CHAPTER 1

INTRODUCTION

Classification analysis of DNA microarrays has grown rapidly since the introduction of microarrays in 1995 [1]. Original methodological research covered cluster analysis, which spanned into principal and independent component analysis, and more recently, support vector machines. Current approaches include deterministic annealing, particle swarm optimization, ant colony optimization, and random linear oracles. Future growth areas in classification will likely address parallelization of multiple random linear oracles for ensemble classifier fusion, as a way to overcome diversity problems with today's ensemble methods.

This book discusses classification analysis of DNA microarrays. Topics are introduced in an order that increases with the complexity of the method covered. This chapter introduces the reader to the terminology and rules of thumb used in classification analysis in general, and for DNA microarrays specifically. As the ideas are presented, linkage of topics with the specific information presented in later chapters will become apparent. This chapter is divided into three main parts: class discovery, dimensional reduction, and class prediction. Class discovery (see Section 1.1) is commonly viewed as a hypothesis-generating step that forms an essential part of *knowledge discovery*. Here, the basic assumption is that there is little information about the cluster structure of the data. Analogously, when new molecular methods are used, the goal is often to look for new patterns among the objects (e.g., microarrays) if the research project represents a novel application of laboratory methods for which there is no precedent. The part on dimensional reduction (Section 1.2) offers an approach for collapsing dimensions into a smaller set of features that describe a majority of variation and informativeness of the original larger set. Later, class prediction (Section 1.3) is introduced as a method for training and testing with a learned procedure in order to predict class labels of new unknown objects. Obviously, the learning parameters and predictive

Classification Analysis of DNA Microarrays, First Edition. Leif E. Peterson.

accuracy change as new unknowns are added to the system, and this *concept drift* problem is a greater concern when operationalizing a method in the real world. At this point, we would expect a classification system to perform quite well during external validity tests to classify objects obtained after training and testing, or objects omitted from training that are part of the original training set.

1.1 CLASS DISCOVERY

In today’s molecular and genomic world, the first step in analyzing a new dataset should be to perform class discovery from a knowledge discovery approach. You can get more ‘‘bang for the buck’’ or a greater *return on investment* if you exploit newly generated data from the laboratory by trying to determine whether there are new diagnostic classes present, or patients (animals) that don’t fit in with the assumed cluster structure. This problem can be partitioned into three categories.

Novel Diagnostic Classes. Never before has there been a better chance to identify new categories of a disease that were previously unknown or unobserved. By augmenting clinical data with DNA microarray data and employing knowledge discovery methods, you may be able to discern new patterns in the data that have never been seen before. Some of the new patterns of objects (patients, animals) may reveal clusters that suggest a new diagnostic class, while others may reflect objects that don’t fit in with the assumed structure. Far too often investigators with little experience in linear and nonlinear class discovery methods are unaware of the rich repertoire of methods available for identifying new patterns in a set of data. Other times, there is such a strong focus on hypothesis-driven methods that there is little time available to perform knowledge discovery.

Comorbidity and Overlap. Comorbidity relates to phenotype and represents overlap of disease or diagnostic classes within a disease. Elderly patients presenting with pneumonia may have cardiovascular disease and electrolyte imbalance. In behavioral genetics, one of the best known examples of comorbidity from a biomolecular perspective occurs with alcohol and nicotine dependence [2]. Additionally, it is also well recognized that bipolar disorder, alcoholism, and stress reactivity are comorbid [3]. Depression and comorbidity with epilepsy is another area where DNA microarrays have been employed [4]. Use of class discovery methods for identifying the presence of comorbidity when using DNA microarrays in animal studies or in clinical research may identify new regions of phenotypic overlap.

Outliers and Heterogeneity. Occasionally there are objects that do not cluster with the majority of data and either form their own small groups or act singly as *outliers* due to their unique feature characteristics. Such objects seldom fit in with other objects in the major clusters or major diagnostic categories and are misclassified in class prediction models. Misclassification is common when using genomic data to cluster objects into the diagnostic categories from which objects were originally drawn. Misclassification can be caused by systematic error in sample collection, use of nonstandard buffers and laboratory methods, lack of equipment calibration, intrinsic errors in instrumentation ($1/f$ pink noise, electronic fluctuations, etc.), or genetic and environmental determinants of disease. There is no a guarantee that all objects will fall into discrete known categories, and this should never be assumed.

Table 1.1 lists several class discovery methods described in this book. Crisp and fuzzy K-means cluster analyses (CKM, FKM) are partitional clustering methods, which group objects together in K clusters. The optimal number of clusters is determined by using *cluster validity*. Self-organizing maps (SOMs) provide an unsupervised method that incorporates a neighborhood function to reward learning. Unsupervised neural gas (UNG) uses prototype learning and a punishment-reward learning method for deriving cluster structure. Hierarchical cluster analysis (HCA) is an *agglomerative* cluster method that starts with individual objects, and adds together objects having similar profiles. Divisive cluster analysis works the other way, starting with the entire set of objects as one cluster, and ending up with smaller clusters at the end. The Gaussian mixture model (GMM) method uses the

TABLE 1.1 Unsupervised Class Prediction Methods

Method	Remarks
Crisp K-means cluster analysis (CKM)	Iterative reduction of object-cluster distance
Fuzzy K-means cluster analysis (FKM)	Object-specific membership function for each class
Self-organizing maps (SOM)	Partitions objects using neighborhood functions
Unsupervised neural gas (UNG)	Partitions objects with rank methods
Hierarchical cluster analysis (HCA)	Natural grouping of objects
Gaussian mixture models (GMM)	Exploits the EM algorithm for training

expectation-maximization (EM) algorithm to determine the cluster structure of microarrays.

1.2 DIMENSIONAL REDUCTION

Gene expression datasets are notorious for having a very large number of features (genes). In most cases, not all the features are needed for class discovery and class prediction. Instead, it is possible to reduce the number of features to a lower number of dimensions that can retain the informativeness about the cluster structure of microarrays while explaining a majority of variation in the original dataset. Linear dimensional reduction in the form of principal components analysis (PCA) is first covered in Part II. This is followed by a chapter on nonlinear manifold learning (NLML) for embedding a lower-dimensional map into the original higher-dimensional sample space.

1.3 CLASS PREDICTION

Class prediction methods address the ability that a classifier can learn information from the features of objects, and then make an accurate prediction to assign objects to their true class. This requires knowledge of the true class of each object, and tabulation of this knowledge is commonly referred to as a *truth table*. During *unsupervised* class prediction, the class prediction method does not consider the misclassification error during the learning stage. In *supervised* class prediction, misclassification error is monitored during the learning process, and parameters are updated to achieve better prediction accuracy.

Supervised Methods. The following supervised classification methods (classifiers) are discussed in this book: linear regression (LREG), decision tree classification (DTC), random forests (RF), *K*-nearest neighbor (KNN), naïve Bayes classifier (NBC), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), Fisher's discriminant analysis (FDA), learning vector quantization (LVQ1), logistic regression (LOG), polytomous logistic regression (PLOG), gradient ascent support vector machines (SVMGA), least-squares support vector machines (SVMLS), artificial neural networks (ANN), kernel regression (KREG), genetic algorithms (GA), covariance matrix self-adaptation (CMSA), particle swarm optimization (PSO), ant colony optimization (ACO), supervised neural gas (SNG), and mixture of experts (MOE). Table 1.2 lists the classifiers used and some remarks regarding their applications.

TABLE 1.2 Supervised Class Prediction Methods

Classifier	Remarks
Linear regression (LREG)	For linearly separable classes
Decision tree classification (DTC)	Classification component of CART[a]
Random forests (RF)	Possibly the least generalization error
K-nearest neighbor (KNN)	Instanced-based learning, "lazy learner"
Naïve Bayes classifier (NBC)	Bayesian classifier, assumes feature independence
Linear discriminant analysis (LDA)	Assumes equal covariance matrices
Quadratic discriminant analysis (QDA)	Assumes unequal covariance matrices
Fisher's discriminant analysis (FDA)	Reduced rank discriminants
Learning vector quantization (LVQ1)	Hebbian nearest prototype learning
Logistic regression (LOG)	2-class maximum likelihood
Polytomous logistic regression (PLOG)	K-class maximum likelihood
Gradient ascent support vector machines (SVMGA)	L_1 soft norm, convex
Least-squares support vector machines (SVMLS)	L_2 soft norm, strictly convex (unique solution)
Artificial neural networks (ANN)	Massively parallel connectionist model
Kernel regression (KREG)	Distance weighted regression
Genetic algorithms (GA)	Genetic selection, crossover, mutation
Covariance matrix self-adaptation (CMSA)	Evolutionary strategy
Particle swarm optimization (PSO)	Swarm intelligence
Ant colony optimization (ACO)	Swarm intelligence with pheromone
Supervised neural gas (SNG)	Prototype learning with error minimization
Mixture of experts (MOE)	Gated mixtures of experts and EM algorithm

[a] Classification and regression trees (algorithm).

1.4 CLASSIFICATION RULES OF THUMB

There are many rules of thumb in machine learning, and classification analysis has its own special rules that have taken form over the last several decades. These are enumerated below, and should be followed to the extent possible because they form a foundation for this book.

1. **Microarrays are objects; genes are features or attributes.** The data used in classification analysis commonly fall into three categories: features, objects, and class labels. *Features* are, by definition, the attributes or characteristics of each object, usually obtained by some sort of measurement or assessment. *Microarrays* are defined as the objects or instances, each of which have features. Thus, an *object* usually consists of a vector of feature values. Objects can also have a *class label*, which for all objects makes up what is called the truth table.
2. **Class discovery does not use class labels.** Class discovery attempts to arrange objects so that a cluster structure is discernible and can provide new insight into the distribution of data in the *sample space*. The fundamental reason for using class discovery is to employ a class discovery technique to partition or arrange objects without considering their true class labels.
3. **Class discovery is commonly independent of feature selection.** The first step in any classification analysis is to determine what features should be used. Class discovery may or may not use all of the available features. However, it is more common to use all of the available features for class discovery, since feature selection is usually done to optimize performance (accuracy) of a class prediction analysis. Expert judgement for selecting ''relevant'' features is probably the most reasonable reason for preselecting features prior to class discovery analysis. With DNA microarrays it is very typical to cluster the objects using all genes in order to reveal patterns in heterogeneity or new clusters from heatmaps. Certainly, if feature selection is performed for class prediction, then the true classes of objects should be observable in heatmaps generated from the selected features. Thus, the only rationale for selecting features for class discovery is to ensure that only features that are relevant to the data and research questions being addressed are used. For this reason, class discovery is commonly performed independently from feature selection.
4. **Class prediction requires known class labels.** Knowing the truth table of class labels for objects is required for class prediction. Machine learning via training and testing of objects is guided from the truth table of known true objects' classes, and prediction accuracy is hinged to the truth table.
5. **Class prediction commonly employs feature selection.** Feature selection is performed in class prediction in order to increase computational efficiency and optimize class prediction accuracy. Computational efficiency is increased when there are fewer features used, and classification accuracy improves when an optimal set of features is selected. This is a basic tenet of statistical modeling, where the best variables are selected when performing function approximation or

developing a predictive model. The most common methods of feature selection are filtering and wrapping. Filtering selects features using scores independently of the classification procedure, while wrapping incorporates the feature selection process directly into the classification analysis. Use of parametric and nonparametric statistical tests to identify features that are significantly different across class labels independently and prior to statistical modeling would be a form of filtering. Stepwise regression, however, involving feature selection during the optimization process is a form of wrapping. It warrants noting that wrapping, or the selection of features during the classification process, can result in biased predictions since the features selected are specific to the classifier used. Occasionally all the available features are used because either they are relevant to the research question or the investigators chose not to perform feature selection.

6. **Dimensional reduction is typically performed on features.** Dimensional reduction methods are typically used for reducing the number of features involved in a ''high-dimensional'' classification analysis. DNA microarrays often result in the *small-sample problem*, where the number of features greatly outweighs the number of objects (i.e., $p \gg n$). Principal component analysis (PCA), *K*-means cluster analysis, and nonlinear manifold learning can be used to reduce the number of features down to a manageable number. An important distinction between dimensional reduction and feature selection is that the former will result in a reduced set of dimensions that are not the same as any of the original features. Any attempt to make inferences on the original features will require a mapping transformation back to the original feature space. Nevertheless, there is great utility in the reduced feature dimensions, since they represent the major source of variation in the *feature space*. Objects can also be used for dimensional reduction. *K*-means cluster analysis is performed when generating *prototypes* for learning vector quantization and *centers* for radial basis function networks and kernel regression. The number of nodes used in self-organizing maps can also be determined by running *K*-means cluster analysis on objects, and then setting the nodes equal to the number of centers generated. Any of the dimensional reduction methods described in this book can be used on a larger set of features to obtain reduced features that are then input into class discovery or class prediction algorithms. Class prediction results after using SOM, UNG, and NLML are provided in the respective chapters for these reduction techniques.
7. **Dimensional reduction on objects is common in cluster analysis.** *K*-means cluster analysis, neural gas, learning vector quantization (LVQ), and radial basis function networks all use methods that

develop *centers* representing multiple groups of objects. In *K*-means cluster analysis the centers have the same dimensions as the number of features and are determined with the average feature values of objects assigned to a given cluster. Neural gas and LVQ use Hebbian learning to derive cluster centers, and in radial basis function networks, the network input node values are based on the distance between each object and every center. What is important is that the user knows that dimensional reduction is being carried out on either the object space or the feature space.

8. **Classification is performed on objects.** It is more common to perform class discovery than class prediction on features, because the use of class labels for features is not nearly as popular as using class labels for objects. For DNA microarrays, class discovery is performed on genes to determine clusters of coregulation or like families of proteins; however, it is less popular to have a research goal to ''better predict'' gene class membership on the basis of the some truth table.
9. **No single feature set is the best: Ugly Duckling Theorem.** The Ugly Duckling Theorem states that classification always assumes inherent bias, and learning is therefore impossible without bias [5]. If *S* represents a swan and *D* represents a duckling, there is no difference between the ugly duckling and two swans when lined up spatially as *SSD*, *SDS*, *DSS*. Therefore, for a specific classifier, there is no reason why a particular set of features should be favored over another.
10. **No single classifier is the best: No-Free-Lunch Theorem.** The No-Free-Lunch Theorem states that ''For any two learning algorithms, independent of class priors and the number of objects: There is no difference in expected error when averaged over all target functions, or averaged over all priors'' [6]. Hence, if there are *X* reasons why the first learning algorithm outperforms the second, then there are *X* different reasons why the second outperforms the first. Learning is impossible without assumptions, so any observed superiority of one classifier over another is due to the nature of the problem. Overall, the No-Free-Lunch Theorem provides justification that no one classifier is better than another.
11. **Simple is better: Occam's Razor.** The Occam's Razor Theorem states that simpler is better; thus, do not use classifiers that are overly complex. However, a dilemma arises because the No-Free-Lunch Theorem states that simple algorithms should not be favored over complex ones. The Occam's Razor viewpoint about the assumption that accuracy always increases with a higher number of objects is likely not true.

12. **Machine learning is not statistical analysis.** Statistical analysis is based on human interaction, and is usually geared toward inferential hypothesis testing; Type I and Type II errors; sample size and statistical power; normality and distributional assumptions; data range, scale, and transformation; and modeling to fit data. On the other hand, most machine learning methods for classification focus more on establishing consistent objectivity over thousands (millions) of repeated procedures, conducting large-scale repetitive analyses that are humanly impossible in the context of performing them manually, and obtaining high performance and reproducibility consistently in regions of complex decision boundaries where humans break down. In addition, more recent research into machine learning and classification has focused on classifier fusion, diversity, robotics, and automation via expert systems.
13. **Machine learning is not computational intelligence.** Classification with machine learning techniques encapsulates classical statistical methods such as logistic regression and discriminant analysis, Hebbian learning in learning vector quantization, winner-take-all and punishment-reward approaches used in self-organizing maps, as well as kernel methods used in radial basis function networks and support vector machines. Computational intelligence involves three areas, namely, artificial neural networks, soft (fuzzy) computing, and evolutionary algorithms. Artificial neural networks are massively parallel connectionist machines whose constituent perceptrons mimic the neuron in the brain. Soft computing or fuzzy methods exploit uncertainty to make sense of ambiguous data. We have shown on more than one occasion that feature fuzzification can improve classification performance. Evolutionary algorithms such as genetic algorithms mutate and exchange chromosomal regions of object data to arrive at an optimal classification configuration.

1.5 DNA MICROARRAY DATASETS USED

Data used for classification analysis in this book were originally available in C4.5 format from the Kent Ridge Biomedical Data Set Repository (`http://sdmc.i2r.a-star.edu.sg/rp`). At present, many of the datasets used are available at the BRB-ArrayTools Data Archive for Human Cancer Gene Expression [7]. The two-class pediatric brain cancer data consisted of 60 arrays (21 failures, 39 survivors) with expression for 7129 genes [8]. The two-class adult prostate cancer dataset consisted of 102 training arrays (52 tumor, and 50 normal) with 12,600 features. The original report for the prostate data supplement was published by Singh et al. [9]. Two breast cancer datasets were used. The first had two classes and consisted of 15 arrays for eight BRCA1-positive women and seven BRCA2-positive women with expression profiles of 3170 genes [10], and the second was also a two-class set including 78 patient arrays and 24,481 features (genes) consisting of 34 cases with distant metastases who relapsed (''relapse'') within 5 years after initial diagnosis and 44 disease-free (''nonrelapse'') cases for more than 5 years after diagnosis [11]. Two-class expression data for adult colon cancer were based on the paper published by Alon et al. [12]. The dataset contains 62 arrays based on expression of 2000 genes in 40 tumor biopsies (''negative'') and 22 normal (''positive'') biopsies from nondiseased colon biopsies from the same patients. An adult two-class lung cancer set including 32 arrays [16 malignant pleural mesothelioma (MPM) and 16 adenocarcinoma (ADCA)] of the lung with expression values for 12,533 genes [13] was also considered. Two leukemia datasets were evaluated; one was a two-class dataset with 38 arrays (27 ALL, 11 AML) containing expression for 7129 genes [14], and the other consisted of three classes for 57 pediatric arrays for lymphoblastic and myelogenous leukemia (20 ALL, 17 MLL, and 20 AML) with expression values for 12,582 genes [15]. The Khan et al. [16] dataset on pediatric small round blue cell tumors (SRBCTs) had expression profiles for 2308 genes and 63 arrays constituting four classes [23 arrays for Ewing sarcoma (EWS), 8 arrays for Burkitt's lymphoma (BL), 12 arrays for NB-neuroblastoma (NB), and 20 arrays for rhabdomyosarcoma) (RMS)].

The entire gene sets listed Table 1.3 were rarely used for unsupervised or supervised classification runs in this book. Instead, gene filtering was

TABLE 1.3 Datasets Used for Classification Analysis

Cancer site	Class	Arrays	Features	Reference
Brain	2	60 (21 failures, 39 survivors)	7,129	8
Prostate	2	102 (52 tumor, 50 normal)	12,600	9
Breast	2	15 (8 BRCA1, 7 BRCA2)	3,170	10
Breast	2	78 (34 relapse, 44 nonrelapse)	24,481	11
Colon	2	62 (40 negative, 22 positive)	2,000	12
Lung	2	32 (16 MPM, 16 ADCA)	12,533	13
Leukemia	2	38 (27 ALL, 11 AML)	7,129	14
Leukemia	3	57 (20 ALL, 17 MLL, 20 AML)	12,582	15
SRBCT	4	63 (23 EWS, 8 BL, 12 NB, 20 RMS)	2,308	16

applied to reduce the number of genes down to a workable level for which there was reduced redundancy and correlation, less noise, and more parsimony. Example 6 (in Chapter 12) describes the main method used for gene filtering.

We did not simulate data for analysis in this book, mainly because the goal was to investigate the characteristics of various classifiers and influence of sample size, statistical significance of features selected, standardization, and fuzzification of features on performance for empirical data. By limiting the coverage to only empirical data, we ensured that the results presented are generalizable to the data considered.

REFERENCES

[1] M. Schena, D. Shalon, R.W. Davis, P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**:467–470, 1995.

[2] N. Ait-Daoud, G.A. Wiesbeck, P. Bienkowski, M.D. Li, R.H. Pfutzer, M.V. Singer, O.M. Lesch, B.A. Johnson. Comorbid alcohol and nicotine dependence: From the biomolecular basis to clinical consequences. *Alcohol Clin. Exp. Res.* **29**(8):1541–1549, 2005.

[3] H. Le-Niculescu, M.J. McFarland, C.A. Ogden, Y. Balaraman, S. Patel, J. Tan, Z.A. Rodd, M. Paulus, M.A. Geyer, H.J. Edenberg, S.J. Glatt, S.V. Faraone, J.I. Nurnberger, R. Kuczenski, M.T. Tsuang, A.B. Niculescu. Phenomic, convergent functional genomic, and biomarker studies in a stress-reactive genetic animal model of bipolar disorder and co-morbid alcoholism. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **147**(2):134–166, 2008.

[4] S. Koh, R. Magid, H. Chung, C.D. Stine, D.N. Wilson. Depressive behavior and selective down-regulation of serotonin receptor expression after early-life seizures: Reversal by environmental enrichment. *Epilepsy Behav*. 10(1):26–31, 2007.

[5] S. Watanabe. *Pattern Recognition: Human and Mechanical*. Wiley, New York, 1985.

[6] D.H. Wolpert, W.G. Macready. No free lunch theorems for optimization. *IEEE Trans. Evolut. Comput.* **1**:67, 1997.

[7] Y. Zhao, R. Simon. BRB ArrayTools data archive for human cancer gene expression: A unique and efficient data sharing resource. *Cancer Inform.* **6**:9–15, 2008.

[8] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.-Y.H. Kim, L.C. Goumnerovak, P.M. Blackk, C. Lau, J.C. Allen, D. ZagzagI, J.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califanokk, G. Stolovitzkykk, D.N. Louis, J.P. Mesirov, E.S. Lander, T.R. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**(6870):436–442, 2002.

[9] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, W.R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**(2):203–209, 2002.

[10] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, M. Raffeld, Z. Yakhini, A. Ben-Dor, E. Dougherty, J. Kononen, L. Bubendorf, W. Fehrle, S. Pittaluga, S. Gruvberger, N. Loman, O. Johannsson, H. Olsson, B. Wilfond, G. Sauter, O-P. Kallioniemi, A.A. Borg, J. Trent. Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* **344**:539–548, 2001.

[11] L.J. van't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, S.H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**:530–536, 2002.

[12] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine. Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **96**(12):6745–6750, 1999.

[13] G.J. Gordon, R.V. Jensen, L.L. Hsiao, S.R. Gullans, J.E. Blumenstock, S. Ramaswamy, W.G. Richards, D.J. Sugarbaker, R. Bueno. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res.* **62**(17):4963–5967, 2002.

[14] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression. *Science* **286**:531–537, 1999.

[15] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub, S.J. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genet.* **30**(1):41–47, 2001.

[16] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, R.S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med.* **7**:673–679, 2001.