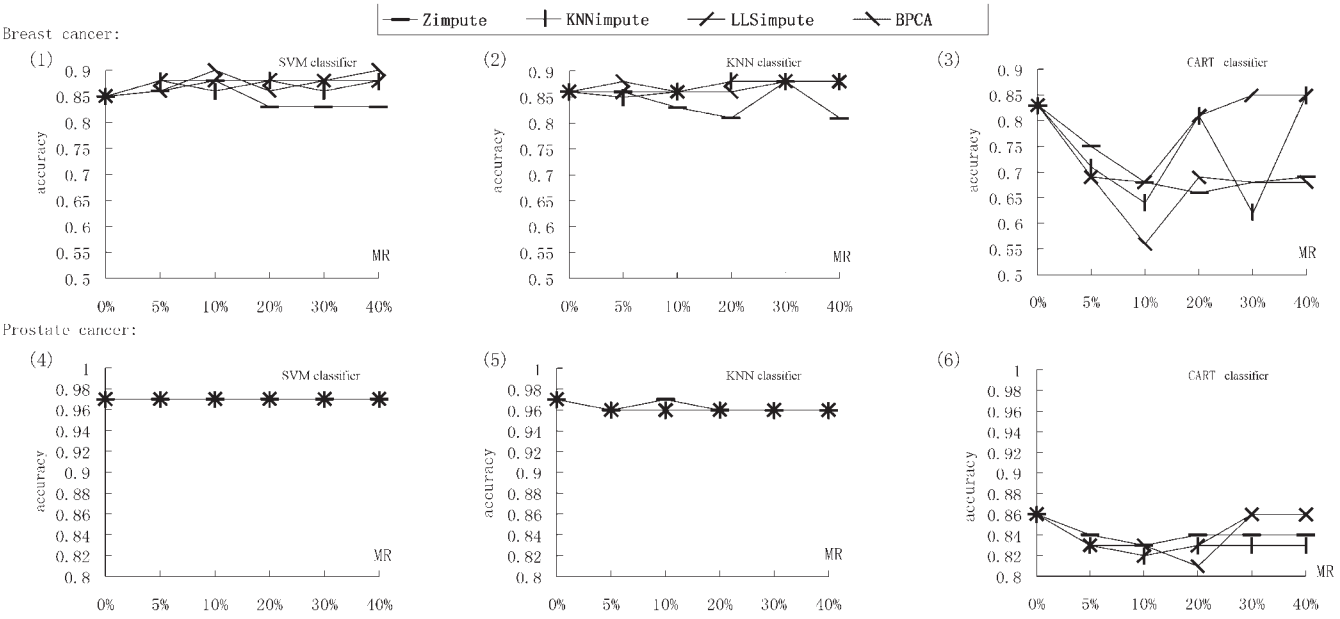

Fig. 2. Accuracy rates of DEG-based classifiers, with varied MV treatments in breast and prostate cancer datasets.

Compared with KNNimpute, LLSimpute and BPCA, when imputing the MVs of genes with larger MR (e.g. MR >5%), Zimpute leads to either a slight decline in terms of accuracy for SVM classifiers (e.g. in the breast cancer dataset), or more obvious drops for the KNN classifiers in four datasets (Fig. 1 and Supplementary Data). The slight better performances of Zimpute for CART classifiers in the prostate cancer and lymphoma datasets (Fig. 1 and Supplementary Data) may possibly owe to the unstable behaviors of the CART algorithm for analyzing high dimensional noisy data. Although BPCA performs better for the KNN classifiers in the breast cancer dataset, in general, the KNNimpute, LLSimpute and BPCA imputation algorithms make little difference in affecting classification performances. It is also worthy to note that in a wide range of MR threshold values the most commonly used KNNimpute method are comparable with other more complicated imputation methods in context of disease classification analysis.

We also evaluate the effects of replacing MVs on the accuracies of the classifiers built on the DEGs. As shown in Figure 2 and

Table 2. Accuracy rates of FEP-based classifiers using KNNimpute to deal with missing data

MR		Breast cancer			Prostate cancer			Lymphoma			Gastric cancer			Liver cancer		
		0%	5%	10%	0%	5%	10%	0%	5%	10%	0%	5%	10%	0%	5%	10%
Mean	SVM	0.73	0.81	0.76	0.92	0.92	0.94	1	1	1	0.92	0.98	0.99	0.93	0.96	0.97
	KNN	0.80	0.81	0.83	0.95	0.92	0.98	1	1	1	0.92	0.96	0.97	0.98	0.96	0.95
	CART	0.71	0.80	0.81	0.94	0.92	0.88	1	0.95	0.95	0.86	0.92	0.91	0.92	0.90	0.90
Median	SVM	0.69	0.83	0.78	0.88	0.89	0.93	1	1	1	0.94	0.97	0.98	0.94	0.91	0.95
	KNN	0.80	0.81	0.78	0.91	0.92	0.96	1	1	1	0.90	0.93	0.95	0.93	0.93	0.93
	CART	0.62	0.76	0.76	0.80	0.86	0.81	0.95	0.90	0.85	0.86	0.93	0.87	0.88	0.85	0.87

the Supplementary Material (Supplementary Data 2), similar to the results when no DEGs are selected, the SVM classifiers are very robust to varied MV imputing methods, while the CART classifiers are sensitive and perform relatively poor compared with the SVM and KNN classifiers. However, even for genes with high percentages of MVs, data imputation does not lead to marked decline in the performance of the KNN classifiers, which is quite different from the situation when no DEGs are selected. Again, in general, the results show that data imputation does not enhance much the accuracy rates of the studied classifiers, even for the breast and prostate datasets that contain high percentages of MVs. As shown in Figure 2 and the Supplementary Material (Supplementary Data 2), KNNimpute, LLSimpute and BPCA imputation algorithms make little difference in affecting classification performances of the SVM and KNN classifiers, while the Zimpute performs the worst.

3.2 Classification based on functional expression profiles

We further evaluate the effects of replacing MVs by KNNimpute method on the accuracy of the classifiers built on two FEP summary measures (arithmetic mean and median). As described in Table 2, for MV treatments for genes with different MRs in five datasets, the CART classifiers perform relatively poor compared with the SVM and KNN classifiers. As demonstrated in Figure 3 and the Supplementary Material (see Supplementary Data 3), in general, performing data imputation does not affect much the performances of the FEP-based SVM and KNN classifiers, for both arithmetic mean and median summary measures, which provide robust and precise classification results comparable with the classifiers based on DEGs. However, for breast dataset, the FEP classifiers perform relatively poor compared with the classifiers based on DEGs. The main reason might be that, for this dataset, among the DEGs selected at MR threshold 0, 5 and 10% respectively, there are only ~300 genes that can be annotated to GO for constructing the FEP, leading to too much information loss. This result indicates that the power of FEP approach is limited by the availability of gene functional knowledge.

In each dataset, although the MR thresholds for treating MVs seriously influence the output of the DEG lists, the most significant modules identified are relatively consistent. For example, using all samples in the prostate cancer dataset, there are 1675 DEGs selected at MR threshold 0% and the number of DEGs increases greatly to 2758 at MR threshold 5%. However, for the 15 modules identified at MR threshold 0, 4 are the same and 3 are of parent–child

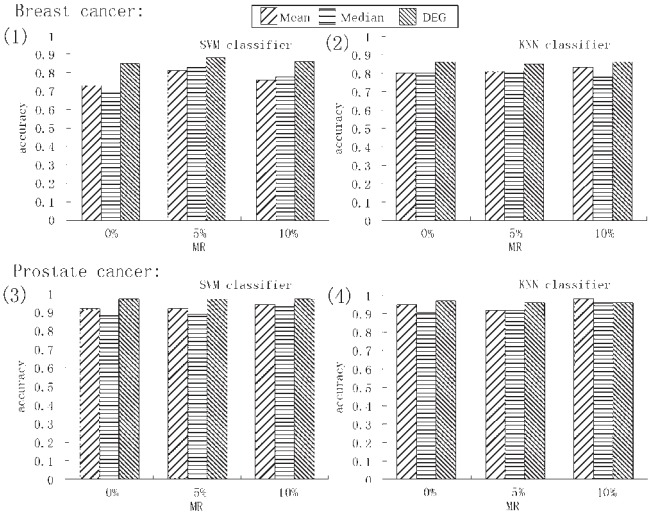

Fig. 3. Accuracy rates of FEP-based classifiers using KNNimpute to deal with missing data in breast and prostate cancer datasets.

relationships with the modules identified at MR threshold 5%. We take two modules identified in the prostate dataset, as examples, to explain the disease relevance of the modules. In the module 'muscle development', IGF1 (insulin-like growth factor 1) is relevant to the degradation of p53 and its expression level can be used to predict the risk of prostate cancer (Chen *et al*., 2005), and fibroblast growth factors (FGF-1, 2, 6, 8, FGF-2) can induce enhanced proliferation and metastasis and are relevant to prostate cancer (Kwabi-Addo *et al*., 2004). For the module 'cell–cell adhesion', it has been suggested that the expression of cell adhesion proteins is of prognostic value for prostate cancer (Kallakury *et al*., 2001). In this module, CD44 is a glycosylated adhesion molecule mediating prostate cell adhesion, and the interactions between CD44 isoforms and cytoskeletal proteins may play a pivotal role in regulating tumor cells during prostate cancer development (Welsh *et al*., 1995).

4 DISCUSSION AND CONCLUSION

Proper data imputation should be determined in light of the overall objective of a study. If the next process is a classification analysis based on some algorithms sensitive to the imputed values, such as the KNN classifier, we recommend estimating and replacing MVs for genes with only a small percentage of MVs (e.g. MR $\leq$5%) and

discarding genes with larger MR, similar to the preprocess protocol in Dudoit *et al*. (2002). However, based on the results of this study, any single exclusion criterion for discarding genes with a certain MR seems arbitrary as a general rule. For the SVM classifiers, there are few differences in accuracy rates at different MR thresholds, reflecting the robust characteristics of the SVM classifiers for treating noisy data. Another possible reason is that the five datasets are non-informative as to the optimal MR cutoff for SVM classifiers. Most imputation methods exploit the co-expression relationships among genes across different samples for estimating MVs. The existence of other genes with similar (high correlation) prediction power can render data imputation less useful. When many MVs are imputed for genes with high MR (e.g. MR >5%), the contribution of noise or uncertainty to disease classification may overwhelm the contribution of signal or additional information, leading to a decrease in classification accuracy for the resulting classifier(s). If we have to keep more genes for other types of analysis such as finding DEGs, it may be necessary to deal with genes having higher MRs (Jornsten *et al*., 2005; Scheel *et al*., 2005). However, one should take precautions when drawing critical biological conclusions from data that are partially estimated.

Among the three kinds of classifiers evaluated in this study, based on either DEGs or all genes, SVM classifiers are robust to varied MV treatments, while the CART classifiers are the unstable ones. The KNN classifiers lie in between, and are rather robust to varied MV treatments, when built on DEGs, but when built on all measured genes, including estimated MVs for genes with larger MR (e.g. MR >5%), they can significantly deteriorate the performances of the KNN classifiers. Nevertheless, based on this study, we cannot conclude that classification based on DEGs is consistently improved over classification based on all the genes, since using DEGs or all genes, the accuracy rates for SVM classifiers are very similar. In contrast, it is evident that generally, the KNNimpute, LLSimpute and BPCA imputation algorithms have little difference in affecting classification performances of the SVM and KNN classifiers, while the Zimpute performs the worst.

In general, doing MV imputation does not improve much the disease classification accuracy, indicating that a sufficient fraction of genes with relatively reliable measurements can hold enough information for disease classification, which might be explained by the systematic and modular characteristics of gene expressions in cancers (Segal *et al*., 2004). For the same reason, the SVM and KNN classifiers based on functional modules can provide precise disease classification results comparable with the classifiers based on DEGs. In general, data imputation does not affect much the performances of the FEP-based SVM and KNN classifiers.

The modules, used as features in the FEP classifiers, are themselves relatively robust to varied MVs treatments and other uncertainties existed in data-preprocess procedures (Hosack *et al*., 2003), making the FEP approach more biologically revealing for disease classification. Furthermore, because it uses summary measures of the gene expressions in the modules, the FEP approach suggests a logical way to integrate cross-platform data, i.e. for data integration with MVs (Warnat *et al*., 2005). Finally, although robust in nature, the modules identified vary to some extent with different MV treatments and DEGs selection thresholds (Pan *et al*., 2005). Therefore, more biologically and statistically sounding strategies for identifying DEGs and modules (Bickel, 2004; Breitling *et al*., 2004) deserve further explorations for more powerful module-based analysis of the microarray data with MVs.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 30170515, 30370388, 30370798, 3057042 and 30571034), the Chinese 863 Program (Grant Nos. 2003AA2Z2051 and 2002AA2Z2052), the Heilongjiang Province Department of Education Outstanding Overseas Scientist grant (Grant No. 1055HG009) and US National Institute of Health SCCOR grant (Grant No. P50 HL077101-01).

Conflict of Interest: none declared.

REFERENCES

Alizadeh,A.A. *et al*. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.

Asyali,M.H. *et al*. (2006) Gene expression profile classification: a review. *Current Bioinform*., **1**, 55–73.

Bickel,D.R. (2004) Degrees of differential gene expression: detecting biologically significant expression differences and estimating their magnitudes. *Bioinformatics*, **20**, 682–688.

Bo,T. and Jonassen,I. (2002) New feature subset selection procedures for classification of expression profiles. *Genome Biol*., **3**, RESEARCH0017.

Braga-Neto,U. *et al*. (2004) Is cross-validation better than resubstitution for ranking genes? *Bioinformatics*, **20**, 253–258.

Breitling,R. *et al*. (2004) Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*, **5**, 34.

Chen,C. *et al*. (2005) Prostate carcinoma incidence in relation to prediagnostic circulating levels of insulin-like growth factor I, insulin-like growth factor binding protein 3, and insulin. *Cancer*, **103**, 76–84.

Chen,X. *et al*. (2003) Variation in gene expression patterns in human gastric cancers. *Mol. Biol. Cell*., **14**, 3208–3215.

Chen,X. *et al*. (2004) Novel endothelial cell markers in hepatocellular carcinoma. *Mod. Pathol*., **17**, 1198–1210.

de Brevern,A.G. *et al*. (2004) Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinformatics*, **5**, 114.

Draghici,S. *et al*. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Dudoit,S. *et al*. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc*., **97**, 77–87.

Furey,T.S. *et al*. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Guo,Z. *et al*. (2005) Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, **6**, 58.

Hartwell,L.H. *et al*. (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.

Hosack,D.A. *et al*. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol*., **4**, R70.

Jornsten,R. *et al*. (2005) DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics*, **21**, 4155–4161.

Kallakury,B.V. *et al*. (2001) Co-downregulation of cell adhesion proteins alpha- and beta-catenins, p120CTN, E-cadherin, and CD44 in prostatic adenocarcinomas. *Hum. Pathol*., **32**, 849–855.

Kim,H. *et al*. (2005) Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, **21**, 187–198.

Kwabi-Addo,B. *et al*. (2004) The role of fibroblast growth factors and their receptors in prostate cancer. *Endocr. Relat. Cancer*, **11**, 709–724.

Lapointe,J. *et al*. (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl Acad. Sci. USA*, **101**, 811–816.

Li,L. *et al*. (2005) A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. *Genomics*, **85**, 16–23.

Norsett,K.G. *et al*. (2004) Gene expression based classification of gastric carcinoma. *Cancer Lett*., **210**, 227–237.

Oba,S. *et al*. (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**, 2088–2096.

Osier,M.V. *et al*. (2004) Handling multiple testing while interpreting microarrays with the Gene Ontology Database. *BMC Bioinformatics*, **5**, 124.

Pan,K.H. *et al*. (2005) Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proc. Natl Acad. Sci. USA*, **102**, 8961–8965.

Scheel,I. *et al*. (2005) The influence of missing value imputation on detection of differentially expressed genes from microarray data. *Bioinformatics*, **21**, 4272–4279.

Segal,E. *et al*. (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet*., **36**, 1090–1098.

Simon,R. *et al*. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl Cancer Inst*., **95**, 14–18.

Troyanskaya,O. *et al*. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Warnat,P. *et al*. (2005) Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, **6**, 265.

Welsh,C.F. *et al*. (1995) Interaction of CD44 variant isoforms with hyaluronic acid and the cytoskeleton in human prostate cancer cells. *J. Cell. Physiol*., **164**, 605–612.

Zhang,H. *et al*. (2003) Cell and tumor classification using gene expression data: construction of forests. *Proc. Natl Acad. Sci. USA*, **100**, 4168–4172.

Zhao,H. *et al*. (2004) Different gene expression patterns in invasive lobular and ductal carcinomas of the breast. *Mol. Biol. Cell*, **15**, 2523–2536.