D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

Contents

Acknowledgements *iii*

Preface *iv*

1 Introduction to Microarray Data Analysis 1
Werner Dubitzky, Martin Granzow, C. Stephen Downes, Daniel Berrar

2 Data Pre-Processing Issues in Microarray Analysis 47
Nicholas A. Tinker, Laurian S. Robert, Gail Butler, Linda J. Harris

3 Missing Value Estimation 65
Olga G. Troyanskaya, David Botstein, Russ B. Altman

4 Normalization 76
Norman Morrison and David C. Hoyle

5 Singular Value Decomposition and Principal Component Analysis 91
Michael E. Wall, Andreas Rechtsteiner, Luis M. Rocha

6 Feature Selection in Microarray Analysis 110
Eric P. Xing

7 Introduction to Classification in Microarray Experiments 132
Sandrine Dudoit and Jane Fridlyand

8 Bayesian Network Classifiers for Gene Expression Analysis 150
Byoung-Tak Zhang and Kyu-Baek Hwang

D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

9 Classifying Microarray Data Using Support Vector Machines 166
Sayan Mukherjee

10 Weighted Flexible Compound Covariate Method for Classifying Microarray Data 186
Yu Shyr and KyungMann Kim

11 Classification of Expression Patterns Using Artificial Neural Networks 201
Markus Ringnér, Patrik Edén, Peter Johansson

12 Gene Selection and Sample Classification Using a Genetic Algorithm and *k*-Nearest Neighbor Method 216
Leping Li and Clarice R. Weinberg

13 Clustering Genomic Expression Data: Design and Evaluation Principles 230
Francisco Azuaje and Nadia Bolshakova

14 Clustering or Automatic Class Discovery: Hierarchical Methods 246
Derek C. Stanford, Douglas B. Clarkson, Antje Hoering

15 Discovering Genomic Expression Patterns with Self-Organizing Neural Networks 261
Francisco Azuaje

16 Clustering or Automatic Class Discovery: non-hierarchical, non-SOM 274
Ka Yee Yeung

17 Correlation and Association Analysis 289
Simon M. Lin and Kimberly F. Johnson

18 Global Functional Profiling of Gene Expression Data 306
Sorin Draghici and Stephen A. Krawetz

19 Microarray Software Review 326
Yuk Fai Leung, Dennis Shun Chiu Lam, Chi Pui Pang

20 Microrray Analysis as a Process 345
Susan Jensen

Index 361

D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

Preface

In the past several years, DNA microarray technology has attracted tremendous interest in both the scientific community and in industry. With its ability to simultaneously measure the activity and interactions of thousands of genes, this modern technology promises unprecedented new insights into mechanisms of living systems. Currently, the primary applications of microarrays include gene discovery, disease diagnosis and prognosis, drug discovery (pharmacogenomics), and toxicological research (toxicogenomics).

Typical *scientific tasks* addressed by microarray experiments include the identification of coexpressed genes, discovery of sample or gene groups with similar expression patterns, identification of genes whose expression patterns are highly differentiating with respect to a set of discerned biological entities (e.g., tumor types), and the study of gene activity patterns under various stress conditions (e.g., chemical treatment). More recently, the discovery, modeling, and simulation of regulatory gene networks, and the mapping of expression data to metabolic pathways and chromosome locations have been added to the list of scientific tasks that are being tackled by microarray technology.

Each scientific task corresponds to one or more so-called *data analysis tasks*. Different types of scientific questions require different sets of data analytical techniques. Broadly speaking, there are two classes of elementary data analysis tasks, *predictive modeling* and *pattern-detection*. Predictive modeling tasks are concerned with learning a classification or estimation function, whereas pattern-detection methods screen the available data for interesting, previously unknown regularities or relationships.

A plethora of sophisticated methods and tools have been developed to address these tasks. However, each of these methods is characterized by a set of idiosyncratic requirements in terms of data pre-processing, parameter configuration, and result evaluation and interpretation. To optimally design and analyze microarray experiments, researchers and developers need a

D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

sufficient overview of existing methodologies and tools and a basic understanding of how to apply them.

We believe that one significant barrier to the widespread effective and efficient use of microarray analysis methods and tools is a lack of a clear understanding of how such techniques are used, what their merits and limitations are, and what obstacles are involved in deploying them. Our goal in developing this book was to address this issue, by providing what is simultaneously a *design blueprint*, *user guide*, and *research agenda* for current and future developments in the field.

As design blueprint, the book is intended for life scientists, statisticians, computer experts, technology developers, managers, and other professionals who will be tasked with developing, deploying, and using microarray technology including the necessary computational infrastructure and analytical tools.

As a user guide, the book seeks to address the requirement of scientists and researchers to gain a basic understanding of microarray analysis methodologies and tools. For these users, we seek to explain the key concepts and assumptions of the various techniques, their conceptual and computational merits and limitations, and give guidelines for choosing the methods and tools most appropriate for the analytical task at hand. Our emphasis is not on a complete and intricate mathematical treatment of the presented analysis methodologies. Instead, we aim at providing the users with a clear understanding and practical know-how of the relevant methods so that they are able to make informed and effective choices for data preparation, parameter setting, output post-processing, and result interpretation and validation. For methodologies where free software exists we will also provide practical tips for obtaining and using the tools.

As a research agenda, this volume is intended for students, teachers, researchers, and research managers who want to understand the state of the art of the presented methodologies and the areas in which gaps in our knowledge demand further research and development. To this end, our aim was to maintain the readability and accessibility of a textbook throughout the chapters, rather than compiling a mere reference manual. Therefore, considerable effort was made to ensure that the presented material, which stresses the applied aspects of microarray analysis, is supplemented by rich literature cross-references to more foundational work.

Clearly, we cannot expect to do justice to all three goals in a single book. However, we do believe that we have succeeded in taking useful steps toward each goal. In doing so, we hope to advance the understanding of both the methodologies and tools needed to analyze microarray data, and the implications for future developments of microarray technology and its support technologies.

The design and subsequent analytical examination of microarray experiments rests on the scientific expertise of the experimenters, their

D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

knowledge of the relevant microarray technology and experimental protocols, and their understanding of analysis methods and tools. The available machinery of microarray analysis methods ranges from classical statistical approaches, to machine learning techniques and to methods from artificial intelligence. Hence, the preparation of this book must draw upon the experts from many diverse subfields in mathematics and computer science. In developing this volume, we have assembled a distinguished set of authors, each recognized as an authority in one or more of these fields. We have asked these authors to present a selected set of state-of-the-art methodologies and tools for analyzing microarrays from a highly practical, user-oriented perspective, emphasizing the how-to aspects of the presented techniques. To support the research agenda of this book, we have also asked the authors to identify where future developments are likely to take place and to provide a rich set of pointers to theoretical works underpinning the presented methods. The result, we hope, is a book that will be valuable for a long time, as summary of where we are, as a practical user guide for making informed choices on actual microarray analysis projects, and as roadmap for where we need to go in order to improve and further develop future microarray analysis technology.

This book contains one introductory chapter and 19 technical chapters, dealing with specific methods or class of methods. As illustrated in Table 1, the technical chapters are roughly grouped into two broad categories, namely *data preparation* and *exploratory data analysis* respectively. Partitioning the chapters into these areas largely mirrors the current state of the art in the field. Different protocols, experimental conditions, analysis goals, data complexity, and sources of systematic variation in microarray experiments normally require different ways for selecting and preparing the raw data obtained from the detection devices. These methods range from missing value imputation and normalization to feature subset selection and data integration. Collectively, we refer to these methods as data preparation or *pre-processing* techniques. Once the final format for the data is achieved, data exploration or analysis can commence. This part of the data analysis process is referred to as *exploratory data analysis*. Typical exploratory analysis tasks include *classification* (or class prediction), *clustering* (or automatic classification), correlation and association analysis, and others.

As much as possible, the chapters are presented in an order that reflects the overall data analysis process. In Chapter 1, we provide an introduction to microarray analysis with the aim of (a) providing an easy-to-understand description of the entire process, and (b) establishing a common terminology. First, we recapitulate the biological and technological background of microarray hybridization experiments. This includes the main types of arrays that exist, aspects of their protocols, and what kind of quantities they are measuring. Second, we categorize the classes of questions life scientists hope to answer with microarray experiments, and what kind of

D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

analytical tasks they imply. Third, we describe the entire process from the inception of a scientific question or hypothesis, to the design and execution of a microarray experiment, and finally to data preparation, analysis, and interpretation. We then discuss some of the conceptual and practical difficulties the experimenter faces when choosing and applying specific data analysis techniques. The remaining chapters are intended to shed more detailed light on these issues.

Table 1. *Roadmap to the content of the book.*

Part	Topic	Chapter#
	Introduction: Overview of microarray analysis process	1
Pre-processing	Foundations, issues, and methods	2
	Missing value imputation	3
	Error handling and normalization	4
	Singular value decomposition, principal component analysis	5
	Feature selection: established and recent techniques	6
Classification	Statistical foundations and methods	7
	Bayesian networks	8
	Support vector machines	9
	Weighted flexible compound covariate method and decision trees	10
	Artificial neural networks	11
	k-nearest neighbor and genetic algorithms	12
Cluster Analysis	Overview and review of some methods	13
	Hierarchical clustering methods	14
	Self-organizing maps	15
	Other non-hierarchical methods	16
Other	Correlation and association analysis methods	17
	Functional interpretation analytical results	18
Tools	Systematic review of free and commercial software	19
	Managing microarray data analysis: workflow and process	20

Chapter 2 addresses the issue of data pre-processing in microarray analysis in general. It is written for the newcomer to this field and explains the basic concepts and provides a useful vocabulary. It discusses the motivation for normalization, data centralization, data re-scaling, and missing value imputation. This chapter represents an introduction to the Chapters 2 to 5.

Chapter 3 presents three different methods for missing value imputation in microarray data. This includes a *k-nearest-neighbor* approach, a method based on *singular value decomposition*, and *row averaging*. Practical guidelines are presented for using publicly available free software tools.

D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

Chapter 4 discusses various sources of errors in microarray data, and then proceeds with a detailed discourse on normalization. In contrast to Chapter 2, the focuses is on mathematical considerations.

Chapter 5 is concerned with a major problem in microarray data analysis – the so-called *large-p-small-n problem* also known as the *curse of dimensionality*. This refers to the fact that for many microarray experiments the number of variables (genes) exceeds the number of observations (samples) by a factor of 10, 100, 1000, or more. Feature selection and dimension reduction methods refer to techniques designed to deal with this "curse". The chapter discusses the use – and misuse – of *singular value decomposition* and *principal component analysis*.

Chapter 6 is a survey of several important feature selection techniques used to ward off the curse of dimensionality. First, it presents classic *filter* and *wrapper* approaches and some recent variants of explicit feature selection. Second, it outlines several feature weighting techniques including WINNOW and Bayesian feature selection. Third, towards the end, the chapter describes some recent work on feature selection for clustering tasks, a subject that has been largely neglected.

Chapter 7 discusses statistical issues arising in the classification of gene expression data. This chapter introduces the statistical foundations of classification. It provides an overview of traditional classifiers, such as linear discriminant analysis and nearest neighbor classifiers, in the context of microarray analyses. The general issues of feature selection and classifier performance assessment are discussed in detail.

Chapter 8 looks at Bayesian networks for the classification of microarray data. It introduces the basic concept of this approach, and reports on a study where the performance of Bayesian networks was compared with other state-of-the-art classifiers.

Chapter 9 describes a classification method that has been gaining increasing popularity in the microarray arena – *support vector machines* (SVMs). It provides an informal theoretical motivation of SVMs, both from a geometric and algorithmic perspective. Instead of focusing on mathematical completeness, the intention of this chapter is to provide the practitioner with some "rules of thumb" for using SVMs in the context of microarray data. Finally, pointers to relevant, publicly available free software resources are given.

Chapter 10 reports on a recent case study of gene expression analysis in lung cancer. The authors describe the *weighted flexible compound covariate* method for classifying the microarray data. They also demonstrate how this relatively new method is related to decision trees.

Chapter 11 deals with a widely used machine learning technique called *artificial neural networks* (ANNs). The authors describe the application of ANNs to micrarray classification task. They discuss how a principal component analysis, *cross-validation* and *random permutation tests* can be

D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

employed to improve and evaluate the predictive performance of ANNs. The problem of extracting important genes from a constructed ANN is also addressed.

Chapter 12 represents the last chapter on classification. It presents the *k-nearest-neighbor* strategy and *genetic algorithms* for classifying microarray data. It discusses the general motivation and the concepts of these methods, and demonstrates their performance on microarray data sets. The authors provide references to publicly available free software resources.

Chapter 13 provides an overview of the major types of clustering problems and techniques for microarray data. It focuses on crucial design and analytical aspects of the clustering process. The authors provide some important criteria for selecting clustering methods. Furthermore, the chapter describes a scheme for evaluating clustering results based on their relevance and validity (both computational and biological).

Chapter 14 addresses hierarchical clustering methods in the context of microarray data. The discussed methods include hierarchical clustering methods, including *adaptive single linkage clustering*, a new method designed to provide adaptive cluster detection while maintaining scalability. Furthermore, the chapter provides examples using both simulated and real data.

Chapter 15 presents *self-organizing maps* (SOMs) for clustering microarray data. It discusses question such as: How do these models work? Which are their advantages and limitations? Which are the alternatives? In answering these questions, this chapter constitutes a rich source of practical guidelines for using SOMs to analyze microarray data.

Chapter 16 examines a number of non-hierarchical clustering algorithms for microarray analysis, namely *cluster affinity search technique*, the famous *k-means* technique, *partitioning around medoids*, and *model-based clustering*. The chapter puts emphasis on the practical aspects of these algorithms, such as guidelines for parameter setting, the specific algorithmic properties, and practical tips for implementation.

Chapter 17 addresses correlation and association analysis methods. It addresses questions that should help the user to assess the limitations and merits of these methods, such as: How to statistically measure the strength between two variables and test their significance? What is correlation, what is association? Which conclusions do correlation and association analysis allow in the context of microarray data?

Chapter 18 discusses the global functional interpretation of gene expression experiments. After a researcher has found differentially expressed genes using one of the above described methods, he must face the challenge of translating his results into a better understanding of the underlying biological phenomena. This chapter shows how this can be achieved.

Chapter 19 provides an overview of both publicly available, free software and commercial software packages for analyzing microarray data. The aim

D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

of this review is to provide an overview of various microarray software categorized by their function and characteristics. This review should be a great help for those who currently consider obtaining such software.

Finally, Chapter 20 describes the microarray data analysis from a process perspective, highlighting practical issues such as project management and workflow considerations.

The book is designed to be used by the practicing professional tasked with the design and analysis of microarray experiments or as a text for a senior undergraduate- or graduate level course in analytical genetics, biology, bioinformatics, computational biology, statistics and data mining, or applied computer science. In a quarter-length course, one lecture can be spent on each chapter, and a project may be assigned based on one of the topics or techniques discussed in a chapter. In a semester-length course, some topics can be covered in greater depth, covering more of the formal background of the discussed methods. Each chapter includes recommendations for further reading. Questions or comments about the book should be directed to the editors by e-mail under *dp.berrar@ulster.ac.uk*, *w.dubitzky@ulster.ac.uk*, or *granzow@4T2consulting.de*. For further details on the editors, please refer to the following URL:

http://www.infj.ulst.ac.uk/~cbbg23/interests.html.

Daniel Berrar

Werner Dubitzky

Martin Granzow

D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

Index

A
accuracy, 42, 68, 69, 71, 72, 295, 302
activation function, 207
adaptive single linkage clustering, 246
adjusted Rand index, 280, 286
affinity, 277
Affymetrix, 78, 82, 88, 160, 307, 310, 313
alien sequences, 57
all-pairs, 181, 182
alternative polyadenylation, 327
alternative promoter usage, 327
alternative splicing, 327
analysis of variance, 59
ANN. *See* artificial neural network
annotation software, 341
ANOVA. *See* analysis of variance
AP. *See* all-pairs
application data, 44
application set, 27, 44
Apriori algorithm, 297, 301, 302
array profile, 22, 23, 28, 32
artificial neural network, 201, 215
association, 37, 40, 42, 43, 45, 289, 290, 292-298, 300-304
association rules, 294, 297, 303
Audic and Claverie's test, 316
audit trail, 89, 346, 349, 350
auditability, 345
average linkage clustering, 246, 250, 253
average linkage hierarchical clustering, 133
B
backward elimination, 112
bagging, 355
Bayes classifier, 159, 161
Bayes decision rule, 197
Bayes rule, 138, 139, 140
Bayesian belief networks. See Bayesian network
Bayesian feature selection, 111, 125, 129
Bayesian information criterion, 285
Bayesian network, 120, 150, 151, 153, 154, 156-163
Bayesian optimal classifier, 126
best-first search, 112
bias, 50-52, 54, 57, 59, 77, 78, 80, 84- 86, 88, 136, 143-145, 148
bias-variance trade-off, 144
BIC. *See* Bayesian information criterion
biclustering, 234
binomial distribution, 313, 314
bioinformatics, 91
black holes, 53
blessing of smoothness, 167, 184
Bonferroni correction, 217
boosting, 146, 148
bottom-up, 246, 252
C
calibration, 208, 210
CART. *See* Classification And Regression Tree
CAST. *See* cluster affinity search technique
category utility, 128
CCM. *See* compound covariate method
cDNA, 9, 10, 14, 15, 18, 46, 51, 52, 56, 132, 133, 136, 148, 187, 200, 267
cell-cycle data, 96, 97, 100, 103
centralization, 55, 56, 57, 58, 60, 61
chaining effect, 251, 252
Chebychev metric, 240, 243
cheminformatics, 348, 357
chi-square, 294, 297
chi-square test, 314-316
chromosome, 219, 221, 222, 223, 224, 225, 226
classification, 100, 106, 110, 113, 114, 118, 119, 122, 124, 125, 127, 128, 132, 133-139, 141, 142, 143-148, 150-154, 157-163, 166-168, 174-177, 179-185, 201-203, 205-215
Classification And Regression Tree, 66, 142 198
Classification Rule with Unbiased Interaction Selection and Estimation, 198.
classification task, 40
classification tree. *See* decision tree
classifier, 134-138, 140-146, 202, 206, 210, 211
CLICK algorithm, 129
cluster, 133, 135, 230-243, 274, 277-282, 284, 287
cluster affinity search technique, 274, 277

D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

clustering, 103, 105, 106, 110, 111, 127-129, 186, 196, 199, 230, 231, 233, 235, 238, 242, 274, 282
clustering task, 41
codebook, 181, 182
coefficient of variation, 87
committee, 159, 210, 212, 213
compactness, 232
complete linkage clustering, 246, 250, 252
compound covariate method, 186, 188, 192, 193, 199
comprehensive software, 330, 331, 333, 334, 341
confidence, 13, 30, 34, 42, 175, 181, 182, 295-297
consensus trees, 244
contingency table, 293-296
continuous, 99, 103, 106, 112, 114, 128
control, 11
convex admissible, 235
correlation, 36, 41, 96, 100-103, 107, 290-294, 299, 300, 303
correlation coefficient, 276, 278
covariance, 81, 127, 140, 219, 248, 251
coverage, 42, 295, 302
cross-validation, 43, 44, 119, 120, 122-124, 141-146, 202, 204, 209-214
curse of dimensionality, 110, 167, 184
Cy3, 8, 29
Cy5, 8, 29

D

DAG. *See* directed-acyclic graph
data filtering, 47, 49, 61, 62
data inspection, 47, 60, 61
Data inspection, 60
data mining, 294, 297
data transformation, 47, 59
decision rule, 118, 126
decision surface, 209
decision tree, 118, 150, 158, 161, 354, 355
dendrogram, 250-259
dependence tree, 120, 121
DHP. *See* Direct Hashing and Pruning
diagonal linear discriminant analysis, 140
diagonal quadratic discriminant analysis, 140
DIC. *See* Dynamic Itemset Counting
differential display, 7
dimensional reduction, 202
dimensionality reduction, 95, 105, 330, 331
Direct Hashing and Pruning, 297
directed-acyclic graph, 152
discretization, 117, 154
discriminant analysis, 133, 134, 135, 139, 141, 146
dissimilarity, 275, 276, 279
DLDA. See diagonal linear discriminant analysis
DNA, 48, 58
double self-organizing map, 268
DQDA. See diagonal quadratic discriminant analysis
Dunn's index, 232, 238
dyes, 8, 29, 51, 56, 57, 59
dye-swap, 78, 85, 86, 89
Dynamic Itemset Counting, 297

E

early stopping, 211
eigenarray, 94
eigenassays, 94, 96, 98, 100, 102
eigengenes, 67, 68, 71, 74, 94, 95-101, 103, 104
eigenvalues, 67, 69
eigenvector, 127
elbow, 98, 103, 104
EM. *See* expectation maximization
empirical error, 174
entropy, 115, 128, 129, 189
epochs, 208, 210, 211, 271
ESTs. *See* expressed sequence tags
Euclidean distance, 67, 141, 189, 193, 220, 248, 264, 276, 278
exons, 5
expectation maximization, 68, 156, 284
experimental study, 298
experimenter bias, 77
explicit feature selection, 111
expressed sequence tags, 327
expression profile, 16, 21, 22, 28, 36, 40, 92, 95, 96, 100, 105
extension/accessory software, 330
external criterion analysis, 280

F

false positive, 217
false positive rate, 190
fast self-organizing feature map, 268
feature reduction, 217

D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

feature selection, 110, 111, 113, 119, 122, 124-126, 128, 129, 134, 136, 141-143, 145, 146
feature weighting, 111
feed-forward network, 201
figure of merit, 281
filter, 111, 112, 118, 122, 123, 142, 176
filtering, 49, 61, 62, 63
Fisher's exact test, 314-316, 338
fitness, 219, 222-224
fluorescent, 53
FOM. *See* figure of merit
forward selection, 112
FP-growth, 297
FPR. *See* false positive rate
G
GA. *See* genetic algorithm
Gaussian distribution, 112, 114
Gaussian kernel, 173, 183
Gaussian maximum likelihood discriminant rule, 139
Gaussian mixture model, 282
Gaussian quadratic classifier, 123
GCS. *See* growing cell structures
gene profile, 16, 21, 23, 35, 37, 40
generalization error, 138, 143, 144, 174
generalized linear model, 124
genetic algorithm, 219
genetic risk group, 302
Gibbs sampler, 126
GLIM. *See* generalized linear model
global normalization, 81
gradient descent algorithm, 208
Gram-Schmidt orthogonalization, 93
group average linkage clustering, 250
growing cell structure, 234, 269
growing self-organizing map, 269
GSOM. *See* growing self-organizing map
H
heterogeneity, 232, 235, 241
heuristic-based clustering, 277, 280
hidden layer, 206
hierarchical agglomerative clustering, 246, 249
hierarchical clustering, 52, 58, 62, 65, 133, 217, 232, 233, 236, 244, 246, 247, 250, 251, 259, 278, 284
hill-climbing, 112, 120, 156
hinge loss function, 173
homogeneity, 281
housekeeping genes, 57, 79
hybridization, 2, 7-10, 12, 14, 26-28, 30, 34, 35, 39, 49, 51, 53, 58, 59, 61, 62, 77, 78, 80, 83, 85, 86, 88
hypergeometric distribution, 313-315
hyperplane, 168-170, 171, 172, 175, 176
I
IGG. *See* incremental grid growing neural network
ill-posed problem, 167
image analysis, 49, 54, 56, 61-63
image analysis software, 328, 329
image processing, 53, 77
incremental grid growing neural network, 269
induction, 110-112, 118, 129
inductive bias, 119, 123
Info Score. *See* mutual information scoring
information gain, 114, 115, 117, 118, 122, 129
input layer, 206
intensity dependent methods, 83
intensity independent methods, 81
internal criterion analysis, 280
Internet, 49
intra-cluster diameter metrics, 232
introns, 5
isolation, 232
K
Kendall's tau, 291, 292, 293, 294, 299
kernel function, 172, 173
k-means, 65, 217, 233, 236, 261, 265, 274, 277-279, 284
k-nearest neighbor, 179, 194, 217, 222
KNN. *See k*-nearest neighbor
KNNimpute, 66-70, 72-74
Kohonen map. *See* self-organizing map
Kohonen Self-Organizing Feature Map. *See* self-organizing map
Kullback-Leibler divergence, 115
L
L_1 measures, 232
Laboratory Information Management System, 339
landing lights, 53
large *p*, small *n*-problem, 137
lattice machine, 42, 43
LDA. *See* linear discriminant analysis
learning, 110-112, 118, 119, 124, 125, 129, 201, 202, 209, 210, 215
learning cycle, 264

D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

learning rate, 264, 265, 269
learning set, 44, 135, 136, 138, 140, 141, 143-145, 216, 217, 220
learning vector quantization, 139
least squares estimates, 66
leave-one-out cross validation, 161
leave-one-out cross-validation, 45, 123, 124, 141, 144, 194
leukemia, 110, 113, 115, 117, 118, 123, 160-163, 166, 174, 179, 180, 218, 234, 242, 243, 262, 265, 266
lift, 145
likelihood, 66, 113, 121, 155, 156, 158, 249, 250-252, 284, 285
linear discriminant analysis, 140, 217
locally weighted regression, 56, 57, 84
log bases, 55
logistic linear classifier, 123
logistic regression, 139
logistic regression classifier, 123, 124
LOOCV. *See* leave-one-out cross-validation
loss function, 137, 138, 146
Lowess. *See* locally weighted regression
Lowess smoothing, 84, 85

M

M vs. *A* plot, 57, 60
M-A plot, 83, 84, 85, 86
machine learning, 110, 111, 202, 234, 289, 299, 357
macro-cluster, 271
Mahalanobis distance, 140
Manhattan metric, 240, 243
margin, 167, 170, 171-174, 177, 179, 184
market basket analysis, 294, 301, 303
Markov blanket, 115-118, 122, 129, 158, 159
Markov Chain Monte Carlo, 126
matrix stability theory, 127
maximum association algorithm, 302
Maximum Entropy Discrimination, 126
MDL. *See* minimum description length
mean squared error, 207, 211, 212
medoids, 279
membership, 234, 241, 244
metric variables, 290, 291, 292
Metropolis-Hasting algorithm, 126
MIAME, 48, 63
Microarray, 47, 48
Microarray Gene Expression Data Group, 319
microarray software, 326, 339
microarrayers, 20
minimum description length, 156
Minimum Information About a Microarray Experiment, 319, 331
misclassification rate, 138
missing value, 13, 28, 31, 32, 65
mixture model, 114, 115
MLP. *See* multilayer perceptron
model application, 43, 44
model construction, 2, 43, 44
model validation, 43
model verification, 43
model-based clustering, 246, 248, 249, 250-252, 274, 277, 279, 282, 284, 285
momentum coefficient, 208, 209
monotone admissible, 235
mRNA, 56, 57, 111, 133, 134, 136
MSE. *See* mean-squared error
multiclass, 180, 181, 183, 184
Multidimensional Scaling, 108, 127, 130
multilayer perceptron, 161, 206
multivariate analysis, 52, 58
mutual information, 154, 155
mutual-information scoring, 187, 188

N

naive Bayes method, 139
nearest neighbor, 66, 123, 134, 137, 139, 141, 142, 145, 146, 148
nearest neighbors, 68, 70, 73, 74
nearest shrunken centroids, 217
near-optimal, 224, 225, 226
near-optimal chromosome, 224, 226
neural networks. *See* artificial neural networks.
neurone, 263
niche, 221, 222, 224
noise, 97
noise-tolerant, 235
noisy data, 74
nominal variables, 290
normality, 55, 61
normalization, 54, 56-59, 76-89
northern blots, 7
NP-hard, 156, 157

O

observational study, 298
oligonucleotide chips, 132, 133
one-versus-all, 181
ORDERED-FS, 122, 123, 129
ordinal, 99, 106

D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

ordinal variables, 290
orthogonality, 203
output layer, 206, 207
OVA. *See* one-versus-all
over-expressed, 154, 155, 160
overfitting, 44, 198, 207, 210, 211, 213, 285

P

paired-slide normalization, 85
PAM. *See* partitioning around medoids
partial derivative, 212
partitioning around medoids, 274
pathway, 308, 317-320
pathway reconstruction software, 330, 338
pattern recognition, 111, 217
pattern-detection, 1, 40
PCA. *See* principal component analysis
Pearson correlation, 67, 128
Pearson's contingency coefficient, 291, 293
Pearson's product-moment correlation coefficient, 290, 291, 294
Pearson's rho. *See* Pearson's product-moment correlation coefficient
perceptron, 206, 207, 215
permutation test, 178
P-metric, 154, 160, 176
point proportion admissible, 235
Poisson distribution, 316
polymerase chain reaction, 45
population, 50, 219, 221
Power, 50
prediction set. *See* application set
predictive modeling, 1, 27
predictor. See classifier. See classifier. See classifier. See classifier. See classifier
pre-processing, 47, 49, 61, 63, 77
principal component analysis, 65, 91, 93, 127, 203, 214
probe, 8-10, 12, 15, 18, 19, 26, 30, 31
projection, 101-103, 107
projection pursuit, 139
protein, 48
prototype, 263- 269

Q

QDA. *See* quadratic discriminant analysis
quadratic discriminant analysis, 139, 140
QUEST. *See* Quick, Unbiased, Efficient Statistical Tree
Quick, Unbiased, Efficient Statistical Tree, 198

R

Rand index. *See* adjusted Rand index
random error, 78
random permutation tests, 202, 211
randomization, 50, 51
randomized inference method, 66
ratio statistics, 87
receiver operating characteristic, 159
recursive feature elimination, 176
recursive partitioning, 217
reference, 10-14, 17, 34
reference chip, 82, 88
regression methods, 81
regularization, 166, 173, 174, 183
regulatory networks, 112, 299
relevance machine, 42
reporter molecules, 8, 11, 12
re-scaling, 55, 58, 61
residual effects, 50
response function, 124-126
reverse-engineering of the genetic networks, 300
reverse-labelling, 51
RFE. *See* recursive feature elimination
RMS. *See* root mean squared
RNA, 52, 56
ROC. *See* receiver operating characteristic
root mean squared, 68
roulette-wheel selection, 222
rule induction, 354

S

S2N. *See* signal-to-noise
SAM. *See* significance analysis of microarrays
Sample preparation, 77
SANN. *See* self-adaptive and incremental learning neural network
scalability, 247
scanner settings, 77
scatter plots, 100, 102
scoring metric, 156, 160
scree plots, 98
search tree, 119-121
second order polynomial kernel, 173
self-adaptive and incremental learning neural network, 268
self-organizing map, 66, 108, 217, 233, 333

D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

self-organizing neural network, 261
sensitivity, 159
separability score method, 301
separation, 281
sequence data, 48
serial analysis of gene expression, 7
signal-to-noise, 175, 176, 178
signature, 22
significance analysis of microarrays, 187, 188
significance threshold, 62
silhouette, 232, 237, 241-243, 281
similarity, 275-279, 281
simplified fuzzy ARTMAP, 234
single linkage clustering, 246, 247, 250, 251-259
single slide normalization, 83
singular value, 92, 93, 95, 96, 98, 99, 104
singular value decomposition, 65-67, 91, 214
singular value spectrum, 98, 99
smoothness, 167, 173
soft margin, 171
SOM. *See* self-organizing map
spatial effects, 77
Spearman's rho, 291-294, 299
specific analysis software, 330, 333
specificity, 159
spherical model, 279, 283, 284, 286
spiking controls, 57
splicing, 5, 6, 19
standardization, 145, 146
statistical design, 50
statistical software, 335
statistics, 111, 112
step size, 208, 209
subsymbolic, 43
sum of squares, 232
supervised learning, 111, 133, 135, 136, 217, 218
supervised network self- organizing map, 268
support, 42, 295, 296, 297
support vector, 172, 217, 218
support vector machine, 42, 43, 119, 126, 134, 161, 166, 167, 185, 202, 217
SVD. *See* singular value decomposition
SVDimpute, 66, 67, 68, 71, 72, 73
SVDPACKC, 107
SWISS-PROT, 318, 320, 322
symbolic, 23, 39, 43, 46
systematic error. *See* bias
systems biology, 94-96, 100, 233

T

TAN. *See* tree-augmented naive Bayes classifier
target, 5, 10-12, 14, 19, 34, 39
terminology, 47, 51, 55
test data, 43
test set, 44, 143-145, 211
Tikhonov regularization principle, 174
time series, 97, 104-106
top-down, 246, 252
top-ranked genes, 224-226
total intensity methods, 79
total least squares, 82
total risk, 138
toxicogenomics, 223
training, 111, 112, 114, 119, 121, 123, 125-127, 155, 156, 159, 160
training data, 43
training set, 44
transcriptional response, 92, 95-97, 100, 101-106
transformation, 54, 55, 58-61, 63
tree-augmented naive Bayes classifier, 159
t-test, 187, 188, 190-192, 195, 199
Tukey's compound covariate, 217
Turnkey system, 330, 331
two-sample *t*-test, 217
type I error, 50

U

uncertainty, 244
under-expressed, 154, 155, 160
univariate mixture model, 113
unsupervised learning, 111, 127, 135, 136, 217

V

validation, 52, 54
validation data, 27, 43, 44
validation set, 27, 44, 45, 144, 208, 210, 211
variance, 136, 141, 143, 144, 148
variation, 47, 49, 56, 60
Voting Gibbs classifier, 126

W

Ward's method, 251
weight decay, 208, 209, 211
weight vector, 125
weighted flexible compound covariate method, 188, 193
weighted gene analysis, 187, 188, 200

D. Berrar, W. Dubitzky, and M. Granzow (Eds.), *A Practical Approach to Microarray Data Analysis*, Copyright 2002, Kluwer Academic Publishers, Boston/Dordrecht/London.

weighted voting average, 179
well-posed problem, 167
WFCCM. *See* weighted flexible compound covariate method
WGA. *See* weighted gene analysis
Wilcoxon, 218
WINNOW, 111, 124, 125
work flow, 348, 349, 351, 353, 358, 359
wrapper, 111, 119, 122, 123, 128, 142, 176

Y

yeast, 50, 96, 104, 105, 262, 274, 278, 285, 286