

iPcc: a novel feature extraction method for accurate disease class discovery and prediction

Xianwen Ren¹, Yong Wang^{2,3}, Xiang-Sun Zhang^{2,3,*} and Qi Jin^{1,*}

¹MOH Key Laboratory of Systems Biology of Pathogens, Institute of Pathogen Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China, ²Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China and ³National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing, 100190, China

Received February 5, 2013; Revised March 24, 2013; Accepted April 10, 2013

ABSTRACT

Gene expression profiling has gradually become a routine procedure for disease diagnosis and classification. In the past decade, many computational methods have been proposed, resulting in great improvements on various levels, including feature selection and algorithms for classification and clustering. In this study, we present iPcc, a novel method from the feature extraction perspective to further propel gene expression profiling technologies from bench to bedside. We define ‘correlation feature space’ for samples based on the gene expression profiles by iterative employment of Pearson’s correlation coefficient. Numerical experiments on both simulated and real gene expression data sets demonstrate that iPcc can greatly highlight the latent patterns underlying noisy gene expression data and thus greatly improve the robustness and accuracy of the algorithms currently available for disease diagnosis and classification based on gene expression profiles.

INTRODUCTION

With the rapid development of high-throughput technologies, gene expression profiling based on microarrays or next-generation sequencing techniques have been widely applied in clinical research (1–9). The big advantage of simultaneously measuring the expression levels of thousands of genes facilitates informative and accurate disease diagnosis and classification. However, it also incorporates many irrelevant genes, producing a feature vector with extremely high dimensionality (10,11). The situation is much exacerbated by sample limitations for algorithmic training, leading to the curse of dimensionality problems.

To surmount the difficulties caused by the curse of dimensionality problem, various approaches have been proposed and developed on different levels of the experimental and analytical protocol. On the experimental level, increasing the number of samples involved in the studies to enhance the statistical power and measuring the expression levels of a set of selected genes involved in a certain signaling pathway are two possible solutions. However, increasing the number of samples is generally difficult because of eligible sample limitations and the huge cost of high-throughput measurements. Furthermore, measuring genes involved in only selected pathways can introduce artificial biases. In fact, the rapid development of high-throughput technologies produces more and more information about samples, allowing unbiased investigation of the molecular truth of various biomedical phenomena. For example, RNA-Seq (deep sequencing the transcriptomes of samples) can detect expression levels of novel genes that are not annotated in the reference genomes, compared with the traditional gene expression profiling microarrays (8).

On the analytical level, adapting and developing algorithms capable of handling high-dimension data sets, computationally selecting a small set of relevant genes from the huge gene list for subsequent diagnosis and prognosis and extracting a small set of virtual genes that are certain functions of the real genes, are representative computational schemes to bypass the curse of dimensionality. In view of machine learning, these three schemes correspond to developing new algorithms of clustering and classification, feature selection and feature extraction. Until now, many algorithms suitable for dealing with high-dimension data sets [e.g. support vector machine (12–14), random forest (15–17), naïve Bayes classifier (18–20) and *k*-means (21–24)] have been applied to disease sample classification and clustering based on gene expression profiles. Various feature selection methods implemented by mathematical programming (25), Bayesian inference (26,27),

*To whom correspondence should be addressed. Tel: +8610 6787 7732; Fax: +8610 6787 7736; Email: zdsys@vip.sina.com
Correspondence may also be addresses to Xiang-Sun Zhang. Tel: +86 10 62561440; Fax: +86 10 62561963; Email: zxs@amt.ac.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

ant colony algorithm (28) and mutual information (29,30) have been proposed and developed, facilitating accurate identification of relevant disease genes. The concept of virtual genes or meta-genes based on principle component analysis and nonnegative matrix factorization was introduced from the feature extraction perspective to reduce the number of dimensions of the data sets in the meta-gene space (31–34). These methods, based on new clustering/classification algorithms, gene selection and virtual genes, greatly improve the accuracy of disease diagnosis and prognosis on gene expression measurement but neglect an important feature space formed by sample correlations.

In this study, we propose a novel method to negotiate the dimensionality problem in clinical gene expression studies. Similar to the available feature extraction methods, we define a series of novel features for samples based on the gene expression profiles by iterative use of Pearson's correlation coefficient. These features form a feature space, referred to as 'correlation feature space'. We demonstrate that the latent structures between samples can be easily highlighted based on the newly defined features, even though many irrelevant genes are included in the original data. Furthermore, we illustrate that the newly defined features can greatly improve the accuracy of the currently available clustering and classification algorithms.

In current biomedical studies, two typical strategies are widely used to analyze gene expression data for clinical purposes: class discovery and class prediction (5). Class discovery tries to discover new disease subtypes based on gene expression patterns, while class prediction tends to assign particular samples to well-defined disease classes. Both strategies have significant potential to improve cancer diagnosis, prognosis and therapies. On real prostate cancer, leukemia and psoriasis data sets, we demonstrate that the newly defined features greatly leverage the power of the current class discovery and class prediction methods. Therefore, we propose a novel type of noise-resistant features based on iterative Pearson correlation coefficients, which is substantially helpful to boost gene expression profiles from bench to bedside.

MATERIALS AND METHODS

Sample correlation feature space defined by gene expression profiles

Given a gene expression data set $\mathbf{X}_{m \times n}$, in which the expression of n genes is measured for m samples and x_{ik} denotes the expression level of gene k in sample i , we first calculate $f_{ij}^{(1)}$, the correlation of the gene expression profiles of samples i and j , by the Pearson correlation coefficient (35):

$$f_{ij}^{(1)} = \frac{\sum_{k=1}^n (x_{ik} - x_{i-})(x_{jk} - x_{j-})}{\sqrt{\sum_{k=1}^n (x_{ik} - x_{i-})^2} \sqrt{\sum_{k=1}^n (x_{jk} - x_{j-})^2}} \quad (1)$$

where x_{ik} and x_{i-} are the expression level of gene k and the average gene expression level of sample i , respectively. The variables x_{jk} and x_{j-} are the expression level of gene k and the average gene expression level of sample j , respectively. Finally, we obtain a symmetric matrix $\mathbf{F}_{m \times m}^{(1)}$ in which $f_{ij}^{(1)}$ is its element. Function $f_{ij}^{(1)}$ measures how the gene expression profiles of samples i and j correlate to each other. However, here we interpret it as a feature of sample i that is scored by sample j . Based on this interpretation, $\mathbf{F}_{m \times m}^{(1)}$ becomes a matrix of m samples and m features. Thereafter, we can define $f_{ij}^{(2)}$ as follows:

$$f_{ij}^{(2)} = \frac{\sum_{k=1}^m (f_{ik}^{(1)} - f_{i-}^{(1)})(f_{jk}^{(1)} - f_{j-}^{(1)})}{\sqrt{\sum_{k=1}^m (f_{ik}^{(1)} - f_{i-}^{(1)})^2} \sqrt{\sum_{k=1}^m (f_{jk}^{(1)} - f_{j-}^{(1)})^2}} \quad (2)$$

We name $f_{ij}^{(1)}$, $j \in \{1, \dots, m\}$ as the first-order correlation features of sample i and $f_{ij}^{(2)}$, $j \in \{1, \dots, m\}$ as the second-order correlation features of sample i . Generally, we can define the t -order correlation features of sample i as follows:

$$f_{ij}^{(t)} = \frac{\sum_{k=1}^m (f_{ik}^{(t-1)} - f_{i-}^{(t-1)})(f_{jk}^{(t-1)} - f_{j-}^{(t-1)})}{\sqrt{\sum_{k=1}^m (f_{ik}^{(t-1)} - f_{i-}^{(t-1)})^2} \sqrt{\sum_{k=1}^m (f_{jk}^{(t-1)} - f_{j-}^{(t-1)})^2}} \quad (3)$$

All the features $f_{ij}^{(1)}, f_{ij}^{(2)}, \dots, f_{ij}^{(t)}, \dots$, form a correlation feature space.

Because the correlation features are defined based on Pearson correlation coefficients, we named the computational method iterative Pearson correlation coefficients (iPcc). It can be seen that correlation features have several obvious properties: (i) $f_{ii}^t = 1$ for any i and t ; (ii) $f_{ij}^t = f_{ji}^t$ for any i, j and t ; and (iii) $|f_{ij}^t| \leq 1$ for any i, j and t . Below we compose three toy gene expression data sets of six genes (for example) to demonstrate numerically other properties of sample correlation features in an oversimplified setting.

In toy Data set 1, three samples are composed in which samples 1 and 2 (s1 and s2) represent two different classes, whereas sample 3 (s3) is more similar to sample 1 (Figure 1). By calculating the first-order correlation features of these three samples, we can observe that s3 has a correlation feature vector of (0.3, -0.3, 1) that is similar to that of s1, which is (1, -1, 0.3). By calculating the second-order correlation features, the feature vector of s3 becomes (0.7, -0.7, 1) while the feature vector of s1 is (1, -1, 0.7). When the fifth-order or higher correlation features are calculated, the feature vectors of s3 and s1 become the same, whereas the feature vectors of s3 and s2 become opposite. In other words, the similarity of s3 and s1 measured by Pearson correlation coefficients based on their features converges from 0.3 to 1, while the similarity of s3 and s2 converges from -0.3 to -1. Thus, in this oversimplified setting, $|f_{ij}^{t+1}| \geq |f_{ij}^t|$ holds true.

In toy Data set 1, the initial feature profile of s3 is closer to that of s1, and the high-order correlation feature profile of s3 converges to the same as that of s1, suggesting that

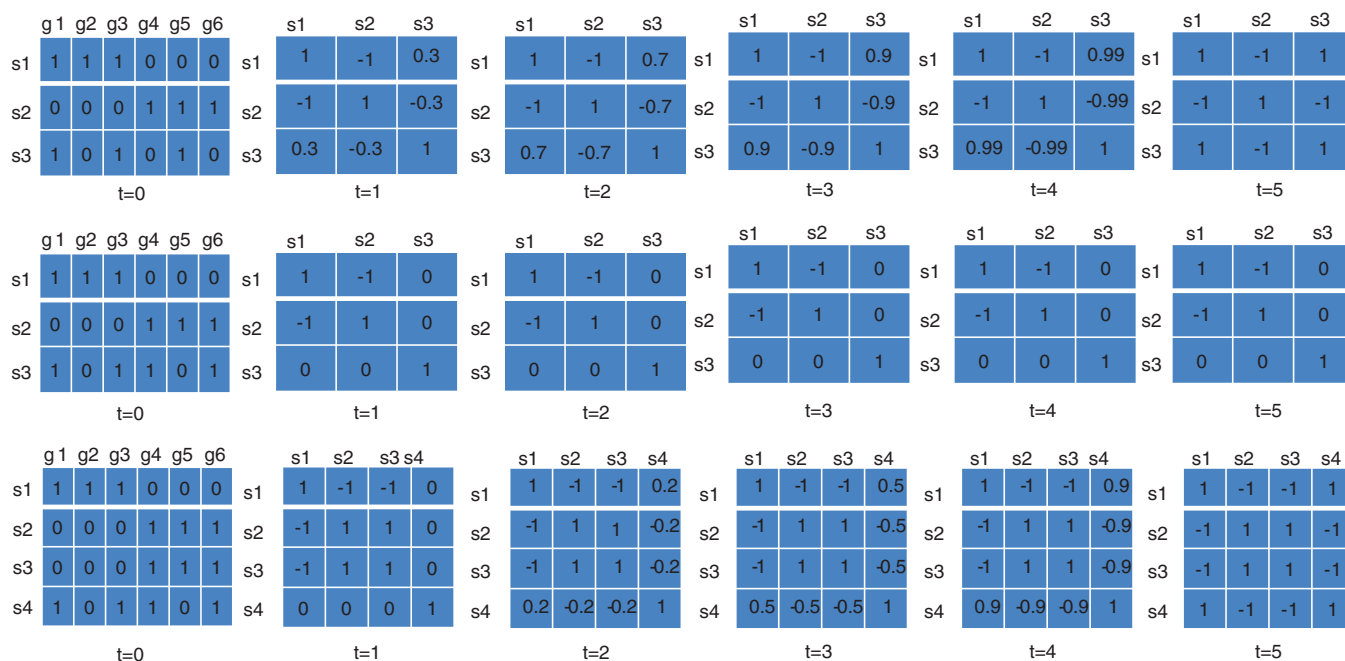


Figure 1. Three toy examples demonstrate correlation features generated by iPcc. In the top line, three samples are depicted by six virtual genes in which s1 and s2 have distinct gene expression profiles, and s3 has a gene expression profile similar to that of s1. Based on the original features ($t = 0$), the first- ($t = 1$), second- ($t = 2$), third- ($t = 3$), fourth- ($t = 4$) and fifth-order ($t = 5$) correlation features were constructed by iPcc. The fifth-order correlation features suggest that s3 belongs to the s1 class. In the middle line, s1 and s2 are the same as those in the top line, whereas s3 is similar to neither s1 nor s2. Those constructed correlation features suggest that s3 is similar to neither s1 nor s2 even with higher orders. In the bottom line, s1 and s2 are still the same as in the top and middle lines, s3 is the same as s2, and s4 is the same as s3 in the middle line. Based on the first-order correlation features, s4 is found to be similar to neither s1 nor s2 and s3. However, on the second-order correlation features, s4 is found to be weakly similar to s1. Increasing the order to the fifth order, s4 is found to have the same fifth-order correlation features as s1, suggesting that iPcc can reveal sample similarity hiding in high-order correlation features.

correlation features can enlarge weak pattern underlying the raw data set. In toy Data set 2, we demonstrate that correlation features can also preserve independent relationships among samples. We compose an initial feature profile for s3 that has equal distance to those of s1 and s2. The first-order correlation feature vector of s3 is (0, 0, 1), which is vertical to both of the feature vectors of s1 and s2. The second-order or higher correlation feature vectors of s3 are the same as the first-order correlation feature vector of s3, suggesting that the independence of s3 from s1 and s2 is conserved during the iterative correlation feature extraction.

In toy Data set 3, we conceive two different classes of unbalanced class sizes (s1 for Class 1, s2 and s3 for Class 2). We further conceive s4 that has equal distance from s1, s2 and s3. The first-order correlation feature vector of s4 is (0, 0, 0, 1), suggesting that s4 forms an independent class from s1, s2 and s3. However, the second-order correlation feature vector of s4 indicates weak similarity to that of s1. The higher-order correlation feature vectors of s4 highlight the similarity of s4 to s1 until their Pearson correlation coefficient converges to 1. Thus, correlation features not only can enlarge the weak patterns underlying the raw data set and preserve the balanced independent relationships, but also can highlight those patterns hidden in the higher feature space.

Simulated and real gene expression data sets

We evaluated our method on both simulated and real gene expression data sets. First, we simulated a series of

artificial gene expression data sets to show the properties and performance of the correlation features against the magnitude of uncertainty of gene expression measurements and the ratio of irrelevant genes to relevant genes. We designed three different sample classes. For each sample class, we generated two relevant genes with expression levels randomly sampled from a normal distribution with mean 1 and given a standard deviation σ from 0.1 to 0.5. We simulated 0-, 10-, 50- and 100-fold irrelevant genes according to a normal distribution with mean 0 and the same standard deviation as those six (2×3) relevant genes and added into the data sets. For each simulated data set, we applied our method and demonstrated the sample relationship by visualizing the sample similarity matrices in Figure 1. Clustering and classification algorithms were used to show the benefits produced by correlation features. We also evaluated iPcc on completely random data sets. We simulated three types of completely random data sets of 150 genes and 150 samples by sampling from the standard normal distribution, the uniform distribution on the interval [0, 1] and a discrete uniform distribution on $\{-1, 0, 1\}$. These completely random data sets provide negative examples to demonstrate the behavior of iPcc in situations no relationships exist.

To evaluate the advantage of correlation features to real-world applications, we examined the clustering and classification performance of a series of widely used

algorithms on three real gene expression data sets. The first real gene expression data set was the gene expression data set of leukemia (5) in which 25 acute myeloid leukemia (AML) samples, 38 B-cell acute lymphoblastic leukemia (ALL) samples and 9 T-cell ALL samples were profiled genome-wide by gene expression microarrays. Among the 72 samples, 7129 probes were available. After removing probes with missing values, three preprocessing steps, including flooring/ceiling, filtering and log10-transformation as described in (36), were applied to select informative probes, resulting in 3571 informative probes. Thus, it is a typical gene expression data set with the curse of dimensionality.

The second data set was a prostate cancer gene expression data set (37) downloaded from the NCBI GEO database (38) with accession number GDS3289, which includes 104 samples and 9483 genes. Among the 104 samples, there are six types of samples: 22 normal samples from benign epithelial cells, 5 prostate cancer samples with atrophic lesions, 20 metastatic prostate cancer samples, 32 localized prostate cancers, 13 prostate cancer samples with prostatic intraepithelial neoplasia and 12 normal samples from benign stromal cells. Accurately revealing sample relationships by either clustering or classification algorithms from the large number of gene expression profiles is important for clinical diagnosis and even therapeutics.

The third data set contains two series of gene expression profiles of normal and psoriasis samples, downloaded from the NCBI database with accession numbers GSE13355 and GSE14905 (39,40). There are 180 samples in GSE13355 (122 normal samples and 58 psoriasis samples) and 82 samples in GSE14905 (49 normal samples and 33 psoriasis samples). In total 54 676 probes were applied to measure the expression levels of human genes. Because the two series of gene expression profiling experiments were conducted in different laboratories, batch effects exist. We used this data set to demonstrate how correlation features reveal the existence of batch effects and how batch effects are removed.

Computational methods for disease class discovery and prediction used in this study

Besides the direct application of iPcc to the gene expression data sets to reveal the underlying patterns between samples, we also illustrate the leverage impact of iPcc on disease class discovery and prediction. For each data set, we used k -means to implement disease class discovery. k -means is a classic clustering algorithm that partition n samples into k clusters in which each sample belongs to the cluster with the nearest mean (24,41). k -means has been applied widely to gene expression data analysis, and various variants have been proposed. Because the computation of k -means is NP-hard, efficient heuristic algorithms are commonly used and generate local optima. Due to the curse of dimensionality caused by the high number of dimensions in gene expression data sets, k -means can give out different local optima during different runs in general. We demonstrate that iPcc can circumvent the curse of dimensionality and greatly enhance the

power and robustness of k -means on gene expression data sets. This feature extraction method can also greatly improve the performance of supervised algorithms for disease class prediction. We take the naïve Bayes classifier (41) as an example to illustrate how iPcc boosts the accuracy of disease class prediction based on gene expression profiling. Finally, we show the great power of iPcc to remove batch effects in gene expression analysis.

RESULTS

iPcc highlights latent structures between samples with noisy simulated gene expression profiles

We first tested the performance of iPcc on the simulated data sets. For the first simulated data set, which has no irrelevant genes, we observed that iPcc can reflect the true sample relationship faithfully with any order of correlation features, the same as the performance of the original features (Figure 2). After adding 10-fold noisy genes (the second simulated data set), we noted that the sample relationship revealed by the original features became more and more obscure when increasing the standard deviation σ . However, iPcc highlighted the latent sample relationship through interrogating the correlation features. When σ was 0.1, 0.2 and 0.3, iPcc perfectly unraveled the predefined sample relationship. As the order of correlation features increased, the patterns between samples became clearer and clearer (Figure 2). When σ was 0.4 or 0.5, there was no obvious pattern to be observed based on the original features. However, as the order of correlation features increased, the patterns underlying those samples were highlighted (Figure 2), although the highlighted patterns deviated a bit from the predefined sample similarity structure. This deviation was caused by the big σ , which redefined the sample similarity structure.

The potential of iPcc to highlight latent sample relationships beneath the high-dimension gene expression data was further demonstrated by increasing the number of noisy genes. In the third simulated data set, 50-fold noisy genes were added. The original features only revealed weak similarity between samples from the same predefined class at $\sigma = 0.1$ and 0.2. When $\sigma = 0.3, 0.4$ or 0.5, the predefined sample classes could not be reflected by the original features at all (Figure 2). However, as the order of correlation features increased, the sample relationships at $\sigma = 0.1$ and $\sigma = 0.2$ were enlarged, and the sample relationship at $\sigma = 0.3$ was unraveled almost perfectly. Although the predefined sample structure was not precisely reproduced at $\sigma = 0.4$ or 0.5, it was dug out from the vagueness.

When the number of noisy genes was 100-fold the number of informative genes (the fourth simulated data set), the original features could only reveal weak correlations between samples from the same classes at $\sigma = 0.1$, and the sample patterns were blurred at $\sigma = 0.2, 0.3, 0.4$ and 0.5. When $\sigma = 0.3, 0.4$ or 0.5, there were no obvious similar samples observed (Figure 2). However, iPcc reproduced the sample relationship almost completely at $\sigma = 0.1, 0.2$ and 0.3 and unraveled the underlying sample

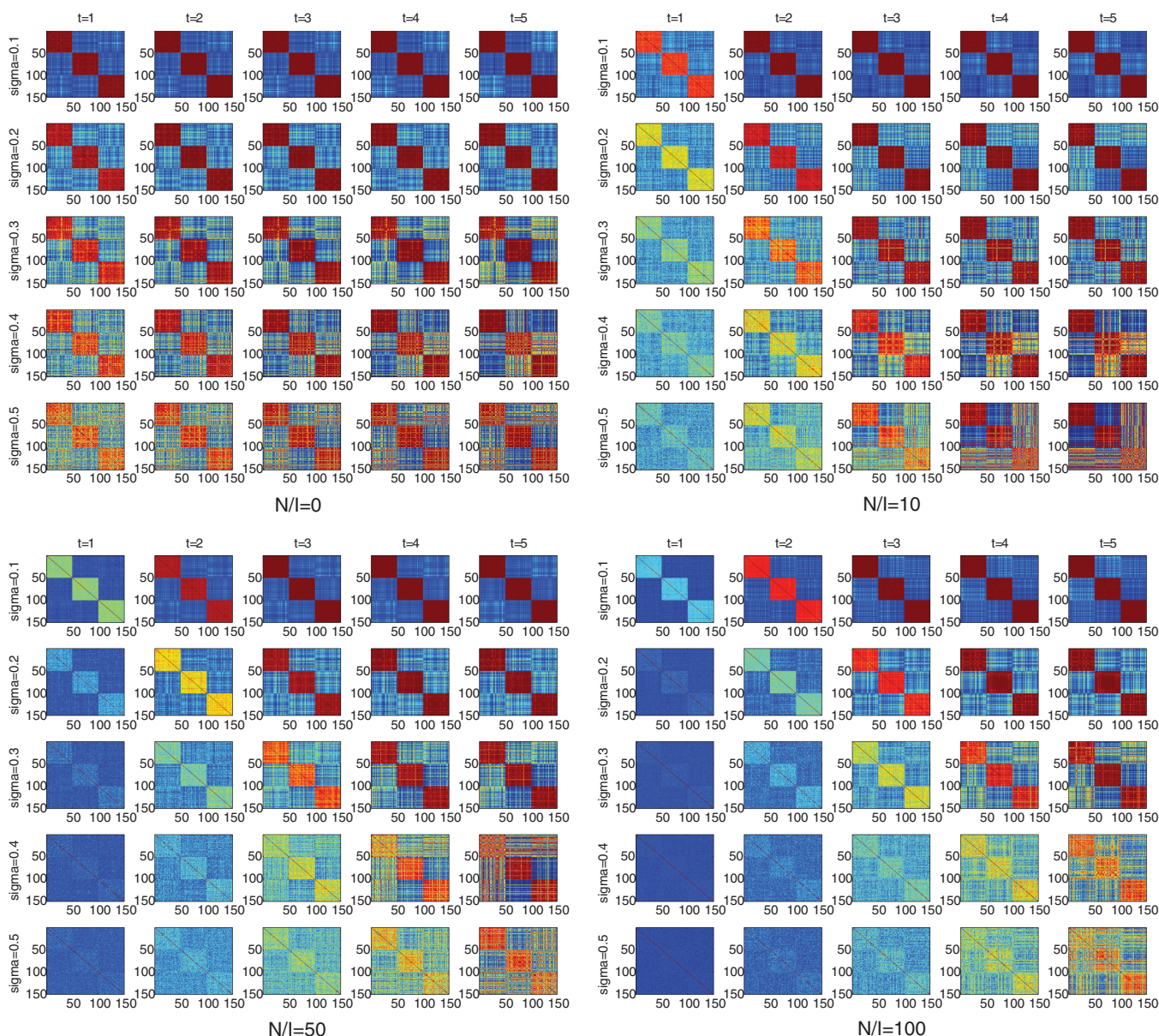


Figure 2. *iPcc* highlights patterns embedded in simulated samples with no, 10-, 50- and 100-fold noisy genes. Three sample classes were simulated, in which samples 1–50 belong to one class, samples 51–100 belong to another class and samples 101–150 belong to the third class. N/I denoted the ratio of the number of noisy genes to the number of informative genes. Sigma is the standard deviation of gene expression during simulation. ' t ' is the order of correlation features.

patterns at $\sigma = 0.4$ and 0.5 with a slight deviation from the predefined sample classes.

***iPcc* improves the accuracy of clustering and classification algorithms on noisy simulated gene expression profiles**

We tested the clustering and classification algorithms *k*-means and the naïve Bayes classifier on the simulated data sets. For all four simulation data sets with 0-, 10-, 50- and 100-fold noisy genes, *k*-means generated more accurate clustering results (average accuracy in 1000 runs) in general based on the correlation features than those based on the original features (Figure 3). When $\sigma = 0.1$, the average accuracy of *k*-means on the original

four simulated data sets was $\sim 93\%$, whereas the highest accuracy of *k*-means on the correlation features of the four data sets reached $\sim 97\%$. When σ increased, the trends still existed and the accuracy increment could even reach 12% (100-fold noisy genes).

We tested the performance of the naïve Bayes classifier by leave-one-out cross-validation evaluations on the four simulated data sets. The naïve Bayes classifier generated more accurate classification results on the correlation features (with order ≥ 2) than those of the original features. When $\sigma = 0.1, 0.2$ or 0.3 , the prediction accuracy approached 1 on both the original features and correlation features for all four simulated data sets. When $\sigma = 0.4$ or 0.5 , the prediction accuracy increased from 86

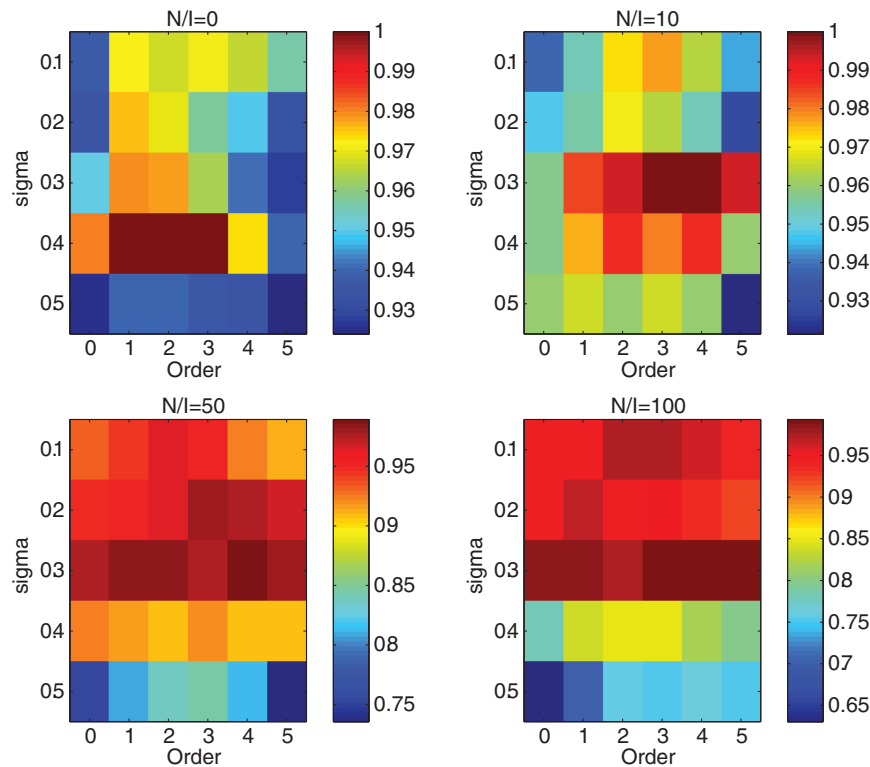


Figure 3. Average clustering accuracy of 1000 k -means ($k = 3$) runs on the simulated data sets with various noisy gene levels and orders of correlation features. Order 0 signifies the original simulated gene expression profiles.

to 90% when 10-fold noisy genes were added, from 88 to 90% when 50-fold noisy genes were added and from 71 to 80% when 100-fold noisy genes were added.

The performance of iPcc on the completely random data sets suggested that the clustering and classification accuracy based on those correlation features generated by iPcc was similar to that on the original features. For each type of statistical distributions, we generated 1000 completely random data sets. For each data set, we calculated the clustering (by k -means) and classification (by the naïve Bayes classifier) accuracy on both the original features and the correlation features. We applied the Student's t -test to those accuracy values and found that no significant differences were observed between the original features and the correlation features for both clustering and classification.

iPcc leverages the power of algorithms for disease class discovery and prediction on a real leukemia data set with two classes

We further tested the performance of iPcc on real gene expression data sets for three different types of diseases: leukemia, prostate cancer and psoriasis. First, we evaluated the sample relationship reflected in the whole gene expression profiles of the leukemia data set (3571 probes in total after preprocessing) by iPcc. The sample similarity matrix revealed by the original 3571 features weakly suggest the existence of two big groups: the AML group (samples 1–24) and the ALL group (samples 25–72), in which sample 25 was wrongly

indicated and is in fact an AML sample (Figure 4). The ALL group was further suggested weakly to be composed of two small subgroups: the B-cell ALL group (samples 25–63) and the T-cell ALL group (samples 64–72). The average intra-class sample similarity (AML–AML and ALL–ALL similarities) was 0.7262, while the average inter-class sample similarity (AML–ALL similarities) was 0.6572.

However, the sample similarity matrix constructed based on the first-order correlation features by iPcc strongly indicated the existence of the AML and ALL groups (Figure 4). In the sample similarity matrix constructed of the original features, the difference between the average intra- and inter-class similarities was 0.0690. However, this value in the similarity matrix constructed by iPcc on the first-order correlation features became 0.5226, with an average intra-class similarity of 0.4274 and an average inter-class similarity of -0.0952 . By increasing the order of the correlation features (Figure 4), iPcc further highlighted the sample similarity from the same disease class and enlarges the sample distinctiveness from different disease classes. When the order of the correlation features was five, the average intra-class sample similarity was 0.8385 and the average inter-class sample similarity was -0.8413 . The difference between the intra- and inter-class sample similarities approached 2.

The great power of iPcc was further demonstrated by its big leverage effect on the accuracy and robustness of clustering algorithms, e.g. k -means. We ran k -means (implemented in Matlab R2011b) 1000 times on the original

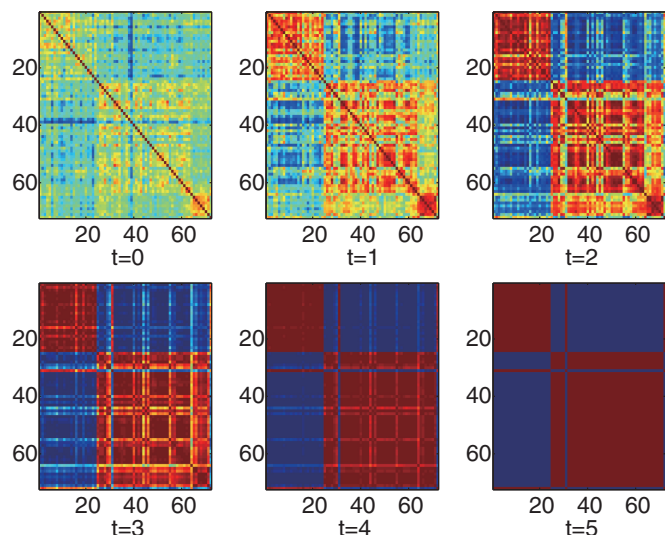


Figure 4. Heatmaps of the leukemia sample relationships on the original, first-, second-, third-, fourth- and fifth-order correlation features. Samples 1–25 are AMLs. Samples 26–63 are B-cell ALLs. And samples 64–72 are T-cell ALLs.

features to cluster the samples into two groups. Compared with the true disease classes, we found that 357 runs produced accuracy of <60%, 70 runs produced accuracy of between 60 and 90% and 573 runs produced accuracy of >90%. However, 1000 runs of k -means on the first-order correlation features constructed by iPcc produced accuracy of >90% (Figure 5). Increasing the order of the correlation features suggested that the accuracy of k -means stabilized ~96%.

Besides the clustering algorithm, iPcc also enhanced the classification algorithms, e.g. the naïve Bayes classifier. We evaluated the prediction accuracy of the naïve Bayes classifier on the original features by the leave-one-out cross-evaluation, which yielded 80% accuracy. The accuracy increased to 96% quickly on the first-order correlation features. On the second-order correlation features, the accuracy even approached 97%. Increasing the order further, the accuracy stabilized at 96% (Figure 6).

iPcc improves the performance of algorithms for disease class discovery and prediction on real prostate cancer data set with multiple classes

Besides the good performance of iPcc on two-class situations, its great power on multi-class gene expression data sets was also demonstrated. We extracted 22 normal samples, 20 metastatic prostate cancer samples and 32 localized prostate cancer samples from the whole prostate cancer data set, selected the top 1000 informative genes by F-test and constructed a three-class gene expression data set. We constructed the sample similarity matrices based on the original features and the first-, second-, third-, fourth- and fifth-order correlation features (Figure 7). We found that the three-class pattern of samples was highlighted with the order of the correlation features increasing. The difference between the average intra- and inter-class sample similarities on the original features was 0.4028 (intra: 0.2890, inter:

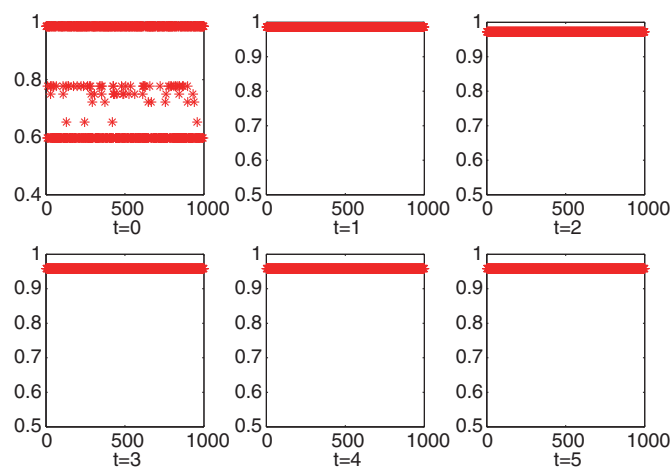


Figure 5. Clustering accuracy of 1000 k -means ($k = 2$) runs on the leukemia data set based on the original, first-, second-, third-, fourth- and fifth-order correlation features.

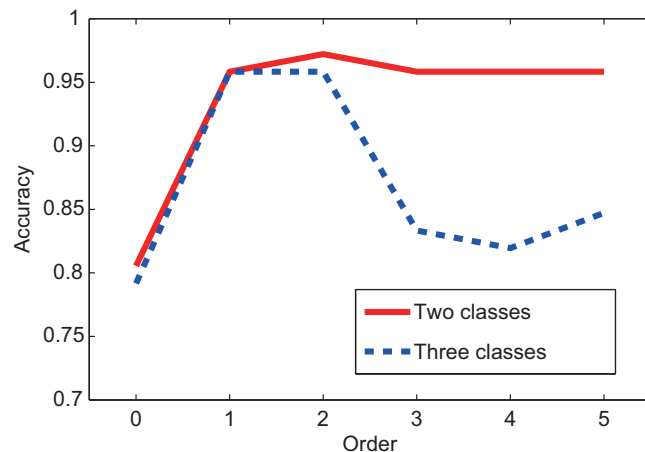


Figure 6. The leave-one-out prediction accuracy by the naïve Bayes classifier on the leukemia data set based on the original, first-, second-, third-, fourth- and fifth-order correlation features.

–0.1138). However, the difference between the first-order correlation features was 0.8518 (intra: 0.5600, inter: –0.2918). The difference on the second-order correlation features was 0.9962 (intra: 0.6520, inter: –0.3442). The difference approached the maximum on the third-order correlation features (1.006, with intra: 0.6592 and inter: –0.3465). Increasing the order further, the difference decreased a bit because the normal samples and the metastatic prostate cancer samples began to show similar correlation features, i.e. three classes begin to converge to two classes.

Furthermore, we examined the impact of iPcc on the performance of k -means clustering. We ran k -means 1000 times on the original features, and found that 105 runs generated results with accuracy of <70%. However, all 1000 runs on the first-order correlation features generated results with accuracy of >90%. Runs of 998 and 1000 produced accuracy of >90% on the second- and third-order correlation features, respectively. On the fourth- and fifth-order correlation features, the accuracy

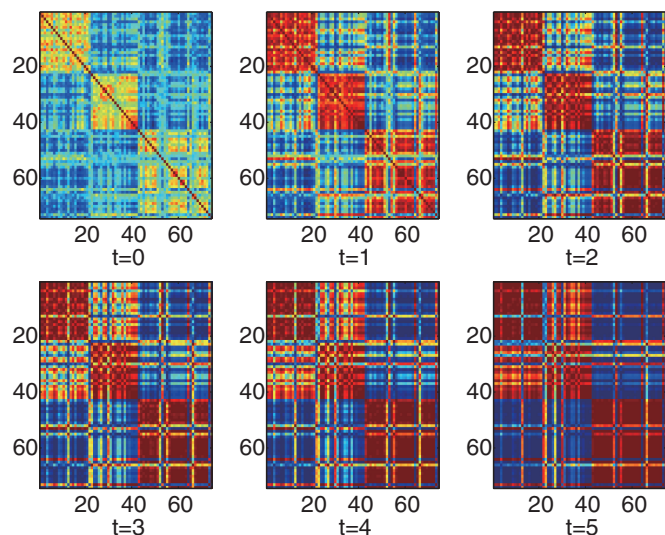


Figure 7. Heatmaps of the prostate cancer sample relationships on the original features, first-, second-, third-, fourth- and fifth-order correlation features. Samples 1–20 are normal samples. Samples 21–42 are metastatic prostate cancer samples, and samples 43–74 are localized prostate cancer samples.

of *k*-means oscillated between 60 and 95% because the three classes began to converge into two classes. The leave-one-out accuracy of the naïve Bayes classifier on the original features was 93%, and the accuracy values on the first-, second-, third-, fourth- and fifth-order correlation features were 93, 90, 90, 90 and 90%, respectively.

In fact, the performance of iPcc in multi-class situations was also demonstrated in the leukemia data set in which the ALL samples can be further divided into B-cell ALLs and T-cell ALLs. The difference between the average intra- (AML–AML, B-cell-ALL–B-cell-ALL and T-cell-ALL–T-cell-ALL) and inter-class (AML–B-cell-ALL, AML–T-cell-ALL and B-cell-ALL–T-cell-ALL) sample similarity values based on the original features was 0.0702. However, the differences on the first-, second-, third-, fourth- and fifth-order correlation features were 0.5002, 0.9755, 1.2217, 1.3316 and 1.3567, respectively. In 1000 *k*-means runs on the original features, 130 runs produced accuracy of >80%, while 247, 114, 276 and 1000 runs produced accuracy of >80% on the first-, second-, third- and fourth-order correlation features. On the fifth-order correlations features, *k*-means generated two clusters although *k* was set as three. This is because B-cell ALLs and T-cell ALLs were merged into one cluster in the high-order correlation features. The prediction accuracy of the naïve Bayes classifier on the original features, the first-, second-, third-, fourth- and fifth-order correlation features for three classes was 79, 96, 96, 83, 82 and 85%, respectively.

iPcc provides a convenient means to deal with batch effects in the real psoriasis data sets

We applied iPcc on the psoriasis data sets to show its performance in situations with batch effects. First, we applied iPcc to examine the sample relationships for

GSE13355 and GSE14905 individually. The results suggested that those psoriasis samples were weakly discriminated from the normal samples based on the whole 54 676 original features for both of the two data sets. On greatly reducing the size of the correlation features, the distinctiveness between psoriasis from normal samples was further underlined and confirmed. Clustering by *k*-means on both the original and correlation features of the GSE13355 data set obtained 99% accuracy. For GSE14905, the clustering accuracy by *k*-means was 98% on the original, the first-, second-, fourth- and fifth-order correlation features and 99% on the third-order correlation features. The leave-one-out prediction accuracy of the naïve Bayes classifier for the GSE13355 data set was 96% on the original features, 99% on the first- and second-order correlation features and 98% on the third-, fourth- and fifth-order correlation features. The leave-one-out prediction accuracy of the naïve Bayes classifier for the GSE14905 data set was 96% on the original, first- and second-order correlation features, and 98% on the third-, fourth- and fifth-order correlation features.

However, we observed that the original features suggested that the sample relationship of the combined data set (concatenating after gene-wise normalization) was dominated by three clusters: the normal samples in GSE13355, the psoriasis samples in GSE13355 and all the samples (both normal and psoriasis) in GSE14905. Correlation features constructed by iPcc confirmed the observation based on the original features, and further suggested that the psoriasis samples in GSE13355 were more similar to those samples in GSE14905 (Figure 8). We clustered the samples by *k*-means on the original features and performed 1000 runs. Five runs obtained accuracy of >90%. One hundred forty-seven runs obtained accuracy of <60%, and 848 runs produced accuracy of between 80 and 90%. On the first-correlation features, the clustering accuracy by *k*-means stabilized at 59%. On the second-, third-, fourth- and fifth-order correlation features, the clustering accuracy by *k*-means stabilized at 81%. The leave-one-out prediction accuracy by the naïve Bayes classifier on the original features was 80%, similar to the clustering accuracy. However, on the first-order correlation features, the leave-one-out prediction accuracy increased to 97%. On the second-order correlation features, the accuracy further increased to 98%. On the third-, fourth- and fifth-order correlation features, the accuracy decreased to 96, 91 and 91%, respectively.

Although the gene-wise normalization step tried to remove batch effects, the above observation suggested that batch effects still influenced the global sample relationship. Here, we show how iPcc removes the batch effects based on correlation features. Because we already knew the experiments were done in two different laboratories, we constructed a sample similarity matrix *S* by setting $s_{ij} = 1$ if the experiments of sample *i* and *j* were done in the same laboratory, otherwise setting $s_{ij} = 0$ if the experiments of sample *i* and *j* were done in different laboratories. We generated the first-order correlation features C_1 and updated C_1 by $C_1 - \lambda \times S$, where $\lambda = 0.3$ in the real computation. Based on the updated

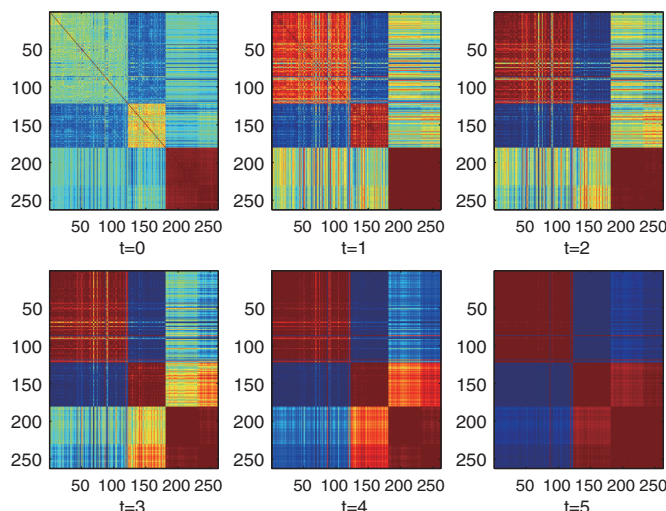


Figure 8. Heatmaps of the psoriasis sample relationships on the original, first-, second-, third-, fourth- and fifth-order correlation features without batch effect removal. Samples 1–122 are normal samples from GSE13355. Samples 123–180 are psoriasis samples from GSE13355. Samples 181–229 are normal samples from GSE14905, and samples 230–262 are psoriasis samples from GSE14905.

C_1 , we further constructed the second-, third-, fourth-, fifth-, sixth- and seventh-order correlation features according to iPcc. The sample relationship revealed by the batch-effect-removed features highlighted the similarity among the normal samples and among those psoriasis samples from different laboratories (Figure 9). One thousand runs of k -means on the third-, fourth-, fifth- and sixth-order updated correlation features obtained stable 86, 94, 97 and 98% accuracy, significantly larger than the clustering accuracy on the original features or the un-updated correlation features. The leave-one-out prediction accuracy by the naïve Bayes classifier also stabilized at 98% on the fourth-, fifth- and sixth-order updated correlation features.

We compared the batch-effect-removal method by iPcc with feature selection methods, which are frequently used to improve the quality of clustering and prediction. We selected the top 1000 informative genes by F-test and evaluated sample relationship, clustering accuracy and prediction accuracy. The average difference between the intra- and inter-class sample similarities was 0.8592 on the original features. One thousand runs of k -means on the original features produced 490 runs with accuracy of 98% and 510 runs with accuracy of 81%. The leave-one-out prediction accuracy of the naïve Bayes classifier on the original features was 98%. Thus, our iPcc-based batch-effect-removal method outperforms, or at least is comparable with, the performance of F-test-based feature selection methods.

By applying iPcc on the 1000 genes selected by F-test, we obtained the intra-class versus inter-class difference values on the first-, second-, third-, fourth- and fifth-order correlation features of 1.2800, 1.3629, 1.4095, 1.4271 and 1.4251, respectively, larger than that on the originally selected genes. One thousand runs of k -means

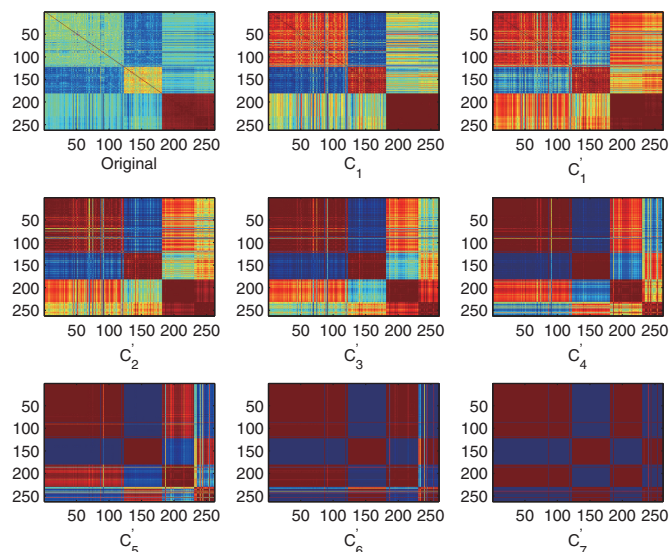


Figure 9. Heatmaps of the psoriasis sample relationships on the original, first-order correlation and batch-effect-removed correlation features. C_1 , the first-order correlation features; C'_1 , the updated first-order correlation features with batch effect removal; C'_2, \dots, C'_7 , the correlation features generated based on C'_1 . Samples 1–122 are normal samples from GSE13355. Samples 123–180 are psoriasis samples from GSE13355. Samples 181–229 are normal samples from GSE14905. Samples 230–262 are psoriasis samples from GSE14905.

on the first-order correlation features yielded 598 runs with an accuracy of 98% and 402 runs with an accuracy of 84%. On the second-order correlation features, the accuracy of 1000 runs of k -means stabilized at 97%. The leave-one-out prediction accuracy of the naïve Bayes classifier on the original features and all the correlation features was 98%. Thus, iPcc can also enhance the power of feature-selection methods.

DISCUSSIONS AND CONCLUSION

With the development of high-throughput technologies, thousands of genes can be measured simultaneously by microarrays or next-generation sequencing technologies. This facilitates comprehensive characterization of the biomedical states of samples but also introduces much irrelevant information. In this study, we proposed a novel feature extraction method, named iPcc, to extract the underlying patterns from noisy data sets through introducing the ‘correlation feature’ concept with iterative Pearson correlation coefficients. Simulations and evaluations on real data sets demonstrate that iPcc greatly improves the disease class discovery and prediction based on the gene expression profiles.

The effectiveness of iPcc may partially originate from reduced dimensionality. In the high-throughput experiments, the number of samples is generally much lower than the number of measured genes, resulting in the ‘curse of dimensionality’. This feature extraction method reduced the dimensionality from thousands of genes to tens of samples. The reduced dimensionality allows rapid and accurate computation of the global optimum of many

clustering and classification algorithms and thus improves accuracy. The effectiveness of iPcc may also originate partially from the introduced ‘correlation features’. Without rephrasing the gene expression levels of samples, correlation features introduced by iPcc used the similarity between samples, which is independent of the precise gene expression levels but is related to the trends of the profiles. These features are similar to clinicians’ clinical experience, which is helpful in diagnosis and prognosis. The effectiveness of iPcc is dependent on the measure of Pearson correlation coefficients. We evaluated those ‘correlation features’ computed by the Spearman correlation coefficients, and found that ‘iScc’ does not have the same effect. We will further investigate the differences between Pearson correlation coefficients and Spearman correlation coefficients in the iterative computation setting. The effectiveness of iPcc also relies on rapid convergence (42,43). For the toy, simulation and real data sets, we observed that those values of correlation features quickly converge to 1 or -1 as the order increases. Although the computation of iPcc is rather simple, we have not found a proper mathematical tool to prove the convergence of iPcc owing to the difficulties posed by the specific features of iPcc, including: (i) the input and output of iPcc are structured matrices; (ii) only a recursion formula is available for iPcc now; (iii) the convergence of iPcc occurs at each element, rather than a common norm of the matrix; (iv) the formula is nonlinear; (v) no constraint exists among the elements. We will continue to seek the mathematical proof of the rapid convergence of iPcc in the future.

Because the computation of iPcc does not require sample class information, it is essentially unsupervised. When confounding factors exist, iPcc can also enlarge the biases embedded within the gene expression data set instead of the subject information, just like the batch effects in the psoriasis data set. Thus, it works best on those data sets with irregular noise. However, if the confounding factors are known, iPcc provides an efficient means to remove the effects of those factors. However, iPcc is not an independent algorithm for disease class discovery or prediction. It provides an effective means to underpin the underlying patterns embedded within the gene expression data sets from the feature extraction perspective. Therefore, it can be used in combination with other clustering, classification, feature selection and feature extraction algorithms, as demonstrated in the results section.

We also observed that multiple sample classes could converge to two classes in the high-order correlation feature space. For example, the B-cell ALLs and T-cell ALLs became undistinguishable in the high-order correlation feature space. This is closely related to the class relationship. Because T-cell ALLs are more similar to B-cell ALLs than to AMLs, iPcc merges T-cell ALLs with B-cell ALLs when the iteration number is adequately large. Because iPcc is unsupervised, it would also enlarge noise in high order of correlation features. Therefore, the optimal iteration number of iPcc is dependent on the specific noise level and the information users needed. Because iPcc generates correlation features, a certain

order of features can also be jointly used with the original and other-order correlation features.

In conclusion, we proposed a novel feature extraction method that is noise-resistant, i.e. iPcc, based on iterative Pearson correlation coefficients for disease class discovery and prediction based on high-throughput gene expression profiles. Evaluations on both simulated and real gene expression data sets suggest that iPcc not only can highlight the patterns underlying high-dimension gene expression files but also can greatly improve the accuracy of disease class discovery and prediction based on clustering and classification algorithms. This feature extraction method is expected to be a useful tool to boost the development of clinical diagnosis and prognosis.

ACKNOWLEDGEMENTS

The authors thank all members of ZHANGroup of AMSS, and the Key laboratory of Systems Biology, CAS for their valuable comments and contributions to discussions.

FUNDING

Funding for open access charge: National Natural Science Foundation of China [31200106, 11131009 and 61171007].

Conflict of interest statement. None declared.

REFERENCES

- Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.
- Beer, D.G., Kardia, S.L.R., Huang, C.C., Giordano, T.J., Levin, A.M., Misek, D.E., Lin, L., Chen, G., Gharib, T.G., Thomas, D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.
- Blaveri, E., Simko, J.P., Korkola, J.E., Brewer, J.L., Baehner, F., Mehta, K., DeVries, S., Koppie, T., Pejavar, S., Carroll, P. *et al.* (2005) Bladder cancer outcome and subtype classification by gene expression. *Clin. Cancer Res.*, **11**, 4044–4055.
- Bresolin, S., Zecca, M., Flotho, C., Trentin, L., Zangrando, A., Sainati, L., Stary, J., de Moerloose, B., Hasle, H., Niemeyer, C.M. *et al.* (2010) Gene expression-based classification as an independent predictor of clinical outcome in Juvenile myelomonocytic leukemia. *J. Clin. Oncol.*, **28**, 1919–1927.
- Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
- Beane, J., Vick, J., Gindin, Y., Anderlind, C., Schembri, F., Liu, G., Wang, S., Levy, S., Massion, P.P., Lenburg, M. *et al.* (2011) RNA-seq identifies novel changes in the airway transcriptome associated with smoking and lung cancer. *Am. J. Respir. Crit. Care Med.*, **183**.
- Beane, J., Vick, J., Schembri, F., Anderlind, C., Gower, A., Campbell, J., Luo, L., Zhang, X.H., Xiao, J., Alekseyev, Y.O. *et al.* (2011) Characterizing the impact of smoking and lung cancer on the airway transcriptome using RNA-Seq. *Cancer Prev. Res.*, **4**, 803–817.

8. Fumagalli,D., Haibe-Kains,B., Michiels,S., Brown,D.N., Gacquer,D., Majaj,S., Salgado,R., Larsimont,D., Detour,V., Piccart,M. *et al.* (2012) Comparison between RNA-Seq and Affymetrix gene expression data. *Cancer Res.*, **72(Suppl.3)**, P3-04-10.
9. Rowley,J.W., Oler,A.J., Tolley,N.D., Hunter,B.N., Low,E.N., Nix,D.A., Yost,C.C., Zimmerman,G.A. and Weyrich,A.S. (2011) Genome-wide RNA-seq analysis of human and mouse platelet transcriptomes. *Blood*, **118**, e101-e111.
10. Nanni,L., Brahmam,S. and Lumini,A. (2012) Combining multiple approaches for gene microarray classification. *Bioinformatics*, **28**, 1151-1157.
11. Sloutsky,R., Jimenez,N., Swamidass,S.J. and Naegle,K.M. (2012) Accounting for noise when clustering biological data. *Brief. Bioinform.*, October 14 (doi: 10.1093/bib/bbs057; epub ahead of print).
12. Brown,M.P.S., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262-267.
13. Buturovic,L.J. (2006) PCP: a program for supervised classification of gene expression profiles. *Bioinformatics*, **22**, 245-247.
14. Furey,T.S., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906-914.
15. Nannapaneni,P., Hertwig,F., Depke,M., Hecker,M., Mader,U., Volker,U., Steil,L. and van Hijum,S.A. (2012) Defining the structure of the general stress regulon of *Bacillus subtilis* using targeted microarray analysis and random forest classification. *Microbiology*, **158**, 696-707.
16. Pang,H., Lin,A., Holford,M., Enerson,B.E., Lu,B., Lawton,M.P., Floyd,E. and Zhao,H. (2006) Pathway analysis using random forests classification and regression. *Bioinformatics*, **22**, 2028-2036.
17. Riddick,G., Song,H., Ahn,S., Walling,J., Borges-Rivera,D., Zhang,W. and Fine,H.A. (2011) Predicting in vitro drug sensitivity using Random Forests. *Bioinformatics*, **27**, 220-224.
18. Li,J., Liu,H., Downing,J.R., Yeoh,A.E. and Wong,L. (2003) Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. *Bioinformatics*, **19**, 71-78.
19. Tan,A.C., Naiman,D.Q., Xu,L., Winslow,R.L. and Geman,D. (2005) Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, **21**, 3896-3904.
20. Tobler,J.B., Molla,M.N., Nuwaysir,E.F., Green,R.D. and Shavlik,J.W. (2002) Evaluating machine learning approaches for aiding probe selection for gene-expression arrays. *Bioinformatics*, **18**, S164-S171.
21. Aronow,B.J., Richardson,B.D. and Handwerger,S. (2001) Microarray analysis of trophoblast differentiation: gene expression reprogramming in key gene function categories. *Physiol. Genomics*, **6**, 105-116.
22. Goh,X.Y., Rees,J.R.E., Paterson,A.L., Chin,S.F., Marioni,J.C., Save,V., O'Donovan,M., Eijk,P.P., Alderson,D., Ylstra,B. *et al.* (2011) Integrative analysis of array-comparative genomic hybridisation and matched gene expression profiling data reveals novel genes with prognostic significance in oesophageal adenocarcinoma. *Gut*, **60**, 1317-1326.
23. Kim,D.W., Lee,K.H. and Lee,D. (2005) Detecting clusters of different geometrical shapes in microarray gene expression data. *Bioinformatics*, **21**, 1927-1934.
24. Osamor,V.C., Adebisi,E.F., Oyelade,J.O. and Doumbia,S. (2012) Reducing the Time Requirement of k-Means Algorithm. *PLoS One*, **7**.
25. Ren,X., Wang,Y., Chen,L., Zhang,X.S. and Jin,Q. (2013) ellipsoidFN: a tool for identifying a heterogeneous set of cancer biomarkers based on gene expressions. *Nucleic Acids Res.*, **41**, e53.
26. Ai-Jun,Y. and Xin-Yuan,S. (2010) Bayesian variable selection for disease classification using gene expression data. *Bioinformatics*, **26**, 215-222.
27. Lee,K.E., Sha,N., Dougherty,E.R., Vannucci,M. and Mallick,B.K. (2003) Gene selection: a Bayesian variable selection approach. *Bioinformatics*, **19**, 90-97.
28. Robbins,K.R., Zhang,W., Bertrand,J.K. and Rekaya,R. (2007) The ant colony algorithm for feature selection in high-dimension gene expression data for disease classification. *Math. Med. Biol.*, **24**, 413-426.
29. Szymczak,S., Nuzzo,A., Fuchsberger,C., Schwarz,D.F., Ziegler,A., Bellazzi,R. and Igl,B.W. (2007) Genetic association studies for gene expressions: permutation-based mutual information in a comparison with standard ANOVA and as a novel approach for feature selection. *BMC Proc.*, **1(Suppl.1)**, S9.
30. Zhu,S., Wang,D., Yu,K., Li,T. and Gong,Y. (2010) Feature selection for gene expression using model-based entropy. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **7**, 25-36.
31. Schachtner,R., Lutter,D., Knollmuller,P., Tome,A.M., Theis,F.J., Schmitz,G., Stetter,M., Vilda,P.G. and Lang,E.W. (2008) Knowledge-based gene expression classification via matrix factorization. *Bioinformatics*, **24**, 1688-1697.
32. Shen,R., Ghosh,D., Chinnaiyan,A. and Meng,Z. (2006) Eigengene-based linear discriminant model for tumor classification using gene expression microarray data. *Bioinformatics*, **22**, 2635-2642.
33. Tamayo,P., Scanfeld,D., Ebert,B.L., Gillette,M.A., Roberts,C.W. and Mesirov,J.P. (2007) Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *PNAS*, **104**, 5959-5964.
34. Tan,Y., Shi,L., Tong,W. and Wang,C. (2005) Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data. *Nucleic Acids Res.*, **33**, 56-65.
35. Rodgers,J. and Nicewander,A. (1988) Thirteen ways to look at the correlation coefficient. *Am. Stat.*, **42**, 59-66.
36. Ren,X., Wang,Y., Wang,J. and Zhang,X.S. (2012) A unified computational model for revealing and predicting subtle subtypes of cancers. *BMC Bioinformatics*, **13**, 70.
37. Tomlins,S.A., Mehra,R., Rhodes,D.R., Cao,X., Wang,L., Dhanasekaran,S.M., Kalyana-Sundaram,S., Wei,J.T., Rubin,M.A., Pienta,K.J. *et al.* (2007) Integrative molecular concept modeling of prostate cancer progression. *Nat. Genet.*, **39**, 41-51.
38. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. and Edgar,R. (2007) NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res.*, **35**, D760-D765.
39. Nair,R.P., Duffin,K.C., Helms,C., Ding,J., Stuart,P.E., Goldgar,D., Gudjonsson,J.E., Li,Y., Tejasvi,T., Feng,B.J. *et al.* (2009) Genome-wide scan reveals association of psoriasis with IL-23 and NF- κ B pathways. *Nat. Genetics*, **41**, 199-204.
40. Swindell,W.R., Johnston,A., Carbajal,S., Han,G., Wohn,C., Lu,J., Xing,X., Nair,R.P., Voorhees,J.J., Elder,J.T. *et al.* (2011) Genome-wide expression profiling of five mouse models identifies similarities and differences with human psoriasis. *PLoS One*, **6**, e18266.
41. Hastie,T., Tibshirani,R. and Friedman,J. (2003) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.
42. McQuitty,L.L. (1968) Multiple clusters, types, and dimensions from iterative intercolumnar correlational analysis. *Multivariate Behav. Res.*, **3**, 465-477.
43. Breiger,R.L., Boorman,S.A. and Arabie,P. (1975) An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *J. Math. Psychol.*, **12**, 328-383.