Available online at www.sciencedirect.com
ScienceDirect
Journal of Biomedical Informatics 41 (2008) 570–579
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Cross-generation and cross-laboratory predictions of Affymetrix microarrays by rank-based methods

Hsi-Che Liu [a,f,g], Chien-Yu Chen [b,*], Yu-Ting Liu [c], Cheng-Bang Chu [c], Der-Cherng Liang [a], Lee-Yung Shih [d], Chih-Jen Lin [e]

[a] Department of Pediatrics, Mackay Memorial Hospital, Taipei, Taiwan
[b] Department of Bio-industrial Mechatronics Engineering, National Taiwan University, No. 1, Roosevelt Rd., Sec. 4, Taipei 106, Taiwan
[c] Graduate School of Biotechnology and Bioinformatics, Yuan Ze University, Chung-Li, Taiwan
[d] Division of Hematology-Oncology, Chang Gung University, Taoyuan, Taiwan
[e] Department of Computer Science, National Taiwan University, Taipei, Taiwan
[f] Mackay Medicine, Nursing, and Management College, Taipei, Taiwan
[g] School of Medicine, Taipei Medical University, Taipei, Taiwan

Received 3 May 2007
Available online 4 December 2007

Abstract

Past experiments of the popular Affymetrix (Affy) microarrays have accumulated a huge amount of public data sets. To apply them for more wide studies, the comparability across generations and experimental environments is an important research topic. This paper particularly investigates the issue of cross-generation/laboratory predictions. That is, whether models built upon data of one generation (laboratory) can differentiate data of another. We consider eight public sets of three cancers. They are from different laboratories and are across various generations of Affy human microarrays. Each cancer has certain subtypes, and we investigate if a model trained from one set correctly differentiates another. We propose a simple rank-based approach to make data from different sources more comparable. Results show that it leads to higher prediction accuracy than using expression values. We further investigate normalization issues in preparing training/testing data. In addition, we discuss some pitfalls in evaluating cross-generation/laboratory predictions. To use data from various sources one must be cautious on some important but easily neglected steps.

Keywords: Affymetrix microarrays; Cross-generation/laboratory prediction; Rank-based normalization

1. Background

Gene expression profiling by DNA microarrays is a useful tool in biological and clinical research. Superior to traditional biological experiments, it compares thousands of genes simultaneously. With fast and systematic analysis of expression values, one can quickly identify significant genes for certain diseases or build models for patient diagnosis/tumor classification.

Though the microarray technology is popular, not many institutions can conduct enough experiments for effective analysis due to the lack of patient samples or the high cost. Studies in recent years have accumulated a huge amount of microarray samples in public databases. If data experimented under similar conditions can be combined together, not only any laboratory can directly apply microarray technology in practical use, but also more extensive and reliable studies are possible. The emergence of common guidelines MIAME (Minimum Information About a Microarray

* Corresponding author. Fax: +886 223627620.
E-mail addresses: hsiche@ms1.mmh.org.tw (H.-C. Liu), cychen@mars.csie.ntu.edu.tw (C.-Y. Chen), s938611@mail.yzu.edu.tw (Y.-T. Liu), s938613@mail.yzu.edu.tw (C.-B. Chu), dcliang@ms2.mmh.org.tw (D.-C. Liang), sly7012@adm.cgmh.org.tw (L.-Y. Shih), cjlin@csie.ntu.edu.tw (C.-J. Lin).

1532-0464/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2007.11.005

Experiment) [3] adapted by leading scientific journals indicates the direction toward the universal use of public data. The comparability of microarray experiments between diversed sources and across distinct technologies is thus an important research issue.

A microarray experiment from raw samples to expression values is a complicated procedure. Expression values from various sources are not easily comparable. Many recent studies explore the cross-platform comparability between cDNA and oligonucleotide arrays, but so far contradictive results have been reported. Even using the same samples, some papers (e.g., [4,5]) conclude that measurements from the two platforms are poorly correlated. Though recent studies (e.g., [6–8]) give more promising results, they still consider that the reproducibility across platforms is not easily available.

For the same type of arrays, comparability issues also occur. In particular, whether results from various generations of popular Affymetrix (Affy) human oligonucleotide arrays can be used together is an issue. Though the same platform tends to produce more consistent expression values, cross-generation and cross-laboratory use of Affy arrays remains a challenging task. This paper intends to have a detailed investigation on this subject. Existing papers of this topic mainly study the following three issues:

1. Whether differentially expressed genes identified across two generations (laboratories) are similar or related.
2. Whether the same samples lead to similar expression values across two generations (laboratories).
3. Whether models built upon data of one generation (laboratory) can differentiate data of another.

This paper focuses on studying the third issue.

Most work studying issue 1 concludes that genes identified across generations (laboratories) are related (e.g., [9]). In contrast, the other two issues are less settled. For issue 2, one of the first studies is [10]. Using the same samples on two generations, it reports that better similarity of the probe sets leads to higher correlation between expression values. References [11,12] further strengthen this finding by showing that considering only probes with overlapping sequences gives excellent comparability. However, even with these studies, some still doubt the reproducibility of expression values across generations, so several papers propose more sophisticated techniques. By calculating expression changes at the probe-level, Elo et al. [13] report that such information gives better comparability. Bhattacharya and Mariani [14] propose regression models, which reflect the relationship between expression values of two generations.

Compared to issue 2, issue 3 concerns more about the practical use of data from different sources. Many applications such as cancer diagnosis and tumor classification are of this type. If data of other laboratories can be used, an institution can classify its patient samples without huge initial experiments/costs. Some have checked issue 3: Bloom et al. [15] collect samples of 21 tumor types across different laboratories and two Affy generations. They normalize expression values of various sources and apply an Artificial Neural Networks (ANN) model. High prediction accuracy (88%) is reported. Jiang et al. [16] consider lung cancer data sets across two generations. They develop special data transformation and report high prediction accuracy. Xu et al. [17] study prostate cancer samples across different laboratories but under the same Affy generation. Using a classifier based on only two genes, a model built on data from three laboratories successfully separates a test set from another laboratory to normal or cancer. Some other related papers are references [1,2].

This paper makes the following two contributions regarding cross-laboratory and cross-generation predictions:

1. We investigate whether expression values are reliable for the prediction tasks. In cross-platform analysis (e.g., cDNA and oligonucleotide), quite a few (e.g., [18,19]) observe inconsistent expression values, so they use information less dependent on the scale of values (e.g., rank levels). While expression values seem to be more consistent if only Affy arrays are considered, it is essential to check which way is better. We propose a rank-based approach and compare it with using expression values.
2. We present a correct way of evaluating cross-laboratory and cross-generation predictions. Some earlier papers (e.g., [15,19,20]) mix data across various generations (or platforms), and then split the set to training/testing. We point out that such a mechanism sometimes overestimates the accuracy. A correct evaluation should have the training set independent of the generation (platform) of the test data.

This paper does not touch cross-platform issues (e.g., cDNA vs. Affy). For a more complete account of cross-platform predictions, see [19] and references therein.

2. Methods

Fig. 1 outlines our approach. Data sets of three cancer types, made of Affy human oligonucleotide microarrays, are from different generations and/or laboratories. The expression values are downloaded from public web sites. In each experiment two sets of the same cancer type are used as training/prediction sets. To take into account the variability of expression values, all samples are normalized so that the mean of expression values over all genes is a specified constant. We identify common genes among different chip generations. Then an optional step is to transform expression values to rank levels. For the training set we apply SAM (Significance Analysis of Microarrays) [21] to obtain differentially significant genes. The classification analysis is performed by applying k-Nearest Neighbors (KNN) [22]. A leave-one-out cross-validation (LOO CV) procedure obtains a suitable parameter k. The entire process is repeated for any pair of sets as training/testing.

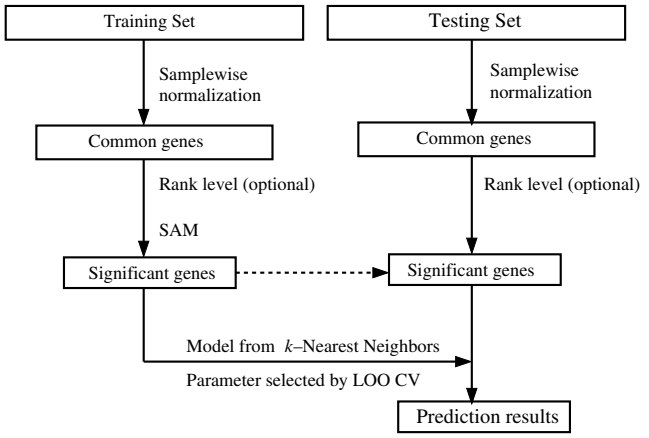

Fig. 1. Workflow of the analysis. The width of each box reflects the number of genes. Details are in the beginning of Section 2. Note that in this paper we compare whether using expression values or transforming them to rank levels is better. Thus "Rank level" is marked as "optional" as it is applied only to the latter.

All experiments are conducted by software packages included in the R-project [23]. Details of our methods are described in the following subsections.

2.1. Microarray data collection and preprocessing

To conduct the integrated, cross-generation and cross-laboratory predictions of Affy human arrays, we select public data of three cancer types. The first type, acute lymphoblastic leukemia (ALL), includes two data sets of different generations published from the same laboratory. The data set of Yeoh et al. [24], from HG-U95Av2 array, is denoted as ALL-95. Another data set of Ross et al. [25], from HG-U133A array, is called ALL-133. It has 132 samples from the original 335 HG-U95Av2 data obtained by Yeoh et al. To generate an independent source, ALL-95 includes only 203 non-replicated cases. The subtypes to be predicted are ALLs with defined recurrent chromosomal aberrations: $t(12;21)$, $t(1;19)$ and hyperdiploid with more than 50 chromosomes (HD $>$ 50). The second cancer type is acute myeloid leukemia (AML). Three data sets are generated by three different institutions using HG-U133A chips: The set Ross et al. [26], abbreviated as AML-1, is a childhood study. The other two studies, AML-2 (Valk et al. [27]) and AML-3 (Gutiérrez et al. [28]), involve adult samples. The predicted subtypes are AMLs with $t(8;21)$, inv(16) and $t(15;17)$. These biologically distinct subtypes are identical both in pediatric and adult AMLs. The last group is breast cancer. Three data sets across three generations of chips (HuGeneFL, HG-U95Av2 and HG-U133A) are collected from three different institutions. They are denoted as Breast-FL (West et al. [29]), Breast-95 (Huang et al. [30]) and Breast-133 (Wang et al. [31]), according to the generation of chip used in each individual study. The estrogen receptor (ER) status (positive or negative) is the variable to be predicted. Table 1 summarizes key characteristics and URL addresses of all the eight microarray data sets.

After expression values are downloaded from the referred public websites, values of each array are rescaled by setting the 2% trimmed mean of all the genes in an array to be 500, as suggested in the Affy Microarray Suite 5.0 (MAS 5.0) program.

2.2. Gene mapping (common probe sets identification)

According to their launch time, the three array generations can be aligned as the order of HuGeneFL, HG-U95Av2, and HG-U133A. Because of multiple design advances used to produce newer Affy human arrays, many probe sets differ between generations of arrays. For a comparative analysis, it is critical to identify a subset of common genes. One approach is to match the UniGene IDs among genes. Each UniGene ID corresponds to a cluster containing sequences that represent a unique gene and its related information [32]. However, different UniGene Builds are used for the three Affy generations. Some UniGene IDs cannot be exactly tracked between two Builds. An alternative method considers LocusLink (currently implemented as Entrez Gene [33]), and it does not suffer from the same problem as much. Another popular method uses matching tables provided by Affymetrix. The matching between two generations of arrays is based on the similarity of sequence information of probe sets (http://www.affymetrix.com/support/technical/comparison_spreadsheets.affx). According to different constructions, there are two mappings called "Good Match" and "Best Match". The latter, obtained under a more stringent criterion than the former, chooses probe sets with the greatest likelihood of representing the same gene across generations [34].

Hwang et al. [11] test methods of UniGene IDs, LocusLink IDs and Best Match to match genes between HG-U95Av2 and HG-U133A arrays. They experimented with 14 samples on both generations of arrays. Correlation coefficients indicate that Best Match demonstrates higher reproducibility than the other two methods. In this study, we thus adopt the method of Best Match for gene mapping between HG-U95Av2 and HG-U133A. Most probe sets between these two generations have one-to-one correspondence. For few multiple (HG-U95Av2)-to-one (HG-U133A) mappings, we select the first HG-U95Av2 probe set in the Affy comparison spreadsheet to make them one-to-one. Though Affymetrix also provides a comparison table for HuGeneFL and HG-U95Av2 (http://www.affymetrix.com/Auth/support/downloads/comparisons/PN600444HumanFLComp.zip), Best Match is not defined and multiple-to-one relations are not directly available. The situation for HuGeneFL and HG-U95Av2 is thus more complicated as multiple-to-multiple relations occur. To generate matched probe sets, we follow previous studies [10,14,16] and have the following procedure. For any given HuGeneFL probe set, from its corresponding

Table 1
Key characteristics of all analyzed data

Study reference[a]	Institution[b]	Microarray generation	No. of samples	Cancer subtypes for analysis {# in the subtype}
Acute lymphoblastic leukemia (ALL)				
Yeoh et al. [24] (ALL-95)	SJCRH	U95Av2	203[c]	*t*(12;21) {59} HD > 50 {47} *t*(1;19) {9}
Ross et al. [25] (ALL-133)	SJCRH	U133A	132	*t*(12;21) {20} HD > 50 {17} *t*(1;19) {18}
Acute myeloid leukemia (AML)				
Ross et al. [26] (AML-1)	SJCRH	U133A	130	*t*(15;17) {15} inv(16) {14} *t*(8;21) {21}
Valk et al. [27] (AML-2)	Erasmus MC	U133A	285	*t*(15;17) {18} inv(16) {19} *t*(8;21) {22}
Gutiérrez et al. [28] (AML-3)	Salamanca	U133A	43	*t*(15;17) {10} inv(16) {4} *t*(8;21) {0}
Breast cancer				
West et al. [29] (Breast-FL)	DUMC	HuGeneFL	49	ER+ {25} ER− {24}
Huang et al. [30] (Breast-95)	KF-SYSCC	U95Av2	89	ER+ {74} ER− {15}
Wang et al. [31] (Breast-133)	Erasmus MC	U133A	286	ER+{209} ER− {77}

[a] URLs of data sets:
ALL-95: http://www.stjuderesearch.org/data/ALL1/index.html
ALL-133: http://www.stjuderesearch.org/data/ALL3/index.html
AML-1: http://www.stjuderesearch.org/data/AML1/index.html
AML-2: ftp://ftp.ncbi.nih.gov/pub/geo/data/geo/raw_data/series/GSE1159/GSE1159_RAW.tar
AML-3: ftp://ftp.ncbi.nih.gov/pub/geo/data/geo/raw_data/series/GSE1729/GSE1729_RAW.tar
Breast-FL: http://data.cgt.duke.edu/west.php
Breast-95: http://data.cgt.duke.edu/lancet.php
Breast-133: ftp://ftp.ncbi.nih.gov/pub/geo/data/geo/by_series/GSE2034_family.soft.gz.

[b] SJCRH, St. Jude Children's Research Hospital; Erasmus MC, Erasmus University Medical Center; Salamanca, Hospital Universitario de Salamanca; DUMC, Duke University Medical Center; KF-SYSCC, Koo Foundation Sun Yat-Sen Cancer Center.

[c] Exclusion of overlapping cases with ALL-133.

ones in HG-U95Av2, we select the one with the highest overlap percentage in the column of "old (HuGeneFL) → new (HG-U95Av2) sequence relationship". This procedure leads to multiple (HuGeneFL)-to-one (HG-U95Av2) matchings. We then select the first of the multiple HuGeneFL probe sets to have one-to-one relationships. This step is the same as how we process HG-U95Av2 and HG-U133A arrays. Finally, we get a list of 5979 common probe sets between HuGeneFL and HG-U95Av2, and a list of 9530 probe sets between HG-U95Av2 and HG-U133A. The intersection of the above two lists gives 5045 probe sets, which are shared across three generations.

2.3. Rank-based normalization

Previous work has shown that considering a gene's rank within a chip instead of using its expression value better eliminates systematic biases and thus improves the classification accuracy [17,19,20]. There are some variants of the rank-based normalization. The simplest one replaces the expression value of a gene with its rank among expression values of a single chip [35,36]. This method is considered in our study. Quantile normalization is another rank-based approach [37,38]. It calculates a value for each rank level by taking the average of the expression values of that particular rank in available arrays, and then replaces the expression value of each gene by the associated reference value of its rank. Median rank scores is also a rank-based approach. This variant derives the median of each gene among the available arrays and sorts those medians as the reference value of a particular rank [19,20]. Tsodikov et al. [35] show that replacing expression values by ranks performs well in terms of selecting differentially expressed genes. Qiu et al. [38] also reveal that this simple scheme outperforms the quantile normalization method in reduction of between-gene dependence and identification of differential

genes. Thus in this work we investigate if a direct replacement of expressions by ranks is effective in cross-generation and cross-laboratory comparisons. Below we provide details of the adopted procedure.

First of all, we obtain common genes from data of each cancer type. The rank-based normalization method then replaces the value of each probe set with its rank in the set of common genes. Next, gene selection is performed with SAM (Significance Analysis of Microarrays) [21] to identify the list of differentially expressed genes based on the training data set of each experiment. When applying SAM for gene selection, the parametric statistical test (*t*-test) is used. For other parameters, "Two class unpaired" is selected as the response type, while the number of permutation and the number of KNN neighbors are set as 300 and 10, respectively. Besides, the logged flag is turned off. The FDR (false discovery rate) is set as 5%. As far as the testing data is concerned, the same procedure of replacing expression values with ranks is applied. After that, the list of differential probe sets selected based on the training data set filters out unwanted probe sets in the testing data set.

2.4. Predictions

In each experiment, one data set from Table 1 is for training and another data set (across generations or laboratories) is for testing. The *k*-Nearest Neighbors (KNN) [22] is employed in the prediction task. For any instance in the test set, KNN predicts its class by the majority class of its *k* closest neighbors in the training set. The distance between any two data instances is by the Euclidean metric. Since the performance of KNN depends on the parameter *k*, in data classification one usually implements a validation procedure to select it. Here we consider leave-one-out cross-validation (LOO CV). For any given *k*, LOO CV sequentially singles one training instance out for validation. That is, KNN predicts this instance by checking its neighbors in the remaining set. The value *k* with the best LOO CV accuracy is then applied to predict the independent testing data. In our experiments, we consider odd integers from 1 to 17 to search for the best *k*. Values beyond this range do not give better LOO CV. In addition to KNN, we also conducted preliminary experiments using support vector machines (SVM). Results are similar, so subsequently we discuss only results of using KNN.

Table 2
A comparison of cross-generation/laboratory predictions: using rank levels and expression values (Exp. val.)

Training → Testing	Cancer subtype	Accuracy		True positive rate	
		Rank	Exp. val.	Rank	Exp. val.
Acute lymphoblastic leukemia (ALL)					
ALL-95 → ALL-133	*t*(12;21)	**96.2**	68.1	**75.0**	**75.0**
	HD > 50	91.6	**92.4**	**94.4**	**94.4**
	t(1;19)	**100**	97.7	**100**	83.3
ALL-133 → ALL-95	*t*(12;21)	**97.0**	93.1	**100**	91.5
	HD > 50	**95.0**	87.1	**78.7**	74.4
	t(1;19)	**99.5**	98.5	**88.8**	77.7
Acute myeloid leukemia (AML)					
AML-1 → AML-2	*t*(15;17)	**99.2**	**99.2**	**94.4**	88.8
	t(8;21)	**100**	99.6	**100**	**100**
	inv(16)	**97.8**	91.2	**100**	36.8
AML-1 → AML-3	*t*(15;17)	**100**	95.3	**100**	80.0
	inv(16)	**97.6**	88.3	**100**	50.0
AML-2 → AML-1	*t*(15;17)	**100**	**100**	**100**	**100**
	t(8;21)	**99.2**	96.9	**95.2**	80.9
	inv(16)	**99.2**	90.7	**92.8**	21.4
AML-2 → AML-3	*t*(15;17)	**97.6**	93.0	**90.0**	**90.0**
	inv(16)	**97.6**	95.3	**75.0**	50.0
AML-3 → AML-1	*t*(15;17)	99.2	**100**	**100**	**100**
	inv(16)	**90.0**	89.2	**7.1**	0
AML-3 → AML-2	*t*(15;17)	**99.6**	97.1	**100**	66.6
	inv(16)	**98.2**	93.3	**94.7**	0
Breast cancer					
Breast-FL → Breast-95	ER−	**86.5**	83.1	73.3	**80.0**
Breast-FL → Breast-133	ER−	**89.2**	81.1	78.0	**80.5**
Breast-95 → Breast-FL	ER−	**87.8**	53.1	**95.8**	41.7
Breast-95 → Breast-133	ER−	**86.4**	75.2	**79.2**	10.4
Breast-133 → Breast-FL	ER−	**85.7**	75.5	**75.0**	54.2
Breast-133 → Breast-95	ER−	**86.5**	83.1	**40.0**	**40.0**

In each row, we boldface the value which gives higher accuracy (true positive rate).

3. Results and discussion

We conduct experiments on three cancer types of data sets, which are summarized in Table 1. This section first compares prediction results under two ways of processing arrays: one directly uses gene expression values, and the other transforms expression values to rank levels. The experimental procedure has several variants of normalization, so subsequently we check their respective effects. Finally, we discuss how the cross-validation analysis might mislead the conclusion about the performance of cross-generation/laboratory predictions.

3.1. A comparison between using expression values and rank information

To perform cross-generation/laboratory predictions, one can prepare training and testing data by directly using expression values of significant genes. However, the scale of each gene may vary due to different chip generations or experimental environments. We can instead use each gene's rank in the same subset of significant genes. Experiments below compare which method is better.

For any cancer type, each experiment considers one subtype as the target prediction label. A data set of this cancer type is used as the training data of two classes: whether an array is associated with the specified subtype or not. For each array in another set (called testing data), we then predict its class label and calculate the accuracy. This procedure is repeated for every two sets of the same cancer type. One exception is that AML-3 has no $t(8;21)$ arrays, so for this subtype AML-3 is not used as the testing set.

Except the difference on using expression values or rank levels, all other settings follow the procedure in Fig. 1. Table 2 gives results of the comparison. Some cancer subtypes have very few arrays in both training and testing sets, so we have the so called "unbalanced problems" in data classification. Accuracy may not be the best evaluation criterion in such situations. Predicting everything not in the specified subtype yields a high but misleading accuracy value. Thus Table 2 also presents the true positive rate, which is

$$\frac{\text{Number of correctly predicted data in the subtype}}{\text{Number of data in the subtype}}.$$

This measurement better reveals the performance on identifying the specified subtype. For example, when training AML-2 (by expression values) to identify the subtype inv(16), the prediction of AML-1 is erroneous (21.4% true positive rate in Table 2), but the accuracy is very high (90.7%). Since breast cancer without ER (ER−) is considered poorly response to treatment, we report ER− as positive prediction.

Table 2 clearly shows that using the ranks of the selected genes within an array consistently produces better results than using the original expression values. For ALL and AML, the prediction by using rank levels is excellent. One exception is to predict inv(16) by using AML-3 as the training set. Since AML-3 contains only four inv(16) arrays, there is no enough information to discriminate this subtype from others.

Both methods give worse accuracy in predicting breast cancer subtypes. As indicated earlier, three data sets of this cancer type are the most heterogeneous. They are cross-generation as well as cross-laboratory, but ALL sets are cross-generation only and AML sets are cross-laboratory only. Training Breast-95 to classify the other two sets gives much lower accuracy than other cases. We suspect the reason is that Breast-95 is the most unbalanced (74 ER+ and 15 ER− arrays) among the three breast cancer sets.

To check if in typical microarray comparisons the proposed rank-based approach is more robust than using expression values, we also conduct experiments on each data set alone, by randomly separating each data set into a training set (two thirds) and an independent testing set (one third). This procedure is repeated ten times in each experiment, and the average accuracy is reported in Table 3. We again find that using ranks leads to higher accuracy.

Though our procedure in Fig. 1 is rather simple, steps such as selecting differential genes are important. Table 4 presents results without removing any genes. The accuracy is generally lower than that in Table 2. Thus even if expression values have been transformed to rank levels, selecting important genes is still essential.

Table 3
A comparison of rank levels and expression values (Exp. val.) by randomly splitting each data set to training (two thirds) and testing (one third)

Data	Cancer subtype	Accuracy		True positive rate	
		Rank	Exp. val.	Rank	Exp. val.
Acute lymphoblastic leukemia (ALL)					
ALL-95	$t(12;21)$	**97.7**	85.7	**99**	80
	HD > 50	**94.4**	94.2	**84.3**	82.5
	$t(1;19)$	**100**	99.1	**100**	80
ALL-133	$t(12;21)$	**98.8**	95.7	**100**	**100**
	HD > 50	**97.4**	94.7	**90**	73.3
	$t(1;19)$	**97.7**	93.6	**100**	81.6
Acute myeloid leukemia (AML)					
AML-1	$t(15;17)$	**100**	97.4	**100**	82
	$t(8;21)$	**96.1**	86.1	**95.7**	65.7
	inv(16)	**96.1**	87.2	**94**	24
AML-2	$t(15;17)$	**99.6**	99.1	**100**	90
	$t(8;21)$	**100**	98.5	**100**	88.7
	inv(16)	**98.1**	95.6	**100**	64.2
AML-3	$t(15;17)$	**100**	96.6	**100**	90
	inv(16)	85.9	**87.9**	0	**10**
Breast cancer					
Breast-FL	ER−	**89.3**	88.0	**88.6**	84.3
Breast-95	ER−	**81.1**	79.7	**20.0**	10.0
Breast-133	ER−	**84.9**	83.4	**67.0**	62.6

In each row, we boldface the value which gives higher accuracy (true positive rate).

3.2. Additional normalization for expression values and rank levels

While Table 2 indicates that rank levels are better than expression values, we investigate if the same conclusion stands after slight changes of the experimental procedure. One issue we intend to study is the effect of gene-wise normalization. That is, after selecting significant genes, for each gene we normalize ranks or expression values in all training arrays to have mean zero and standard deviation one. In data classification such a procedure is called feature scaling (normalization). The purpose is to avoid the possible dominance of genes having large values. The same scaling factors are then employed to normalize the testing data. Table 4 lists accuracy with and without gene-wise normalization. Using expression values, the accuracy with

Table 4
A comparison between variants of the experimental procedure: The first two rows list different settings

Gene selection		Y	Y		Y	Y	
Genewise normalization		Y			Y		
Training → Testing	Subtype	Rank			Exp. val.		
Acute lymphoblastic leukemia (ALL)							
ALL-95 → ALL-133	*t*(12;21)	96.9	96.2	87.1	**98.4**	68.1	80.3
	HD > 50	93.1	91.6	93.9	**97.7**	92.4	81.8
	t(1;19)	**100**	**100**	86.3	99.2	97.7	86.3
ALL-133 → ALL-95	*t*(12;21)	**97.5**	97.0	92.6	**97.5**	93.1	79.3
	HD > 50	94.5	**95.0**	94.0	78.3	87.1	82.2
	t(1;19)	**99.5**	**99.5**	**99.5**	**99.5**	98.5	98.0
Acute myeloid leukemia (AML)							
AML-1 → AML-2	*t*(15;17)	**99.6**	99.2	99.2	98.9	99.2	89.8
	t(8;21)	**100**	**100**	99.6	**100**	99.6	91.9
	inv(16)	98.2	97.8	88.0	**98.5**	91.2	92.9
AML-1 → AML-3	*t*(15;17)	97.6	**100**	**100**	95.3	95.3	86.0
	inv(16)	97.6	97.6	95.3	**100**	88.3	88.3
AML-2 → AML-1	*t*(15;17)	**100**	**100**	99.2	**100**	**100**	93.8
	t(8;21)	**99.2**	**99.2**	98.4	98.4	96.9	84.6
	inv(16)	96.9	**99.2**	**99.2**	98.4	90.7	88.4
AML-2 → AML-3	*t*(15;17)	**97.6**	**97.6**	95.3	95.3	93.0	81.3
	inv(16)	97.6	97.6	95.3	**100**	95.3	90.6
AML-3 → AML-1	*t*(15;17)	99.2	99.2	95.3	97.6	**100**	78.4
	inv(16)	89.2	90.0	90.7	**92.3**	89.2	89.2
AML-3 → AML-2	*t*(15;17)	97.8	**99.6**	96.4	97.1	97.1	93.6
	inv(16)	94.0	**98.2**	94.7	95.0	93.3	93.3
Breast cancer							
Breast-FL → Breast-95	ER−	84.3	**86.5**	68.5	84.3	83.1	80.9
Breast-FL → Breast-133	ER−	85.7	**89.2**	57.3	85.0	81.1	82.2
Breast-95 → Breast-FL	ER−	81.6	**87.8**	59.2	51.0	53.1	51.0
Breast-95 → Breast-133	ER−	80.1	**86.4**	73.1	72.7	75.2	73.1
Breast-133 → Breast-FL	ER−	**87.8**	85.7	85.7	53.1	75.5	81.6
Breast-133 → Breast-95	ER−	82.0	**86.5**	85.4	82.0	83.1	85.4

In each row, we boldface the value which gives the highest accuracy and underline the one with the lowest accuracy.

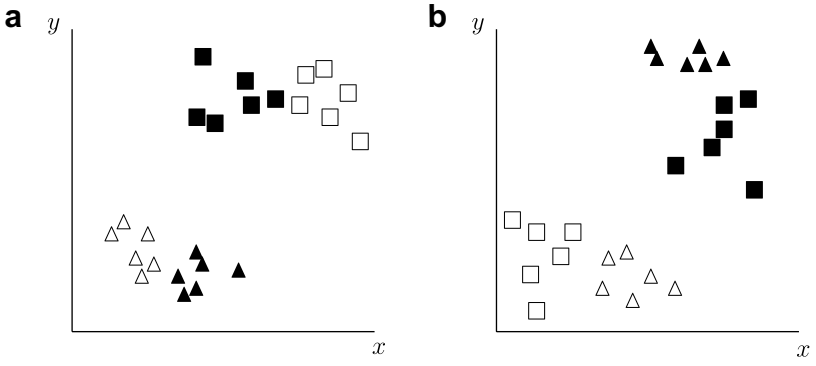

Fig. 2. Distributions of arrays from two different data sets. Data set A: △ (class I), □ (class II). Data set B: ▲ (class I), ■ (class II). (a) Ideal situation; (b) possible practical situation: information from one set cannot discriminate positive/negative arrays of another set. However, the LOO CV accuracy of the combined set is high.

normalization is slightly improved for ALL and AML, but is worse for some cases of breast cancer. Thus one cannot conclude that this normalization is always helpful. For rank levels, the accuracy with/without gene-wise normalization is very similar. It consistently outperforms expression values no matter the gene-wise normalization is performed or not. Overall this normalization has a bigger effect on using expression values than rank levels.

3.3. Pitfalls of reporting cross-validation accuracy

To evaluate the performance of cross-laboratory/generation predictions, earlier we trained data from one source and classified another. However, some studies report the prediction accuracy based on a cross-validation analysis (e.g., [15,19,20]). They pool arrays from different data sets and randomly split the combined set to training and validation sets. Here we argue that the performance of cross-validation analysis should be used in a more careful way when dealing with data sets from different sources. In this type of studies, we expect that arrays from distinct data sets can be merged in the way shown in Fig. 2(a). That is, arrays associated with the same class but different data sets are clustered together. Thus for any given array one can correctly predict its class label. However, it is observed in our study that arrays from two sources may appear in a situation similar to Fig. 2(b). In this example, arrays of one data set are close to each other, so it is difficult to separate arrays in the combined set via their class labels. Training one data set to classify another would result in erroneous predictions. However, for this example, LOO CV accuracy via KNN (or other classification methods) is excellent. For any array singled out for validation, its closest neighbors are arrays in the same class of the same set. Therefore, cross-validation accuracy may overestimate the practical performance of cross-generation/laboratory predictions. In such studies training and testing sets should be from independent sources.

To illustrate the possible overestimation of cross-validation accuracy, Table 5 reports LOO CV accuracy on the combined set of every two sources of the same cancer type. We compare it with the two accuracy values of training one and predicting another. Here expression values are used, so the last column of Table 5 is the same as the result in Table 2. In all situations, CV values are either equally good or much better. Therefore, experimental results based on cross-validation analyses may mislead the predicting power of the data.

Table 5
A comparison of two evaluation methods for cross-generation/laboratory predictions

Leave-one-out				Independent		
Combined set	Subtype	Acc.	TP	Training → Testing	Acc.	TP
Acute lymphoblastic leukemia (ALL)						
ALL-95 + ALL-133	*t*(12;21)	89.5	83.5	ALL-95 → ALL-133	**68.1**	**75.0**
				ALL-133 → ALL-95	93.1	91.5
	HD > 50	96.1	83.0	ALL-95 → ALL-133	92.4	94.4
				ALL-133 → ALL-95	**87.1**	**74.4**
	t(1;19)	99.1	92.5	ALL-95 → ALL-133	97.7	**83.3**
				ALL-133 → ALL-95	98.5	**77.7**
Acute myeloid leukemia (AML)						
AML-1 + AML-2	*t*(15;17)	99.2	93.9	AML-1 → AML-2	99.2	**88.8**
				AML-2 → AML-1	100	100
	t(8;21)	97.3	83.7	AML-1 → AML-2	99.6	100
				AML-2 → AML-1	96.9	80.9
	inv(16)	97.8	84.8	AML-1 → AML-2	**91.2**	**36.8**
				AML-2 → AML-1	**90.7**	**21.4**
AML-1 + AML-3	*t*(15;17)	100	100	AML-1 → AML-3	95.3	**80.0**
				AML-3 → AML-1	100	100
	inv(16)	93.6	66.6	AML-1 → AML-3	**88.3**	**50.0**
				AML-3 → AML-1	89.2	**0**
AML-2 + AML-3	*t*(15;17)	99.3	96.4	AML-2 → AML-3	**93.0**	**90.0**
				AML-3 → AML-2	97.1	**66.6**
	inv(16)	98.4	95.6	AML-2 → AML-3	95.3	**50.0**
				AML-3 → AML-2	**93.3**	**0**
Breast cancer						
Breast-FL + Breast-95	ER−	87.7	74.4	Breast-FL → Breast-95	83.1	80.0
				Breast-95 → Breast-FL	**53.1**	**41.7**
Breast-95 + Breast-133	ER−	81.9	43.5	Breast-95 → Breast-133	**75.2**	**10.4**
				Breast-133 → Breast-95	83.1	40.0
Breast-133 + Breast-FL	ER−	84.2	69.3	Breast-133 → Breast-FL	**75.5**	**54.2**
				Breast-FL → Breast-133	81.1	80.5

Left: LOO CV accuracy by combining data from two sources. Right: accuracy of predicting one data set after training another. Here expression values are used, so the last column is the same as that of Table 2. Acc. and TP mean accuracy and true positive rate, respectively. Accuracy (true positive rate) significantly lower than that by LOO CV is bold-faced.

4. Conclusions

We conduct a detailed study on cross-generation and cross-laboratory predictions of Affy microarray data. A focus is on investigating if using expression values is suitable. Experiments show that an alternative way of using simple rank levels gives more stable prediction results. We also discuss some pitfalls in evaluating cross-generation/laboratory predictions.

The framework proposed in this paper is rather simple. As more studies involve such cross-generation and cross-laboratory predictions, we expect our approach to be very useful. For example, existing data can be trained to predict arrays from a new generation of Affymetrix human oligonucleotide array, Plus 2.0. Future work includes experiments on more cancer types or future Affy generations.

Acknowledgment

The study was supported by the grants from Mackay Memorial Hospital (MMH-E-95009) and NHRI-EX96-9434SI.

References

[1] Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics 2005;21(20):3896–904.
[2] Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mRNA comparisons. Stat Appl Genet Mol Biol 2004;3:19.
[3] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29:365–71.
[4] Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS. Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics 2002;18(3):405–12.
[5] Mah N, Thelin A, Nikolaus TLS, Kühbacher T, Gurbuz Y, Eickhoff H, et al. A comparison of oligonucleotide and cDNA-based microarray systems. Physiol Genom 2004;16:361–70.
[6] Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, et al. Multiple-laboratory comparison of microarray platforms. Nat Methods 2005;2:345–50.
[7] Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J. Independence and reproducibility across microarray platforms. Nat Methods 2005;2(5):337–44.
[8] Bammler T, Beyer R, Bhattacharya S, Boorman G, Boyles A, Bradford B, et al. Standardizing global gene expression analysis between laboratories and across platforms. Nat Methods 2005;2(5):351–6.
[9] Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci USA 2004;101(25):9309–14.
[10] Nimgaonkar A, Sanoudou D, Butte1 AJ, Haslett JN, Kunkel LM, Beggs AH, et al. Reproducibility of gene expression across generations of Affymetrix microarrays. BMC Bioinformatics 2003;4:27.
[11] Hwang KB, Kong SW, Greenberg SA, Park PJ. Combining gene expression data from different generations of oligonucleotide arrays. BMC Bioinformatics 2004;5:159.
[12] Kong SW, Hwang KB, Kim RD, Zhang BT, Greenberg SA, Kohane IS, et al. CrossChip: a system supporting comparative analysis of different generations of Affymetrix arrays. Bioinformatics 2005;21(9):2116–7.
[13] Elo LL, Lahti L, Skottman H, Kylaniemi M, Lahesmaa R, Aittokallio T. Integrating probe-level expression changes across generations of Affymetrix arrays. Nucleic Acids Res 2005;33(22):e193.
[14] Bhattacharya S, Mariani TJ. Transformation of expression intensities across generations of Affymetrix microarrays using sequence matching and regression modeling. Nucleic Acids Res 2005;33(18):e157.
[15] Bloom G, Yang IV, Boulware D, Kwong KY, Coppola D, Eschrich S, et al. Multi-platform, multi-site, microarray-based human tumor classification. Am J Pathol 2004;164:9–16.
[16] Jiang H, Deng Y, Chen HS, Tao L, Sha Q, Chen J, et al. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics 2004;5:81.
[17] Xu L, Tan AC, Naiman DQ, Geman D, Winslow RL. Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. Bioinformatics 2005;21(20):3905–11.
[18] Tothill RW, Kowalczyk A, Rischin D. An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. Cancer Res 2005;65(10):4031–40.
[19] Warnat P, Eils R, Brors B. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. BMC Bioinformatics 2005;6:265.
[20] Tödling J, Spang R. Assessment of five microarray experiments on gene expression profiling of breast cancer. Poster Presentation RECOMB 2003. Available from: http://citeseer.ist.psu.edu/611350.html.
[21] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001;98:5116–21.
[22] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inform Theory 1967;13:21–7.
[23] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2005. Available from: http://www.R-project.org.
[24] Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, et al. Pediatric lymphoblastic leukemia by gene expression profiling. Cancer Cell 2002;1:133–43.
[25] Ross ME, Zhou X, Song G, Shurtleff SA, Girtman K, Williams WK, et al. Classification of pediatric acute lymphoblastic leukemia by gene expression profile. Blood 2003;102:2951–9.
[26] Ross ME, Mahfouz R, Onciu M, Liu HC, Zhou X, Song G, et al. Gene expression profiling of pediatric acute myelogenous leukemia. Blood 2004;104:3679–87.
[27] Valk PJ, Verhaak RG, Beijen MA, Erpelinck CA, van Doorn-Khosrovani SB, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. N Engl J Med 2004;350:1617–28 .
[28] Gutiérrez NC, López-Pérez R, Hernández JM, Isidro I, González B, Delgado M, et al. Gene expression profile reveals deregulation of genes with relevant functions in the different subclasses of acute myeloid leukemia. Leukemia 2005;19:402–9.
[29] West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA 2001;98(20):11462–7.
[30] Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, et al. Gene expression predictors of breast cancer outcomes. Lancet 2003;361(9369):1590–6.
[31] Wang Y, Klijn JGM, Zhang Y, Sieuwerts AM, Look MP, Yang F, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005;365(9460):671–9.
[32] Pontius J, Wagner L, Schuler G. UniGene: a unified view of the transcriptome. In: NCBI Handbook. Bethesda, MD: National Center for Biotechnology Information; 2003.
[33] Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 2005;33:D54–8.

[34] Affymetrix. User's guide to product comparison spreadsheets 2003. Available from: http://www.affymetrix.com/support/technical/manual/.
[35] Tsodikov A, Szabo A, Jones D. Adjustments and measures of differential expression for microarray data. Bioinformatics 2002;18(2):260–1.
[36] Szabo A, Boucher K, Carroll W, Klebanov L, Tsodikov A, Yakovlev A. Variable selection and pattern recognition with gene expression data generated by the microarray technology. Math Biosci 2002;176:71–98.
[37] Bolstad B, Irizarry R, Astrand M, Speed T. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19(2):185–93.
[38] Qiu X, Brooks AI, Klebanov L, Yakovlev A. The effects of normalization on the correlation structure of microarray data. BMC Bioinformatics 2005;6:120.