Chapter 12

Gene Selection and Sample Classification Using a Genetic Algorithm/k-Nearest Neighbor Method

Leping Li and Clarice R. Weinberg
Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC 27709, USA, *e-mail: {li3,weinberg}@niehs.nih.gov*

1. INTRODUCTION

Advances in microarray technology have made it possible to study the global gene expression patterns of tens of thousands of genes in parallel (Brown and Botstein, 1999; Lipshutz et al., 1999). Such large scale expression profiling has been used to compare gene expressions in normal and transformed human cells in several tumors (Alon et al., 1999; Gloub et al., 1999; Alizadeh et al., 2000; Perou et al., 2000; Bhattacharjee et al., 2001; Ramaswamy et al., 2001; van't Veer et al., 2002) and cells under different conditions or environments (Ooi et al., 2001; Raghuraman et al., 2001; Wyrick and Young, 2002). The goals of these experiments are to identify differentially expressed genes, gene-gene interaction networks, and/or expression patterns that may be used to predict class membership for unknown samples. Among these applications, class prediction has recently received a great deal of attention. Supervised class prediction first identifies a set of discriminative genes that differentiate different categories of samples, e.g., tumor versus normal, or chemically exposed versus unexposed, using a learning set with known classification. The selected set of discriminative genes is subsequently used to predict the category of unknown samples. This method promises both refined diagnosis of disease subtypes, including markers for prognosis and better targeted treatment, and improved understanding of disease and toxicity processes at the cellular level.

1.1 Classification and Gene Selection Methods

Pattern recognition methods can be divided into two categories: *supervised* and *unsupervised*. A supervised method is a technique that one uses to develop a predictor or classification rule using a learning set of samples with known classification. The predictive strategy is subsequently validated by using it to classify unknown samples. Methods in this category include *neighborhood analysis* (Golub et al., 1999), *support vector machines* (SVM) (Ben-Dor et al., 2000; Furey et al., 2000; Ramaswamy et al., 2001), *k-nearest neighbors* (KNN) (Li et al., 2001a & 2001b), *recursive partitioning* (Zhang et al., 2001), *Tukey's compound covariate* (Hedenfalk et al., 2001), *linear discriminant analysis* (LDA) (Dudoit et al., 2002; Li and Xiong, 2002), and "*nearest shrunken centroids*" (Tibshirani et al., 2002). Unsupervised pattern recognition largely refers to clustering analysis for which class information is not known or not required. Unsupervised methods include *hierarchical clustering* (Eisen et al., 1998), *k-means clustering* (Tavazoie et al., 1999), and the *self-organizing map* (Toronen et al., 1999). Reviews of the classification methods can be found in Brazma and Vilo (2000), Dudoit et al. (2002) and Chapter 7 in this volume.

Usually a small number of variables (genes) are used in the final classification. Reducing the number of variables is called *feature reduction* in pattern recognition (e.g.; see Chapter 6). Feature reduction in microarray data is necessary, since not all genes are relevant to sample distinction. For certain methods such as LDA, feature reduction is a must. For other methods such as SVMs, ill-posed data (where the number of genes exceeds the number of samples) are more manageable (e.g.; see Chapter 9).

The most commonly used methods for selecting discriminative genes are the standard two-sample *t*-test or its variants (Golub et al, 1999; Hedenfalk et al., 2001; Long et al., 2001; Tusher et al., 2001). Since typical microarray data consist of thousands of genes, a large number of *t*-tests are involved. Clearly, multiple testing is an issue as the number of chance findings, "false positives", can exceed the number of true positives. A common correction to individual *p* values is the Bonferroni correction. For a two-sided *t*-test, an adjusted significance level is $\alpha^*=\alpha/n$, where n is the number of genes, and α is the unadjusted significance level. When the sample size is small, as the case for most microarray data, the variances may be poorly estimated. One way to address this problem is to "increase" the sample size by using genes with similar expression profiles in variance estimation (Baldi and long 2001; Tusher et al., 2001). Furthermore, *t*-test depends on strong parametric assumptions that may be violated and are difficult to verify with small sample size. To avoid the need for parametric assumptions, one may use permutation techniques (Dudoit et al., 2000; Tusher et al., 2001; Pan et al, 2002). Other methods for selecting

differentially expressed genes include Wilcoxon rank sum test (Virtaneva et al., 2001). A comparative review of some of these methods can be found in (Pan, 2002).

Besides the *t*-test and its variants, one can use a classification method to select discriminative genes. For example, Li and Xiong (2002) used LDA in a stepwise fashion, sequentially building a subset of discriminative genes starting from a single gene. In SVM, Ramaswamy et al. (2001) started with all genes to construct a support vector and then recursively eliminated genes that provided negligible contribution to class separation (the smallest elements in a support vector *w*). The GA/KNN approach (Li et al., 2001a & 2001b) utilizes KNN as the discriminating method for gene selection.

1.2 Why the k-Nearest Neighbors Method?

Many supervised classification methods perform well when applied to gene expression data (see, e.g.; Dudoit et al., 2002). We chose KNN as the gene selection and classification method for the following reasons.

KNN is one of the simplest non-parametric pattern recognition methods. It has been shown to perform as well as or better than more complex methods in many applications (see, e.g.; Vandeginste et al., 1998; Dudoit et al., 2002). Being a non-parametric method, it is free from statistical assumptions such as normality of the distribution of the genes. This feature is important, since the distributions of gene expression levels or ratios are not well characterized and the distributional shapes may vary with the quality of either arrays themselves or the sample preparation.

Like many other supervised methods, KNN is inherently multivariate, taking account of dependence in expression levels. It is known that the expression levels of some genes may be regulated coordinately and that the changes in expression of those genes may well be correlated. Genes that are jointly discriminative, but not individually discriminative, may be co-selected by KNN.

Perhaps most importantly, KNN defines the class boundaries implicitly rather than explicitly, accommodating or even identifying distinct subtypes within a class. This property is particularly desirable for studies of cancer where clinical groupings may represent collections of related but biologically distinct tumors. Heterogeneity within a single tumor type has been shown in many tumors including leukemia (Golub et al., 1999), lymphoma (Alizadeh et al., 2000), and breast cancer (Perou et al., 2000). When applied to a leukemia data set, the GA/KNN method selected a subset of genes that not only discriminated between acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) but also unmasked clinically

meaningful subtypes within ALL (T-cell ALL versus B-cell ALL) – even though gene selection only used the ALL-AML dichotomy (Li et al., 2001a).

Finally, the resulting classification of KNN is qualitative and requires none of the hard-to-verify assumptions about the within-class variances or shift alternatives that are used in many other statistical methods. Typical microarray data contain many fewer samples than genes, and the variance-covariance matrix becomes singular (linear dependences exist between the rows/columns of the variance-covariance matrix), restricting attention to certain linear combinations of genes with non-zero eigenvalues. Moreover, methods that require variance-covariance estimation suffer in the face of outlying observations, disparate covariance structures, or heterogeneity within classes.

1.3 Why a Genetic Algorithm?

In KNN classification, samples are compared in multi-dimensional space. However, considering all possible subsets of genes from a large gene pool is not feasible. For instance, the number of ways to select 30 from 3000 is approximately $6.7 \cdot 10^{71}$. Thus, an efficient sampling tool is needed. A natural choice would be a *genetic algorithm* (GA). A GA is a stochastic optimization method. First described by John Holland in the 70's (Holland, 1975), GAs mimic Darwinian natural selection (hence "genetic") in that selections and mutations are carried out to improve the "fitness" of the successive generations (Holland, 1975; Goldberg, 1989). It starts with a population of "chromosomes" (mathematical entities). Usually, the chromosomes are represented by a set of strings, either binary or non-binary, constituting the building blocks of the candidate solutions. The better the fitness of a chromosome, the larger its chance of being passed to the next generation. Mutation and crossover are carried out to introduce new chromosomes into the population (e.g.; see Judson et al., 1997). Through evolution, a solution may evolve. After it was introduced, GA has been used in many optimization problems ranging from protein folding (Pedersen and Moult, 1996) to sequence alignment (Notredame et al., 1997). For reviews, see Forrest (1997) and Judson (1997). Although, it has been demonstrated that GAs are effective in searching high-dimensional space, they do not guarantee convergence to a global minimum, given the stochastic nature of the algorithm. Consequently, many independent runs of GAs are needed to ensure the convergence.

2. THE GA/KNN METHOD

2.1 Overall Methodology

The GA/KNN (Li et al., 2001a & 2001b) is a multivariate classification method that selects many subsets of genes that discriminate between different classes of samples using a learning set. It combines a search tool, GA, and a non-parametric classification method, KNN. Simply speaking, we employ the GA to choose a relatively small subset of genes for testing, with KNN as the evaluation tool. Details of the GA and KNN are given below.

For high dimensional microarray data with a paucity of samples, there may be many subsets of genes that can discriminate between different classes. Different genes with similar patterns of expression may be selected in different, but equally discriminative, subsets. Consequently, it is important to examine as many subsets of discriminative genes as possible. When a large number of such subsets has been obtained, the frequency with which genes are selected can be examined. The selection frequency should correlate with the relative predictive importance of genes for sample classification: the most frequently selected genes should be most discriminative whereas the least frequently selected genes should be less informative. The most frequently selected genes may be subsequently used to classify unknown samples in a test set.

2.2 KNN

Suppose that the number of genes under study is N and that $q<<N$ is the number of genes in a much smaller subset. Let $G_m = (g_{1m}, g_{2m}, \ldots, g_{im}, \ldots, g_{qm})$ where g_{im} is the expression value (typically *log* transformed) of the ith gene in the mth sample; $m = 1,...,M$. In the KNN method (e.g.; Massart et al., 1988), one computes the distance between each sample, represented by its vector G_m, and each of the other samples (see, e.g.; Table 12.1). For instance, one may employ the Euclidean distance. When values are missing, methods for missing value imputation can be found in Chapter 3. A sample is classified according to the class membership of its k nearest neighbors, as determined by the Euclidean distance in q-dimensional space. Small values of 3 or 5 for k have been alleged to provide good classification. In a classic KNN classification, an unknown sample is classified in the group to which the majority of the k objects belong. One may also apply a more stringent criterion to require all k nearest neighbors to agree in which case a sample would be considered unclassifiable if the k nearest neighbors do not all belong to the same class.

Figure 12.1 displays an example. The unknown, designated by **X**, is classified with the triangles, because its 3 nearest neighbors are all triangles.

Table 12.1. An example of two genes (g1 and g2) and 10 samples (S1-S10).

Sample	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Class	△	△	△	△	△	○	○	○	○	○
g1	$g_{1,1}$	$g_{1,2}$	$g_{1,3}$	$g_{1,4}$	$g_{1,5}$	$g_{1,6}$	$g_{1,7}$	$g_{1,8}$	$g_{1,9}$	$g_{1,10}$
g2	$g_{2,1}$	$g_{2,2}$	$g_{2,3}$	$g_{2,4}$	$g_{2,5}$	$g_{2,6}$	$g_{2,7}$	$g_{2,8}$	$g_{2,9}$	$g_{2,10}$

Note that the data are well clustered, because each observation has a class that agrees with the class of its 3 nearest neighbors.

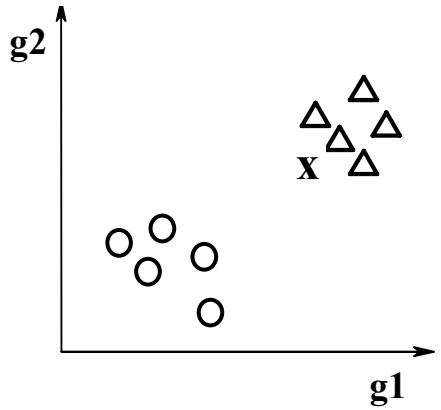

Figure 12.1. KNN classification. For clarity, only two dimensions are shown (q=2), that is, each sample is represented by a vector of two genes (g1 and g2). Triangles and circles represent two distinct classes. A 3-NN classification would assign the unknown sample ***X*** to the class of triangle.

2.3 A Genetic Algorithm

2.3.1 Chromosomes

In GAs, each "chromosome" (a mathematical entity, not the biological chromosome) consists of q distinct genes randomly selected from the gene "pool" (all genes studied in the experiment). Thus, a chromosome can be viewed as a string containing q gene index labels. An example is shown in Figure 12.2. In the example, genes 1, 12, 23, 33, and so on, are selected. The set of q genes in the chromosome constitutes a candidate solution to the gene selection problem, as the goal of each run of the GA is to identify a set of q discriminative genes. Typically, q=20, 30 or 40 should work well for most microarray data sets. A set of such "chromosomes" (e.g.; 100) constitutes a "population" or "niche". We work with 10 such niches in parallel.

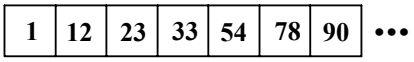

Figure 12.2. An example of a chromosome in GA.

2.3.2 Fitness

The "fitness" of each chromosome is subsequently evaluated by its ability to correctly classify samples using KNN. For each chromosome (a set of q selected genes), we compute the pair-wise Euclidean distances between the samples in the q-dimensional space. The class membership of a sample is then declared by its k-nearest neighbors. If the actual class membership of the sample matches its KNN-declared class, a score of one is assigned to that sample; otherwise, a score of zero is assigned. Summing these scores across all samples provides a fitness measure for the chromosome. A perfect score would correspond to the number of samples in the training set.

2.3.3 Selection and Mutation

Once the fitness score of each chromosome in a niche is determined, the fittest chromosomes, one from each niche, are combined and used to replace the corresponding number of the least fit chromosomes (the lowest scoring chromosomes) in *each* niche. This enrichment strategy allows the single best chromosome found in each niche to be shared with all the other niches. For a typical run with 10 niches, each of which consists of 100 chromosomes, the 10 least fit chromosomes in each niche are replaced by the 10 best chromosomes, one from each niche.

Next, the chromosomes in each niche are ranked, with the best chromosome assigned a rank of 1. The single best chromosome in a niche is passed deterministically to the next generation for that niche *without* subsequent mutation. This guarantees that the best chromosome at each generation is preserved. The remaining chromosomes in the niche are chosen by sampling all chromosomes including the best chromosome in the niche with probability proportional to the chromosome's fitness. This is the so-called "*roulette-wheel selection*" in which the high scoring chromosomes are given high probability of being selected whereas the low scoring chromosomes are given low, but non-zero, probability of being passed to the next generation. Including less fit chromosomes may prevent the search from being trapped at a local minimum. Chromosomes selected based on this sampling strategy are next subject to mutation.

Once a chromosome is selected for mutation, between 1 and 5 of its genes are randomly selected for mutation. The number of mutations (from 1 to 5) is assigned randomly, with probabilities, 0.53125, 0.25, 0.125, 0.0625,

and 0.03125 ($1/2^r$, where r is 1 to 5; 0.03125 is added to the probability r=1 so that the total probability is equal to 1.0), respectively. In this way, a single replacement is given the highest probability while simultaneous multiple replacements have lower probability. This strategy prevents the search from behaving as a random walk as it would if many new genes were introduced at each generation. Once the number of genes to be replaced in the chromosome has been determined, these replacement genes are randomly selected and replaced randomly from the genes not already in the chromosome. An example of a single point mutation is shown in Figure 12.3.

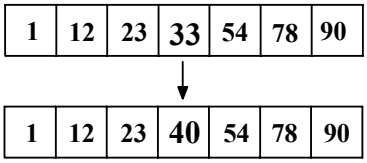

Figure 12.3. A single point mutation. For simplicity, only 7 genes are shown. Upon a single point mutation, gene 33 is replaced by gene 40.

2.3.4 Termination Criterion

Niches are allowed to evolve by repeating the above steps until at least one of the chromosomes achieves a targeted fitness criterion. A targeted fitness criterion is considered to be reached when most of the samples (e.g.; 90% of them) have been correctly classified. Because we do not require perfect classification, gene selection may be less sensitive to outliers or occasional misclassified samples in the data. A less stringent criterion is also computationally faster.

Intuitively, the more distinct classes, the more difficult it will be to find a subset of discriminative genes. For toxicogenomics data or tumor data, multiple classes are not uncommon. For those datasets, the above 90% requirement may be too stringent. For instance, Ramaswamy et al. (2001) did gene expression profiling on 218 tumor samples, covering 14 tumor types, and 90 normal tissue samples using oligonucleotide arrays. When we applied the GA/KNN method to the training set (144 samples and 14 classes), requiring 90% of the 144 samples to be correctly classified was not possible. For such circumstances, one should start with a test run to see how the fitness score evolves from generation to generation. One might choose a fitness score based on what can be achieved in 20 to 40 generations as the targeted fitness value, to balance the computation speed and discrimination power. It should be pointed out that gene selection is relatively insensitive to this choice of the targeted fitness criterion. The other cases where a less

stringent criterion may be needed are time-course and dose-response microarray data, where there are again multiple, potentially similar classes.

We refer to a chromosome that achieves this targeted fitness score as a *near-optimal chromosome*. When a near-optimal chromosome evolves in any niche, that chromosome is retrieved and added to a list; then the entire niche is re-initialized. Because typical microarray data consist of a large number of genes and a small number of samples, for a given data set there may exist many different subsets of genes (near-optimal chromosomes) that can discriminate the classes of samples very well. Hence, the GA/KNN procedure must be repeated through many evolutionary runs, until many such near-optimal chromosomes (e.g.; 10,000) are obtained. Once a large number of near-optimal chromosomes have been obtained, genes can be ranked according to how often they were selected into these near-optimal chromosomes. The most frequently selected genes should be more relevant to sample distinction whereas the least frequently selected genes should be less informative.

It may not be practically possible or necessary to obtain a very large number of near-optimal chromosomes. However, one should check to see if one has sampled enough of the GA solution space for results to stabilize. To do that, one may divide the near-optimal solutions into two groups of equal size and compare their frequency distributions and ranks for the top genes. A tight diagonal line indicates that the ranks for the top genes are nearly reproducible, suggesting that enough near-optimal solutions have been obtained to achieve stability (Figure 12.4).

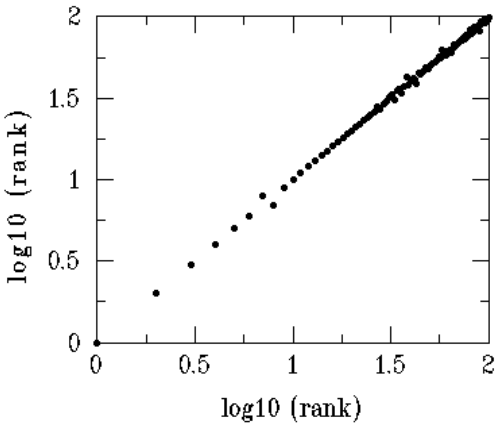

Figure 12.4. An example of plot of the log_{10}-transformed ranks of the 100 top-ranked genes from two independent runs of the GA/KNN procedure. The genes were ranked according to frequency of occurrence in the 500,000 near-optimal chromosomes, with the most frequent gene assigned rank 1 (0 after transformation). Similar result was obtained using fewer near-optimal chromosomes (e.g.; 10,000).

2.4 Statistical Analysis of the Near-optimal Chromosomes

The next step is to develop a predictive algorithm to apply to the test set, by selecting a certain number of top-ranked genes and using those genes, with the KNN method, on the test set samples. A simple way to choose the number of discriminative genes is to take the top 50. Although fewer genes (e.g.; 10) may be preferred in classification, for microarray data, a few more genes might be useful. More genes might provide more insight for the underlying biology. With more genes, the classification should be less sensitive to the quality of data, since the current microarray technology is not fully quantitative. Alternatively, one may choose the number of top-ranked genes that give optimal classification for the training set (Li et al., 2001b). It may also be helpful to plot the Z score of the top-ranked genes (Figure 12.5). Let $Z = \frac{S_i - E(S_i)}{\sigma}$, where S_i is the number of times gene i was selected, $E(S_i)$, is the expected number of times for gene i being selected, σ is the square root of the variance. Let A = number of near-optimal chromosomes obtained (not necessarily distinct) and P_i = q/number of genes on the microarray, the probability of gene i being selected (if random). Then, $E(S_i) = P_i \times A$, and $\sigma = \sqrt{P_i \cdot (1 - P_i) \cdot A}$. A sharp decrease in Z score may suggest that only a few of the top-ranked genes should be chosen as the discriminative genes.

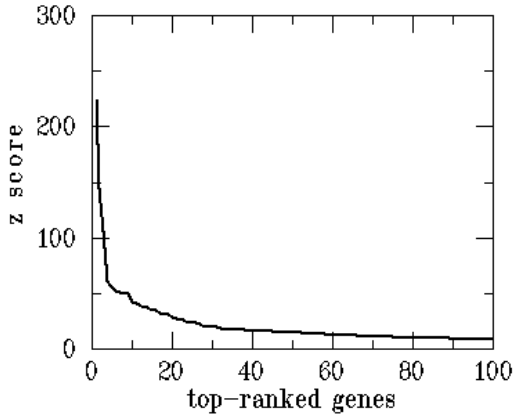

Figure 12.5. A plot of Z scores for 100 top-ranked genes for the breast cancer data set (Hedenfalk et al., 2001). The Z scores decrease quickly for the first 5 to 10 genes. The decrease is much slower after 30 genes. In this case, it seems reasonable to choose 20 to 30 top-ranked genes as the most discriminative genes.

2.5 Comparison between Near-optimal Chromosomes and the Top-ranked Genes

As pointed out earlier, for high-dimensional microarray data with a paucity of samples, many subsets of genes that can discriminate between different classes of samples may exist. Different genes with similar patterns of expression may be selected in different, but equally discriminative subsets, especially when a qualitative classification method, such as KNN, is used. The overlap between q top-ranked genes and each of the near-optimal chromosomes (q genes in length) can be low. For instance, for the breast cancer data set (Hedenfalk et al., 2001), we obtained 500,000 near-optimal chromosomes that can distinguish between *BRCA1* and *BRCA2* tumors. Among the 500,000 near-optimal chromosomes, only 13% of them had 6 or more genes listed among the 30 top-ranked genes. Moreover, classifications in a leave-one-out cross-validation procedure (e.g.; Chapter 7) using the individual near-optimal chromosomes revealed bad performance (data not shown). On the other hand, empirically, we found that substantially larger separation between *BRCA1* and *BRCA2* samples was achieved with the 30 top-ranked genes than with any individual near-optimal chromosome (data not shown). These results suggest that the top-ranked genes do much better than any of the individual near-optimal chromosomes for sample classification.

Although ranking the genes by selecting those individual genes that occur most frequently in near-optimal chromosomes may seem to sacrifice correlation structure, this selection process appears to retain aspects of multivariate structure important for class discrimination. Heuristically, when a subset of genes can discriminate among classes jointly, but not singly, that subset of genes should tend to appear together in near-optimal chromosomes and, consequently, each gene in the jointly discriminative subset may tend to have high frequency of occurrence.

2.6 Computation Cost

The GA/KNN method is computationally intensive, as it searches for many near-optimal solutions (chromosomes). For a typical run, as many as 10,000 near-optimal solutions may be needed. For a small data set with 10 samples in each of two categories, obtaining that many near-optimal solutions can be achieved in a few hours or less. However, for a large data set with multiple classes (e.g.; the MIT's 14 categories tumor data set) (Ramaswamy et al., 2001), it may take a few days to complete the GA/KNN on a Linux machine with reasonable speed.

2.7 Availability

The GA/KNN method will be available on the Web site: http://dir.niehs.nih.gov/microarray/datamining/ for downloading in September 2002.

3. CONCLUDING REMARKS

In summary, the GA/KNN method is non-parametric, multivariate, and able to accommodate (and potentially detect) the presence of heterogeneous subtypes within classes. As the quantitative aspects of the microarray technology improve and computational methods that mine the resulting large data sets are developed further, the technology will have a great impact on biology, toxicology, and medicine.

4. ACKNOWLEDGEMENTS

We thank David Umbach and Shyamal Peddada for insightful discussions and careful reading of the manuscript. LL also thanks Lee Pedersen and Thomas Darden for advice and support.

5. REFERENCES

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Staudt LM, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000; 403:503-11.

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 1999; 96:6745-50.

Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics 2001; 17:509-19.

Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z. Tissue classification with gene expression profiles. J Comput Biol 2000; 7:559-83.

Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA 2001; 98:13790-5.

Brazma A, Vilo J. Gene expression data analysis. FEBS Lett 2000; 480:17-24.

Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. Nat Genet 1999; 21(1 Suppl):33-7.

Dudoit S, Yang YH, Callow MJ, Speed T. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report, Number 578, 2000, Department of Statistics, University of California, Berkeley, California.

Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002; 97:77-87.

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998; 95:14863-8.

Forrest S. Genetic algorithms: principles of natural selection applied to computation. Science 1993; 261:872-8.

Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000; 16:906-14.

Goldberg, David E., *Genetic algorithms in search, optimization, and machine learning*. Massachusette: Addison-Wesley, 1989.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999; 286:531-7.

Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J. Gene-expression profiles in hereditary breast cancer. N Engl J Med 2001; 344:539-48.

Holland, John H., *Adaptation in Natural and Artificial Systems*., Ann Arbor: University of Michigan Press, 1975.

Judson, R. "Genetic algorthms and their use in chemistry." In *Reviews in computational chemistry*, Kenny B. Lipowitz and Donald B. Boyd, eds. New York: VCH publishers, vol 10, 1997.

Li L, Darden TA, Weinberg CR, Levine AJ, Pedersen LG. Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. Comb Chem High Throughput Screen 2001; 4:727-39.

Li L, Weinberg CR, Darden TA, Pedersen LG. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 2001; 17:1131-42.

Li W, Xiong M. Tclass: tumor classification system based on gene expression profile. Bioinformatics 2002, 18:325-326.

Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. Nat Genet 1999; 21(1 Suppl):20-4.

Long AD, Mangalam HJ, Chan BY, Tolleri L, Hatfield GW, Baldi P. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. J Biol Chem 2001; 276:19937-44.

Massart, D.L.; Vandeginste, B.G.M.; Deming, S.N.; Michotte, Y.; Kaufman; L. *Chemometrics*: *a textbook (Data Handling in Science and Technology, vol 2)*, Elsevier Science B.V: New York, 1988.

Notredame C, O'Brien EA, Higgins DG. RAGA: RNA sequence alignment by genetic algorithm. Nucleic Acids Res 1997; 25:4570-80.

Ooi SL, Shoemaker DD, Boeke JD. A DNA microarray-based genetic screen for nonhomologous end-joining mutants in Saccharomyces cerevisiae. Science 2001; 294:2552-6.

Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 2002; 18:546-54.

Pedersen JT, Moult J. Genetic algorithms for protein structure prediction. Curr Opin Struct Biol 1996; 6:227-31.

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Aksien LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature 2000; 406: 747-52.

Raghuraman MK, Winzeler EA, Collingwood D, Hunt S, Wodicka L, Conway A, Lockhart DJ, Davis RW, Brewer BJ, Fangman WL. Replication dynamics of the yeast genome. Science 2001; 294:115-21.

Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA 2001; 98:15149-54.

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet 1999; 22:281-5.

Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 2002; 99:6567-72.

Toronen P, Kolehmainen M, Wong G, Castren E. Analysis of gene expression data using self-organizing maps. FEBS Lett 1999; 451:142-6.

Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001; 98:5116-21.

Vandeginste, B.G.M., Massart, D.L., Buydens, L.M.C., De Jong, S., Lewi, P.J. and Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics. Vol 20B*. The Netherlands: Elsevier Science, 1998.

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS,

Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415:530-6.

Virtaneva K, Wright FA, Tanner SM, Yuan B, Lemon WJ, Caligiuri MA, Bloomfield CD, de La Chapelle A, Krahe R. Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. Proc Natl Acad Sci USA 2001; 98:1124-9.

Wyrick JJ, Young RA. Deciphering gene expression regulatory networks. Curr Opin Genet Dev 2002; 12:130-6.

Zhang H, Yu CY, Singer B, Xiong M. Recursive partitioning for tumor classification with gene expression microarray data. Proc Natl Acad Sci USA 2001; 98:6730-5.