DRUG DISCOVERY TODAY TARGETS

Determination of tumour marker genes from gene expression data

Miroslava Cuperlovic-Culf, Nabil Belacel and Rodney J. Ouellette

Cancer classification has traditionally been based on the morphological study of tumours. However, tumours with similar histological appearances can exhibit different responses to therapy, indicating differences in tumour characteristics on the molecular level. Thus, development of a novel, reliable and precise method for classification of tumours is essential for more successful diagnosis and treatment. The high-throughput gene expression data obtained using microarray technology are currently being investigated for diagnostic applications. However, these large datasets introduce a range of challenges, making data analysis a major part of every experiment for any application, including cancer classification and diagnosis. One of the major concerns in the application of microarrays to tumour diagnostics is the fact that the expression levels of many genes are not measurably affected by carcinogenic changes in the cells. Thus, a crucial step in the application of microarrays to cancer diagnostics is the selection of diagnostic marker genes from the gene expression profiles. These molecular markers give valuable additional information for tumour diagnosis, prognosis and therapy development.

Cancer is a genetic disease developed through the accumulation of abnormalities and aberrations in gene expression. Several different tumour-specific mutations, DNA amplifications and translocations lead to the development of different cancers with diverse clinical behaviour in terms of both therapy response and disease progression. Thus, it is of fundamental importance to precisely and accurately assign a given tissue sample to a diagnostic category. Currently, cancer diagnoses used to determine prognoses and guide therapy decisions are based on morphological features of the tumour, which are sometimes complemented by single-gene or single-protein assays. Examples include the prostate-specific antigen for prostate cancer diagnosis, CA 125 for ovarian cancer diagnosis and the carcinoembryonic antigen for colorectal cancer. However, tumours with similar histopathological appearances can follow significantly different clinical courses and possibly respond differently to therapy [1].

Molecular tumour classification on the basis of genomics experiments offers hope for a more individualized and more accurate diagnosis, prognosis and determination of treatment options [1,2]. The histological origin of primary tumours, their metastatic potential and optimal treatment might be discernable by gene expression analysis of tumour biopsies. As tumours are caused by genetic alterations, detailed gene expression data from genomic measurements are expected to be sufficient for the development of novel tumour classifications based on molecular characteristics. This novel classification could lead to a more complete understanding of molecular variations among tumours and, hence, better diagnoses and treatment strategies for the disease [3].

Miroslava Cuperlovic-Culf*
Rodney J. Ouellette
Beauséjour Medical Research Institute
Nabil Belacel
National Research Council, Canada,
Institute for Information, Technology-e-Health
*e-mail:
miroslavac@health.nb.ca

1359-6446/04/$ – see front matter ©2005 Elsevier Ltd. All rights reserved. PII: S1359-6446(04)03393-3

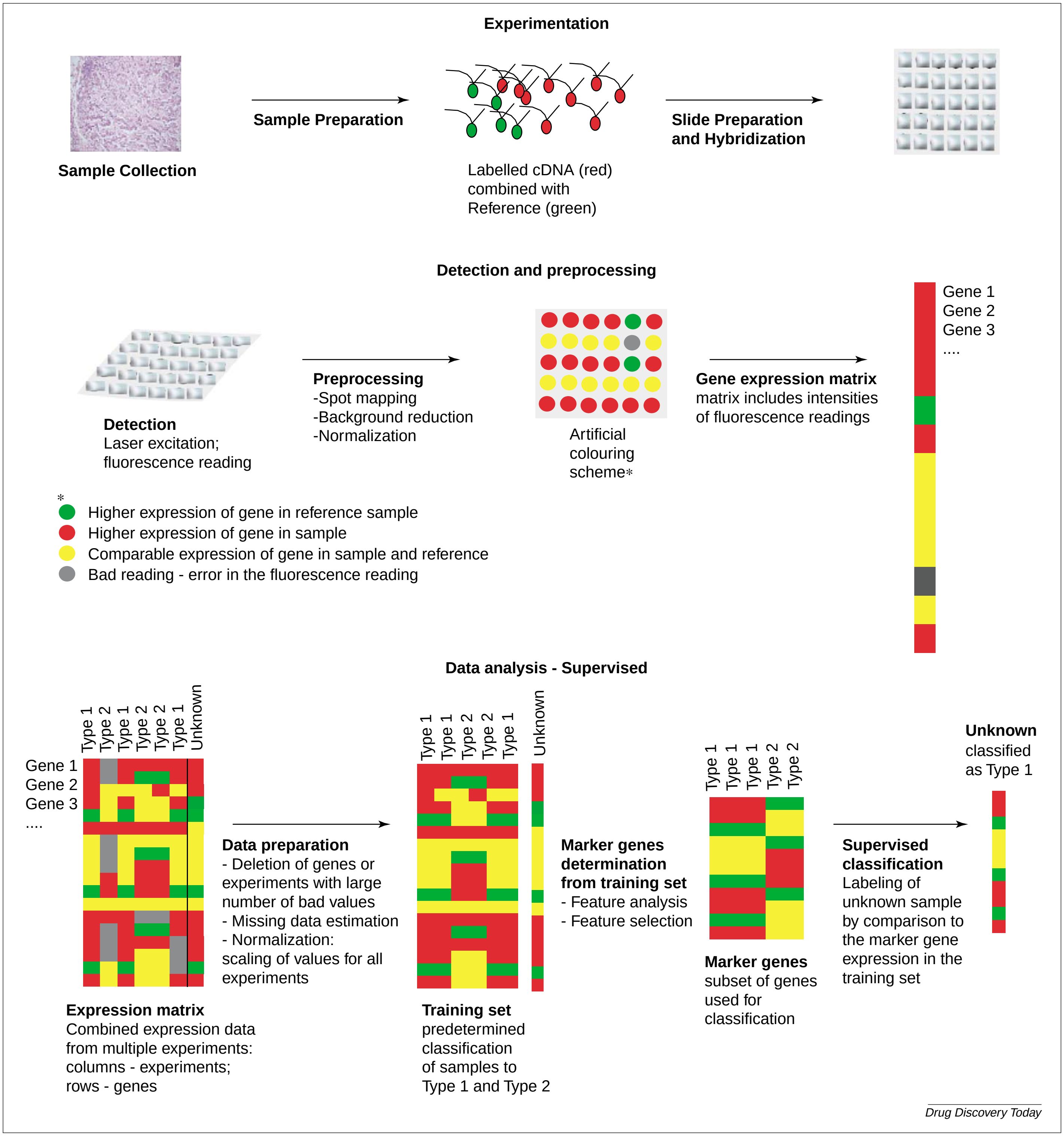

FIGURE 1

Schematic representation of microarray technology aimed at sample classification. The steps shown lead from sample collection (either tissues or cell culture samples), through RNA extraction, labelling and hybridization to DNA arrays, to detection and pre-processing (intensity determination). Data analysis includes the assembling of a gene expression matrix (database) from multiple experiments and the preparation of the data in the matrix for analysis. The missing values for some genes in some experiments (caused by errors in the array or hybridization) can be estimated from values of other genes. Once the gene expression matrix is prepared, a subset of classified samples is assembled into a training set that is used for the determination of marker genes and for classification. The gene expressions of marker genes from the training set are compared with their expressions in the unknown sample set. The unknown sample is then assigned to a class (type) with the most similar expression pattern. Although this figure represents spotted DNA microarray technology, the presented steps are comparable for any DNA microarray platform. Other array methods such as comparative genomic hybridization, cell arrays, protein arrays and glycol-chips follow the same logic.

Microarrays allow the simultaneous study of gene expression of all or a large portion of a genome of interest.

In a typical microarray experiment, total RNA or mRNA is extracted from a sample (tissues or cells), labelled by

BOX 1

Terminology used in microarray data analysis

Features in terms of cancer DNA microarray experiments correspond to the expression measurements of different genes;

Classes correspond to different tumour types (e.g. ER^+ versus ER^- breast tumours; malignant versus benign tumours)

Unsupervised analysis (also known as cluster analysis, class discovery, unsupervised pattern recognition) includes methods used for the grouping of either samples or genes according to gene expression measurements. Unsupervised methods are used when no preliminary information is available about sample groups. Unsupervised analysis involves estimating the number of classes (groups or clusters) and assigning an object to these classes. This type of analysis is ideal for the discovery of novel classes.

Supervised analysis (also known as classification, discriminant analysis, class prediction and supervised pattern recognition) defines methods for sample grouping or classification. In supervised analysis, a subset of data (training set) includes samples that were previously classified by an external supervisor. The task is then to understand the basis for the classification from the training set and to build a classifier that will be used for the classification of unlabelled object. In the case of tumour classification, samples in the training set would be diagnosed by a pathologist before microarray analysis.

Feature analysis is defined as calculating (i.e. determining quantitatively) the difference between gene expression for a gene in different groups of samples.

Feature selection is used for the determination of genes with the most significant difference in gene expression between groups of samples. Feature selection is based on the results obtained in feature analysis. In terms of tumour diagnosis, feature selection is the identification of marker genes that characterize different tumour classes or have good predictive power for an outcome of interest.

reverse transcription typically using fluorescently-labelled nucleotides and hybridized to a previously prepared array of synthetic oligonucleotides or cDNAs, with each spot on the array being complementary to one gene. After hybridization and washing, the arrays are scanned using one or multiple fluorescence frequencies and the fluorescence signal intensities at each spot are determined by image-analysis software. The obtained expression measurements for all genes in all experiments are then organized into an expression matrix (Figure 1). Prior to the application of data, measurements for spots of poor quality (either owing to errors in printing or hybridization), are removed from the expression measurements in the matrix. When a range of experiments is performed using the same array layout and many samples, missing data for a gene in one experiment can be estimated from expression measurements determined for other genes with a similar profile over other measured samples. Absolute intensities determined in an experiment depend upon a range of factors other than gene expression (e.g. scanner setting and sample volumes). Therefore, data from a microarray experiment have to be normalized; that is, all the experimental measurements have to be scaled to have an equal mean or median, or have equal values for certain genes (housekeeping genes or genes added in known quantities as controls). This normalization makes comparison between experiments possible. Data from different experiments can be used for classification only after these steps have been executed.

Since the advent of DNA array technology, researchers have been exploring the possibilities of using expression array analysis as a quantitative (based on numeric expression values), non-subjective diagnostic tool for tumours. Recent data from several groups revealed novel fundamental and reproducible principles in disease classification using gene expression profiling [4]. Gene expression profiling using DNA arrays has great potential as a systematic approach for discovering new, more accurate stratifications of tumours or for assigning tumour samples to predetermined classes [2]. Microarray gene expression measurements have already been used in the classification of many different cancers including acute leukemia [5], breast cancer [6,7], melanoma [8], colon cancer [9] and synovial sarcoma [10].

Currently, attempts are underway to optimize and standardize each step in the microarray procedure for clinical use [11]. Once collected and properly optimized, microarray data from clinical samples can be used for sample analysis using either unsupervised or supervised approaches.

The unsupervised analysis (Box 1) of clinical samples enables identification of groups of statistically related samples or genes based on gene expression profiles. Unsupervised methods are particularly useful when data are analyzed in an exploratory fashion or for the determination of novel sample types [12,13].

The supervised methods (Box 1) provide more powerful tools for the classification of a sample [12]. In the supervised analysis, the initial step is to assemble a subset of samples, called the training or learning set, which were previously diagnosed by an external supervisor (such as pathologist). Following the microarray experiment for these samples as described in Figure 1, a search is performed of all gene expression measurements from all samples in the training set to determine a list of marker genes that are most distinct in their expression between groups with different labelling (*i.e.* diagnosis). Expressions of these genes in an unknown sample are then used for the diagnosis by comparing their expression levels to those of marker genes in groups of known sample types in the training set. A major question in microarray studies is how to describe genes associated with specific pathological states or clinical parameters (feature analysis) and how to select the most significant genes (feature selection). The subset of genes selected for their large difference in expression between sample types, referred to as marker genes or clinical markers, can form the basis for diagnostic tests, particularly if they can be reliably assayed. In addition, significant changes in the expression of clinical marker genes in tumours are possible targets for drug development.

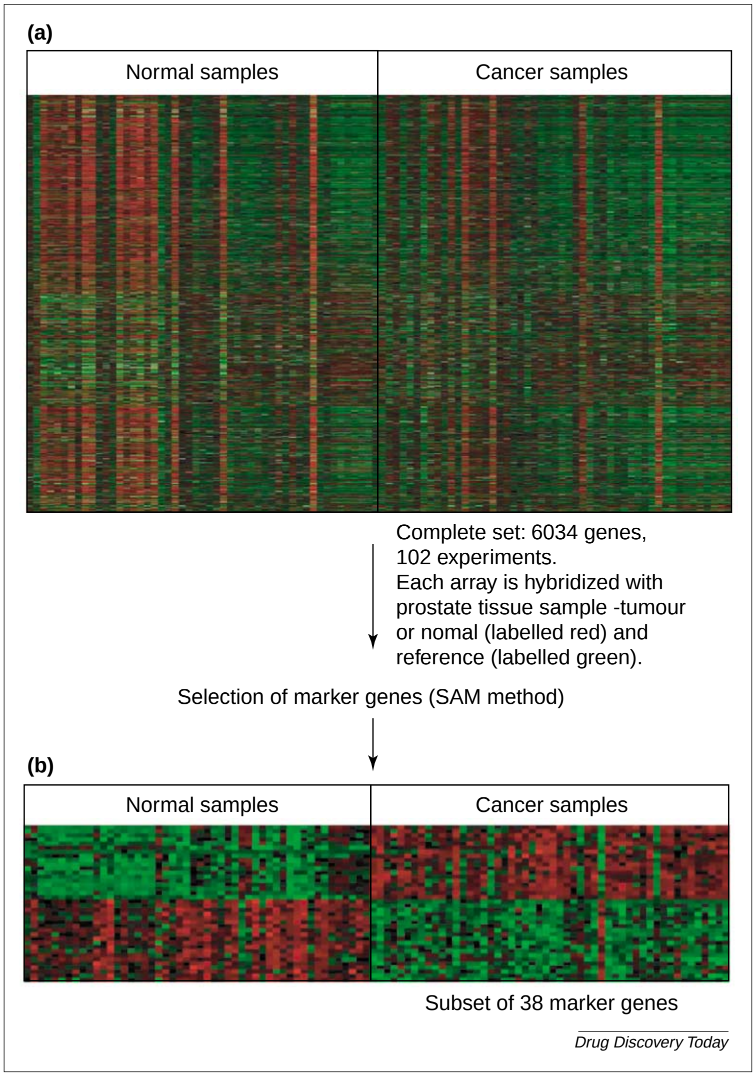
(a)
Normal samples
Cancer samples
Complete set: 6034 genes, 102 experiments. Each array is hybridized with prostate tissue sample -tumour or nomal (labelled red) and reference (labelled green).
Selection of marker genes (SAM method)
(b)
Normal samples
Cancer samples
Subset of 38 marker genes
Drug Discovery Today

FIGURE 2

An example of the importance of feature analysis and selection in sample classification. (a) shows the complete set of gene expression measurements for 50 normal and 52 tumour prostate samples following gene expression matrix preparation (including missing data simulation and normalization) as obtained by Singh *et al.* [15]. The experiments are performed in a two fluorescence label format (procedure explained in Figure 1) with samples always labelled with the same dye (red). All experiments also included a common reference labelled with another fluorescence dye (green). The samples are grouped as normal and cancer tissue samples but the sample type can not be inferred from this complete gene expression matrix. Determination of marker genes for these to tissue types can be performed using various statistical methods. **(b)** shows the subset of marker genes determined using SAM method. The difference between the two sample types becomes even visually transparent when using the gene expression data for the subset of marker genes.

In the context of microarray applications, feature analysis can be defined as the method used to assign a numerical value describing each gene's difference in expression measurements between groups of sample types. In the same context, feature selection defines methods for the determination and selection of a group of genes with the most significant differences in expression measurements between sample types, based on the result of feature analysis. As the number of marker genes for a particular classification project in microarray studies might be very small relative to the total number of genes represented on the array, both feature analysis and selection procedures are essential. Indeed, the accuracy of tumour classification from gene expression has been shown to depend more upon the selection of diagnostic genes than on the classification procedure [14]. Classification is by no means a new subject in the statistical literature, but the large and complex multivariate datasets generated by microarray experiments raise new methodological and computational challenges. A large number of techniques have been developed or altered for microarray applications. This review focuses on feature analysis and selection methods previously used in microarray classifications for cancer diagnosis and prognosis.

An example of the necessity for feature analysis and selection in the context of tumour classification is given in Figure 2. The dataset contributed by Singh and coworkers [15] includes gene expression measurements for 52 prostate tumours and 50 non-tumour prostate samples measured relative to a common reference. Pre-processing and filtering resulted in the set of 6034 genes (Figure 2a). However, this set does not show a clear difference between normal and cancer tissues because of the large number of genes with expression levels determined by factors other then sample type. However, once a subset of marker genes is identified using feature analysis and selection methods, the difference between the two sample types becomes apparent (Figure 2b).

Methods used for the selection of diagnostic genes in cancer

Among the thousands of genes featured on a microarray chip, only a small number are important for the distinction of tumour classes. Therefore, in the majority of tumour classification applications, feature analysis and selection are performed as the initial step in the analysis. In this way, irrelevant attributes (i.e. genes with expression levels not significantly different between groups of different tumour types) are excluded from classification, thereby reducing both the noise of the dataset as well as the time needed to perform the classification.

Microarray datasets are often created from groups of tumour samples with unknown or undetermined types, and the classification of these samples is performed using unsupervised methods. Unsupervised tools used for sample classification can also be used for gene selection and those most frequently used include various clustering and biclustering approaches [16,17] as well as various projection methods (for example, see [8,18,19]).

Although unsupervised methods can be used for both classification and gene selection, statistical methods applied in the supervised mode provide a much more powerful tool for gene analysis and selection [8].

Some of the most widely used methods (with references showing their application) and software tools are listed in Table 1. Some of the software packages shown (rows b and c) include many other data analysis methods outside the scope of this review.

TABLE 1

Data analysis methods used in feature analysis and selection.

A. Methods and corresponding software tools	**Reference**
Regularized *t* test Cyber T (http://visitor.ics.uci.edu/genex/cybert/)	[48]
SAM (http://www-stat.Stanford.edu/~tibs/SAM)	[31]
PAM (http://www-stat.stanford.edu/~tibs/PAM/)	
REF (fpn.mit.edu/SvmFu/)	[14]
ANOVA (http://www.jax.org/staff/churchill/labsite/software/anova/index.html)	[37–39]
B. Non-commercial software packages for data analysis	
Bioconductor (http://www.bioconductor.org) - based on the R statistical software	
BRB ArrayTools - add-on to Microsoft Excel	[49]
TMeV - range of tools for data analysis	[50]
RankGene -range of tools for feature selection and ranking	[46]
C. Commercial data analysis software packages	
Resolver (http://www.rosettabio.com/products/resolver/default.htm)	
GeneSpring (http://www.silicongenetics.com/cgi/SiG.cgi/index.smf)	
Spotfire (http://www.spotfire.com/)	
Acuity (http://www.axon.com/gn_Acuity.html)	

Group A represents software tools developed to accompany a particular method. Groups B and C include software packages, both non-commercial (B) and commercial (C), which include a range of tools for microarray data analysis including many different tools for feature analysis and selection. The list is not exhaustive; rather it is included so as to display some popular software options to the readers.

Statistical methods for gene analysis and selection

Feature analysis and feature selection are the first steps performed in supervised analysis. Some of the most frequently used algorithms for feature analysis and selection are presented below. All mathematical definitions as well as statistical requirements for the use of these algorithms are given in Box 2.

Fold change method

The simplest, non-statistical test method used for the selection of differentially expressed genes is the fold change method. In this method, the ratios between expression level logarithms in two conditions are evaluated. All genes with a ratio of expression level higher than an arbitrary cut-off value are considered to be differentially expressed [20]. The fold change method in its original form can be strongly biased by an inappropriate normalization. This problem has been addressed by the development of intensity-specific thresholds [21]. However, as this simple method is not a statistical test, it has no associated value indicating a level of confidence in the designation of genes as being differentially expressed.

t-statistics and variations

The introduction of various statistical methods relating a gene's expression level to the covariates or responses, and ranking genes accordingly, allows for a more accurate feature analysis and for the most exact feature selection [5].

The *t* test is one of the simplest statistics-based methods used in microarray analysis, both for estimating the accuracy of results from replicated experiments and for the detection of differentially expressed genes.

For the feature analysis of systems with two sample types, the traditional two sample (paired) *t* test is a standard, straightforward and popular method [22]. In most tumour classification studies, however, the samples are independent (i.e. taken from different patients rather then from the same patient at different times), which makes the paired *t* test inappropriate. The two-way *t* test for independent samples (Student *t* test) allows for the determination and selection of an expression pattern that has a maximal difference in the mean level of expression between two groups of independent sample types with a minimal variation of expression within each group. Therefore, the two-way *t* test has been used frequently for the determination of differentially expressed genes in microarray feature analysis and selection [23–25]. The difference in gene expression between sample types is expressed as the *P* value which shows the chance that random sampling would result in the observed difference. The Student *t* test determines the significance of the difference between the means of two independent samples and it is a good choice when i) the two samples are independently and randomly drawn from the source populations; ii) the measurements for both samples have an equal interval; and iii) the source population(s) can be reasonably supposed to have a normal distribution.

Both Student and paired *t* tests assume that features within each group have similar variances. This is rarely seen in microarray experiments. In addition, sample sizes are usually small in comparison to the number of features (genes) [26]. Thus, the variation to the *t* test methods developed for independent samples with unequal variances would be much more appropriate. An example of such test is the Welch (Satterthwaite's) *t* test [26]. This test is based

BOX 2

a. Mathematical definitions of described statistical methods for feature analysis and selection

Dataset $I=[x_{ij}];i-genes;j\in(1,M)samples;\quad U=[x_{ij}];i-genes;j\in(1,N)samples$

Student *t* test $t(i)=\frac{\mu_I(i)-\mu_U(i)}{s(i)(1/N+1/M)}$; pooled variance: $s(i)=\sqrt{\frac{(M-1)s_I^2+(N-1)s_U^2}{M+N-2}}$

SAM $t(i)=\frac{\mu_I(i)-\mu_U(i)}{s(i)-s_0}$; $s(i)=\sqrt{a[\sum(x_j(i)-\mu_I(i))^2+\sum(x_j(i)-\mu_U(i))^2]}$

Welch *t* test $t(i)=\frac{\mu_I(i)-\mu_U(i)}{\sqrt{s_I^2/N+s_U^2/M}}$

SNR $P(i)=\frac{\mu_I(i)-\mu_U(i)}{\sigma_I(i)+\sigma_U(i)}$

ANOVA $\log x_{ijkg}=\mu+$ effects of array, dye, variation and gene

Pearson Correlation Coef. $r(i)=\frac{N\sum x_{iI}x_{iU}-\sum x_{iI}\sum x_{iU}}{\sqrt{(N\sum x_{iI}^2-(\sum x_{iI})^2(N\sum x_{iU}^2-(\sum x_{iU})^2}}$

b. Data-set properties required for application of presented statistical tests

Required properties	Student *t* test	Welch *t* test	SAM	SNR	Wilcoxon	One-way ANOVA	Pearson correlation coefficient
Error independent	Yes	Yes	Yes	Yes	Yes	Yes	Irrelevant
Data paired	No	No	No	No	No	No	No (same number of data points)
Identical distribution	Yes	Yes	Yes	Yes	Irrelevant	Yes	Yes
Gaussian distribution	Yes	Yes	Yes	Yes	Irrelevant	Yes	Yes
Symmetrical around median	Yes	Yes	Yes	Yes	Yes	Yes	Irrelevant
Equal variance	Yes	No	Yes	No	Irrelevant	Yes	Yes

on the difference between sample means and assumes a normal distribution of the samples but allows for different variances of samples. Many other variations of the two-sample *t* test have been used, the choice being dependent upon the sample size and variance of the two sample types.

Signal to noise ratio test

The signal to noise ratio (SNR) test identifies expression patterns with a maximal difference in the mean level of expression between two groups and minimal variation of expression within each group, while not assuming the equality of standard deviations (variances). The SNR combined with different feature selection methods has been the method of choice in most classification studies performed at the Whitehead Institute, MIT [5,27,28], as well as by several other groups [10].

In this procedure, genes are first ranked according to their expression levels using SNR test statistics (Box 2). The top ranking genes can then be selected in several ways. One method was successfully applied to multiclass cancer diagnosis [29]. In this method, SNR ratios are recalculated for each permutation of sample labels (tumour type); that is, each sample is shifted from its original class to the other class group and the SNR is recalculated for these new groups. A gene was then considered a statistically significant marker if the observed SNR for the original set exceeds the permuted SNR in at least 99% of permutations.

Another method for the selection of marker genes from SNR data implements neighbourhood analysis; it was implemented by the group at the Whitehead Institute initially for the classification of leukemia data [5]. The marker genes selected by SNR test should be the ones that best resemble a binary expression profile (i.e. an idealised expression pattern in which the expression level is uniformly high in one class type and low in the other). The SNR ratio measures how well the expression profile of a real gene approximates the ideal marker gene profile. Thus, in this method, the genes with the highest absolute values of SNR are chosen to build binary classifiers [30].

Significance analysis of microarray

The two-sample type significance analysis of microarray (SAM) procedure identifies genes with statistically significant changes in expression by assimilating a set of gene-specific *t* test-like calculations. Each gene is assigned a score on the basis of its change in gene expression relative to the standard deviation of repeated measurements for that gene [31]. The basic statistical procedure used in SAM is similar to a *t* test. SAM determines the difference between means, in units of standard deviations. However, a weighting term is added to the pooled standard deviation in SAM to prevent overestimating the significance of genes with a small variance (Box 2). To find significant changes in gene expression, genes were first ranked by the magnitude of the relative difference t(i) (Box 1). Then, similarly as in the SNR procedure, a relative difference is recalculated for each permutation of sample labels (tumour type) giving *p*-values (equal to number of samples for two-sample type sets of $t_p(i)$). The expected relative difference $t_E(i)$ is calculated as the average of all $t_p(i)$ values — SAM values for sets with one sample label permuted at the time. Genes with significant changes in expression are finally determined from the scatter plot of $t(i)$ versus $t_E(i)$ as the genes with values displaced from the linear function.

Wilcoxon rank sum test

The previously described tests are termed parametric statistical methods; that is, they assume that data follows a particular distribution of values (such as a normal distribution). If no assumption can be made about the shape of the population distribution, non-parametric statistical methods must be used. The Wilcoxon rank sum test is a non-parametric analogue to the *t* test for two independent samples. It is also used for the determination of equality of the means of two non-normal samples. The test is based on the rank of the individual data within the gene data ordered according to the gene expression measurements rather than actual expression values. As it operates on rank-transformed data, it is a robust choice for microarray systems, which are often non-normal and contain outliers [32]. In this test, the ranking of data is performed first by giving the highest rank (equal to the number of features) to the feature with the highest value and rank of one to the feature with lowest value. The ranks for remaining features are assigned using the rank-sum test as explained in details in Biostatistics textbooks [22].

The Wilcoxon method, as well as related non-parametric tools, have been used in several cancer classification studies [32–34].

Multi-feature methods

All of the previously described algorithms make an implicit orthogonality assumption; that is, they disregard the correlation between features (genes). Guyon *et al.* [14] have proposed the application of a Recursive Feature Elimination (RFE) method for feature selection by feature (gene) removal. RFE is an example of a backward feature selection method where, rather then choosing a subset of significant genes, the non-significant genes are chosen and eliminated, usually one gene at a time in the recursive (i.e. cyclic) fashion. The RFE process starts by performing classification using all available features. Then, features are ranked according to their significance for classification, calculated as the effect of removing one feature at a time on the classification. The feature with the smallest rank (i.e. with the smallest significance for classification) is discarded. The process continues by considering the remaining features until the maximum accuracy for classification of the training set is achieved. This method was used by Rifkin *et al.* [35] for the multiclass supervised classification of several different types of cancer.

Multi-condition methods

Datasets comprising more than two different types of samples (i.e. three or more class labels) represent a multi-class classification problem. Clearly, in these multi-condition or multi-type systems, gene analysis and selection has to be based on the comparison of means of more than two distributions. Most methods described so far can be used for feature analysis and selection even in multi-type systems, primarily by determining the differences in gene expression in a sample relative to all the other samples [i.e. one versus all (OVA) methods]. However, to look at gene expression differences of individual samples in these systems, a more general concept of relative expression is needed [36].

ANOVA

The *t* test methodology can be generalized in the case of comparison of more than two distributions into one-way analysis of variance (ANOVA). A one-way ANOVA model allows the comparison of the means of an arbitrary number of groups, each of which follows a normal distribution. ANOVA is an approach that can be applied to spotted microarray data of any experimental design and at many different steps in the data analysis process. The ANOVA approach gives an estimate of the relative expression for each gene in each sample [36–39]. The basic concept in ANOVA is that, given an appropriate experimental design, variability in the quantity being measured can be partitioned into various identifiable sources. ANOVA indicates whether variability caused by a particular experimental factor is statistically significant compared with the random variability. Feature analysis using ANOVA is performed by testing each assayed gene independently for a difference in expression between experimental groups. The output of the analysis is a probability (*p*) that a difference in expression could have been observed randomly. A statistical analysis of this kind can reveal genes that show small but significant changes in expression over different sample types.

ANOVA is implemented in a Matlab package [37] as well as in many other packages (listed in Table 2b and c). ANOVA is also included in the package MAANOVA [40], which includes set of functions written (in Matlab, R and Java) for the ANOVA on microarray data.

Correlation coefficient analysis

An alternative method for the selection of prognostic genes is the determination of a correlation between the prognostic category and the logarithmic expression ratio across samples (i.e. whether there is a relationship between sample type and the gene expression pattern). In this case, genes with greatest correlation coefficients between two sample types (i.e. closer relation to the established classification) are more likely candidates for reporting prognosis. Various methods for the calculation of correlation coefficients have been used for the supervised analysis of cancers [6,7,41], the most popular being the Pearson correlation coefficient method (for example, see [22]). Improved methods for the calculation of correlation coefficients in the case of multiple testing have previously been described in detail [42].

Conclusions

Many methods and software tools have been developed and applied for feature analysis and selection in microarray tumour classification and characterization. Recent efforts in gene selection resulted in the determination of 67 genes that appear to be either more or less active in various cancer cells [43]. Following confirmation using other molecular and biochemical techniques, these genes might provide a resource for the development of diagnostic tests as well as potential therapies.

A vast number of methods that can be used for gene selection present an even more interesting and still largely unanswered question regarding the relationship between marker gene determination and the method used for classification [44]. In addition, the relationship between the microarray platform used for the gene expression measurement and the accuracy of tumour classification, which includes feature analysis and selection, presents an additional problem that has received only limited attention [45].

A recent paper by Li and co-workers [46] provides a comparison of the performance of feature selection methods implemented in Rankgene software combined with various multiclass classification methods. The conclusions of this paper were that no method is optimal for all datasets. In addition, the accuracy of the classification was more dependent upon the classification method than the feature analysis and selection methodology used to determine a subset of diagnostic genes. The latter conclusion is in contradiction with the observations of Guyon *et al.* [14] that the gene selection method impacts the accuracy of classification results more than the classification method used. All of the gene selection methods investigated by Li *et al.* [46] assume orthogonality between features and this is probably causing the discrepancy between these two results. Furthermore, all feature analysis and selection methods chosen by Li *et al.* [46] provided a sufficient number of optimal genes for classification. However, further research studying the relationship between a larger group of more diverse methods for feature selection and various subsequently used classification methods for both binary and multi-class problems is essential.

Other methods utilizing array technology, such as comparative genomic hybridization (used for the determination of DNA copy number) arrays and various methods in proteomics and glycomics, can be used for tumour classification. Although the primary focus of this review was DNA microarrays, the feature selection and analysis tools developed in the context of DNA microarrays can be directly translated to other methods.

While this review was being written, the first FDA approval was issued for the application of gene expression analysis in clinical diagnosis (Roche Molecular Systems Inc.). Although this approval was not for cancer diagnosis, it shows that microarray technology is mature enough for clinical applications. However, the application of microarrays in tumour classification still awaits the determination of an optimal set of marker genes. Determination of marker genes as well as optimal tumour classification will not be possible without designing the appropriate clinical trials in combination with genomics that use optimal laboratory and statistical methods and that incorporate biological knowledge [47].

Acknowledgements

M. Cuperlovic-Culf and R. Ouellette would like to acknowledge the generous support of the Atlantic Innovation Fund (ACOA) and the Georges L. Dumont Hospital Foundation. N. Belacel would like to acknowledge the support from National Research Council of Canada, e-health, Saint John. We are grateful to D. Richard and A. Culf for constructive criticism of the manuscript.

References

1 Chung, C.H. *et al.* (2002) Molecular Portraits and the Family Tree of Cancer. *Nat. Genet.* 32, 533–540

2 Liotta, L. and Petricoin, E. (2000) Molecular Profiling of Human Cancer. *Nature Gen.* 1, 48–56

3 Dudoit, S. and Fridlyand, J. (2003) Classification in microarray experiments. In *Statistical Analysis of Gene Expression Microarray Data*. (Speed, T. ed.), Chapman & Hall/CRC Press

4 Liu, E.T. (2003) Classification of Cancers by Expression Profiling. *Curr. Opin. Genet. Dev.* 13, 97–103

5 Golub, T.R. *et al.* (1999) Molecular Classification of Cancer : Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* 286, 531–537

6 Veer, L.J. *et al.* (2002) Gene Expression Profilling Predicts Outcome of Breast Cancer. *Nature* 415, 530–536

7 Vijver, M.J. *et al.* (2002) A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *N. Engl. J. Med.* 347, 1999–2009

8 Bittner, M. *et al.* (2000) Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling. *Nature* 406, 536–540

9 Alon, U. *et al.* (1999) Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. *Proc. Natl. Acad. Sci. U. S. A.* 96, 6745–6750

10 Nagayama, S. *et al.* (2002) Genome-Wide Analysis of Gene Expression in Synovial Sarcoma Using cDNA Microarray. *Cancer Res.* 62, 5859–5866

11 Hackett, J.L. and Lesko, L.J. (2003) Microarray Data – the US, FDA, Industry and Academia. *Nat. Biotechnol.* 21, 742–743

12 Quackenbush, J. (2001) Computational Analysis of Microarray Data. *Nat. Rev. Genet.* 2, 418–427

13 Belacel, N. *et al.* (2004) Fuzzy J-Means and VNS methods for clustering genes from microarray data. *Bioinformatics* 20, 1690–1701

14 Guyon, I. *et al.* (2002) Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.* 46, 389–422

15 Singh, D. *et al.* (2002) Gene expression correlates clinical prostate cancer behaviour. *Cancer Cell* 1, 203–209

16 Eisen, M.B. *et al.* (1998) Cluster Analysis and Display of Genome-Wide Expression Patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868

17 Tamayo, P. *et al.* (1999) Interpreting Patterns of Gene Expression with Self-Organizing Maps : Methods and Application to Hematopoietic Differentiation. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2907–2912

18 Alter, O. *et al.* (2000) Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10101–10106

19 Hastie, T. *et al.* (2000) Gene Shaving as a Method for Identifying Distinct sets of Genes with Similar Expression Patterns. *Genome Biol* 1,3.1-3.21.

20 DeRisi, J.L. *et al.* (1997) Exploring the Metabolic and Genetic Control of Gene Expression on a Genomics Scale. *Science* 278, 680–686

21 Yang, I.V. *et al.* (2002) Within the Fold: Assessing Differential Expression Measures and Reproducibility in Microarray Assays *Genome Biol.* 3, research0062

22 Rosner, B. (2000) *Fundamentals of Biostatistics*, Duxbyrt

23 Ma, X. *et al.* (2003) Gene Expression Profiles of Human Breast Cancer Progession. *Proc. Natl. Acad. Sci. U. S. A.* 100, 5974–5979

24 Bueno, R. *et al.* (2004) A Diagnostic Test for Prostate Cancer from Gene Expression Profiling Data. *J. Urol.* 171, 903–906

25 Amatschek, S. *et al.* (2004) Tissue-Wide Expression Profiling Using cDNA Substraction and Microarrays to Identify Tumor-Specific Genes. *Cancer Res.* 64, 844–856

26 Pan, W. (2002) A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments. *Bioinformatics* 18, 546–554

27 Savege, K.J. *et al.* (2003) The Molecular Signature of Mediastinal Large B-Cell Lymphoma Differs from that of other Diffuse Large B-Cell Lymphomas and Shares Features with Classical Hodkin Lymphoma. *Blood* 102, 3871–3879

28 Yeang, C-H. *et al.* (2001) Molecular Classification of Multiple Tumor Types. *Bioinformatics* 17(Suppl. 1), S316–S322

29 Ramaswamy, S. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U. S. A.* 98, 15149–15154

30 Yeang, C. *et al.* (2001) Molecular classification of multiple tumor types. *Bioinformatics* 17, S316–S322

31 Tusher, V. *et al.* (2001) Significance Analysis of Microarrays Applied to the Ionizing Radiation Response. *Proc. Natl. Acad. Sci. U. S. A.* 98, 5116–5121

32 Troyanskaya, O.G. *et al.* (2002) Nonparametric methods for identifying differentially Expressed Genes in Microarray Data. *Bioinformatics* 18, 1454–1461

33 Park, P. *et al.* (2001) A Nonparametric Scoring Algorithm for Identifying Informative Genes from Microarray Data. *Pac. Symp. Biocomput.* 6, 52–63

34 Dettling, M. and Buhlman, P. (2002) Supervised Clustering of Genes. *Genome Biol.* 3, 69

35 Rifkin, R. *et al.* (2003) An Analytical Method for Multiclass Molecular Cancer Classification. *SIAM Review* 45, 706–723

36 Cui, X. and Churchill, G.A. (2003) Statistical Tests for Differential Expression in cDNA Microarray Experiments. *Genome Biol.* 4, 210

37 Kerr, M.K. *et al.* (2000) Analysis of Variance for Gene Expression Microarray Data. *Jour Comp Biol* 7, 819–837

38 Lee, M.L. *et al.* (2002) Models for Microarray Gene Expression Data. *J. Biopharm. Stat.* 12, 1–19

39 Yang, Y. *et al.* (2003) Statistical Methods for Analyzing Microarray Feature Data with Replications. *J. Comput. Biol.* 10, 157–169

40 Wu, . *et al.* (2002) *MAANOVA: A Software Package for the Analysis of Spotted cDNA Microarray Experiments in The analysis of gene expression data: methods and software*, Springer

41 West, M. *et al.* (2001) Predicting the Clinical Status of Human Breast Cancer by using Gene Expression Profiles. *Proc. Natl. Acad. Sci. U. S. A.* 98, 11462–11467

42 Dudoit, S. *et al.* (2002) Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments. *Statistica Sinica* 12, 111–139

43 Rhodes, D.R. *et al.* (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9309–9314

44 Statnikov, A. *et al.* A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* (in press)

45 Wang, J. *et al.* (2004) Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. *Bioinformatics* 20, 3166–3178

46 Li, T. *et al.* (2004) A Comparative Study of Feature Selection and Multiclass Classification Methods for Tissue Classification Based on Gene Expression. *Bioinformatics* 20, 2429–2437

47 Lonning, P.E. *et al.* (2005) Genomics in breast cancer – therapeutic implications. *Nature Clinical Practice Onco* 2, 26–32

48 Baldi, P. and Long, A.D. (2001) A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized *t* test and Statistical Inferences of Gene Changes. *Bioinformatics* 17, 509–519

49 Simon, R.M. *et al.* (2003) *Design and analysis of DNA microarray investigations*, Spring Verlang

50 Saeed, A.I. *et al.* (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34, 374–378

Related articles in other Elsevier journals

Genomic and proteomic technologies for individualisation and improvement of cancer treatment
Julia Wulfkuhle *et al.* (2004) *Eur. J. Cancer 40, 2623–2632*

Cancer diagnosis and microarrays
Uli Schmidt and C. Glenn Begley (2003) *Int. J. Biochem. Cell Biol. 35, 119–124*

Tissue microarray study for classification of breast tumors
Dao-Hai Zhang *et al.* (2003) *Life Sci. 73, 3189–3199*