
Contents

List of Figures	xxv
List of Tables	xxxv
Preface	xxxix
1 Introduction	1
1.1 Bioinformatics – an emerging discipline	1
2 The cell and its basic mechanisms	5
2.1 The cell	5
2.2 The building blocks of genomic information	13
2.2.1 The deoxyribonucleic acid (DNA)	13
2.2.2 The DNA as a language	19
2.2.3 Errors in the DNA language	23
2.2.4 Other useful concepts	24
2.3 Expression of genetic information	28
2.3.1 Transcription	30
2.3.2 Translation	32
2.3.3 Gene regulation	35
2.4 The need for high-throughput methods	36
2.5 Summary	37
3 Microarrays	39
3.1 Microarrays – tools for gene expression analysis	39
3.2 Fabrication of microarrays	41
3.2.1 Deposition	41
3.2.1.1 The Illumina technology	42
3.2.2 <i>In situ</i> synthesis	48
3.2.3 A brief comparison of cDNA and oligonucleotide technologies	55
3.3 Applications of microarrays	57
3.4 Challenges in using microarrays in gene expression studies	58

3.5	Sources of variability	63
3.6	Summary	67
4	Reliability and reproducibility issues in DNA microarray measurements	69
4.1	Introduction	69
4.2	What is expected from microarrays?	70
4.3	Basic considerations of microarray measurements	70
4.4	Sensitivity	72
4.5	Accuracy	73
4.6	Reproducibility	77
4.7	Cross-platform consistency	78
4.8	Sources of inaccuracy and inconsistencies in microarray measurements	82
4.9	The MicroArray Quality Control (MAQC) project	85
4.10	Summary	87
5	Image processing	89
5.1	Introduction	89
5.2	Basic elements of digital imaging	90
5.3	Microarray image processing	95
5.4	Image processing of cDNA microarrays	96
5.4.1	Spot finding	99
5.4.2	Image segmentation	100
5.4.3	Quantification	106
5.4.4	Spot quality assessment	111
5.5	Image processing of Affymetrix arrays	113
5.6	Summary	115
6	Introduction to R	119
6.1	Introduction to R	119
6.1.1	About R and Bioconductor	119
6.1.2	Repositories for R and Bioconductor	120
6.1.3	The working setup for R	121
6.1.4	Getting help in R	122
6.2	The basic concepts	122
6.2.1	Elementary computations	122
6.2.2	Variables and assignments	125
6.2.3	Expressions and objects	126
6.3	Data structures and functions	128
6.3.1	Vectors and vector operations	128
6.3.2	Referencing vector elements	131

6.3.3	Functions	133
6.3.4	Creating vectors	135
6.3.5	Matrices :	137
6.3.6	Lists	141
6.3.7	Data frames	141
6.4	Other capabilities	144
6.4.1	More advanced indexing	144
6.4.2	Missing values	145
6.4.3	Reading and writing files	148
6.4.4	Conditional selection and indexing	150
6.4.5	Sorting	151
6.4.6	Implicit loops	154
6.5	The R environment	159
6.5.1	The search path: <code>attach</code> and <code>detach</code>	159
6.5.2	The workspace	161
6.5.3	Packages	163
6.5.4	Built-in data	165
6.6	Installing Bioconductor	165
6.7	Graphics	167
6.8	Control structures in R	169
6.8.1	Conditional statements	170
6.8.2	Pre-test loops	171
6.8.3	Counting loops	172
6.8.4	Breaking out of loops	173
6.8.5	Post-test loops	173
6.9	Programming in R versus C/C++/Java	174
6.9.1	R is “forgiving” – which can be bad	174
6.9.2	Weird syntax errors	175
6.9.3	Programming style	179
6.10	Summary	182
6.11	Solved Exercises	183
6.12	Exercises	191
7	Bioconductor: principles and illustrations	193
7.1	Overview	193
7.2	The portal	194
7.2.1	The main resource categories	195
7.2.2	Working with the software repository	195
7.3	Some explorations and analyses	197
7.3.1	The representation of microarray data	197
7.3.2	The annotation of a microarray platform	199
7.3.3	Predictive modeling using microarray data	203
7.4	Summary	205

8 Elements of statistics	207
8.1 Introduction	207
8.2 Some basic concepts	208
8.2.1 Populations versus samples	208
8.2.2 Parameters versus statistics	209
8.3 Elementary statistics	211
8.3.1 Measures of central tendency: mean, mode, and median	211
8.3.1.1 Mean	211
8.3.1.2 Mode	212
8.3.1.3 Median, percentiles, and quantiles	213
8.3.1.4 Characteristics of the mean, mode, and median	214
8.3.2 Measures of variability	215
8.3.2.1 Range	215
8.3.2.2 Variance	216
8.3.3 Some interesting data manipulations	218
8.3.4 Covariance and correlation	219
8.3.5 Interpreting correlations	223
8.3.6 Measurements, errors, and residuals	230
8.4 Degrees of freedom	231
8.4.1 Degrees of freedom as independent error estimates	232
8.4.2 Degrees of freedom as number of additional measurements	233
8.4.3 Degrees of freedom as observations minus restrictions	233
8.4.4 Degrees of freedom as measurements minus model parameters	234
8.4.5 Degrees of freedom as number of measurements we can change	234
8.4.6 Data split between estimating variability and model parameters	235
8.4.7 A geometrical perspective	235
8.4.8 Calculating the number of degrees of freedom	236
8.4.8.1 Estimating k quantities from n measurements	236
8.4.9 Calculating the degrees of freedom for an $n \times m$ table	237
8.5 Probabilities	241
8.5.1 Computing with probabilities	243
8.5.1.1 Addition rule	243
8.5.1.2 Conditional probabilities	244
8.5.1.3 General multiplication rule	247
8.6 Bayes' theorem	247
8.7 Testing for (or predicting) a disease	250
8.7.1 Basic criteria: accuracy, sensitivity, specificity, PPV, NPV	251

8.7.2	More about classification criteria: prevalence, incidence, and various interdependencies	253
8.8	Summary	257
8.9	Solved problems	257
8.10	Exercises	258
9	Probability distributions	261
9.1	Probability distributions	261
9.1.1	Discrete random variables	262
9.1.2	The discrete uniform distribution	265
9.1.3	Binomial distribution	266
9.1.4	Poisson distribution	275
9.1.5	The hypergeometric distribution	278
9.1.6	Continuous random variables	281
9.1.7	The continuous uniform distribution	282
9.1.8	The normal distribution	283
9.1.9	Using a distribution	287
9.2	Central limit theorem	291
9.3	Are replicates useful?	292
9.4	Summary	294
9.5	Solved problems	295
9.6	Exercises	296
10	Basic statistics in R	299
10.1	Introduction	299
10.2	Descriptive statistics in R	300
10.2.1	Mean, median, range, variance, and standard deviation	300
10.2.2	Mode	304
10.2.3	More built-in R functions for descriptive statistics	305
10.2.4	Covariance and correlation	307
10.3	Probabilities and distributions in R	308
10.3.1	Sampling	308
10.3.2	Empirical probabilities	309
10.3.3	Standard distributions in R	315
10.3.4	Generating (pseudo-)random numbers	316
10.3.5	Probability density functions	316
10.3.6	Cumulative distribution functions	317
10.3.7	Quantiles	319
10.3.7.1	The normal distribution	321
10.3.7.2	The binomial distribution	324
10.3.8	Using built-in distributions in R	326
10.4	Central limit theorem	330
10.5	Summary	336

10.6 Exercises	337
11 Statistical hypothesis testing	339
11.1 Introduction	339
11.2 The framework	340
11.3 Hypothesis testing and significance	343
11.3.1 One-tailed testing	344
11.3.2 Two-tailed testing	348
11.4 "I do not believe God does not exist"	350
11.5 An algorithm for hypothesis testing	352
11.6 Errors in hypothesis testing	353
11.7 Summary	357
11.8 Solved problems	358
12 Classical approaches to data analysis	361
12.1 Introduction	361
12.2 Tests involving a single sample	362
12.2.1 Tests involving the mean. The t distribution.	362
12.2.2 Choosing the number of replicates	368
12.2.3 Tests involving the variance (σ^2). The chi-square distribution	372
12.2.4 Confidence intervals for standard deviation/variance	376
12.3 Tests involving two samples	377
12.3.1 Comparing variances. The F distribution.	377
12.3.2 Comparing means	383
12.3.2.1 Equal variances	386
12.3.2.2 Unequal variances	389
12.3.2.3 Paired testing	389
12.3.3 Confidence intervals for the difference of means $\mu_1 - \mu_2$	390
12.4 Summary	391
12.5 Exercises	394
13 Analysis of Variance – ANOVA	395
13.1 Introduction	395
13.1.1 Problem definition and model assumptions	395
13.1.2 The "dot" notation	399
13.2 One-way ANOVA	400
13.2.1 One-way Model I ANOVA	400
13.2.1.1 Partitioning the Sum of Squares	401
13.2.1.2 Degrees of freedom	403
13.2.1.3 Testing the hypotheses	403
13.2.2 One-way Model II ANOVA	407

13.3	Two-way ANOVA	410
13.3.1	Randomized complete block design ANOVA	411
13.3.2	Comparison between one-way ANOVA and randomized block design ANOVA	414
13.3.3	Some examples	415
13.3.4	Factorial design two-way ANOVA	419
13.3.5	Data analysis plan for factorial design ANOVA	424
13.3.6	Reference formulae for factorial design ANOVA	425
13.4	Quality control	425
13.5	Summary	428
13.6	Exercises	429
14	Linear models in R	433
14.1	Introduction and model formulation	433
14.2	Fitting linear models in R	435
14.3	Extracting information from a fitted model: testing hypotheses and making predictions	439
14.4	Some limitations of linear models	440
14.5	Dealing with multiple predictors and interactions in the linear models, and interpreting model coefficients	443
14.5.1	Details on the design matrix creation and coefficients estimation in linear models	445
14.5.2	ANOVA using linear models	447
14.5.2.1	One-way Model I ANOVA	448
14.5.2.2	Randomized block design ANOVA	453
14.5.3	Practical linear models for analysis of microarray data	454
14.5.4	A two-group comparison gene expression analysis using a simple <i>t</i> -test	455
14.5.5	Differential expression using the limma library of Bioconductor	457
14.5.5.1	Two group comparison with single-channel data	457
14.5.5.2	Multiple contrasts with single-channel data	459
14.6	Summary	460
15	Experiment design	463
15.1	The concept of experiment design	464
15.2	Comparing varieties	464
15.3	Improving the production process	466
15.4	Principles of experimental design	468
15.4.1	Replication	468
15.4.2	Randomization	471
15.4.3	Blocking	472

15.5	Guidelines for experimental design	472
15.6	A short synthesis of statistical experiment designs	474
15.6.1	The fixed effect design	475
15.6.2	Randomized block design	476
15.6.3	Balanced incomplete block design	476
15.6.4	Latin square design	477
15.6.5	Factorial design	477
15.6.6	Confounding in the factorial design	480
15.7	Some microarray specific experiment designs	481
15.7.1	The Jackson Lab approach	481
15.7.2	Ratios and flip-dye experiments	484
15.7.3	Reference design versus loop design	486
15.8	Summary	489
16	Multiple comparisons	491
16.1	Introduction	491
16.2	The problem of multiple comparisons	492
16.3	A more precise argument	499
16.4	Corrections for multiple comparisons	501
16.4.1	The Šidák correction	501
16.4.2	The Bonferroni correction	502
16.4.3	Holm's step-wise correction	503
16.4.4	The false discovery rate (FDR)	504
16.4.5	Permutation correction	505
16.4.6	Significance analysis of microarrays (SAM)	507
16.4.7	On permutation-based methods	508
16.5	Corrections for multiple comparisons in R	508
16.6	Summary	513
17	Analysis and visualization tools	515
17.1	Introduction	515
17.2	Box plots	516
17.3	Gene pies	520
17.4	Scatter plots	521
17.4.1	Scatter plots in R	525
17.4.2	Scatter plot limitations	526
17.4.3	Scatter plot summary	529
17.5	Volcano plots	529
17.5.1	Volcano plots in R	530
17.6	Histograms	533
17.6.1	Histograms summary	542
17.7	Time series	542
17.8	Time series plots in R	543

17.9 Principal component analysis (PCA)	550
17.9.1 PCA limitations	558
17.9.2 Principal component analysis in R	559
17.9.3 PCA summary	560
17.10 Independent component analysis (ICA)	563
17.11 Summary	564
18 Cluster analysis	567
18.1 Introduction	567
18.2 Distance metric	568
18.2.1 Euclidean distance	569
18.2.2 Manhattan distance	570
18.2.3 Chebychev distance	572
18.2.4 Angle between vectors	573
18.2.5 Correlation distance	573
18.2.6 Squared Euclidean distance	574
18.2.7 Standardized Euclidean distance	575
18.2.8 Mahalanobis distance	577
18.2.9 Minkowski distance	577
18.2.10 When to use what distance	578
18.2.11 A comparison of various distances	580
18.3 Clustering algorithms	581
18.3.1 <i>k</i> -means clustering	584
18.3.1.1 Characteristics of the <i>k</i> -means clustering	586
18.3.1.2 Cluster quality assessment	588
18.3.1.3 Number of clusters in <i>k</i> -means	593
18.3.1.4 Algorithm complexity	593
18.3.2 Hierarchical clustering	594
18.3.2.1 Inter-cluster distances and algorithm complexity	596
18.3.2.2 Top-down versus bottom-up	597
18.3.2.3 Cutting tree diagrams	599
18.3.2.4 An illustrative example	601
18.3.2.5 Hierarchical clustering summary	603
18.3.3 Kohonen maps or self-organizing feature maps (SOFM)	605
18.4 Partitioning around medoids (PAM)	614
18.5 Biclustering	616
18.5.1 Types of biclusters	617
18.5.2 Biclustering algorithms	619
18.5.3 Differential biclustering	620
18.5.4 Biclustering summary	621
18.6 Clustering in R	621
18.6.1 Partition around medoids (PAM) in R	629
18.6.2 Biclustering in R	632

18.7 Summary	632
19 Quality control	635
19.1 Introduction	635
19.2 Quality control for Affymetrix data	636
19.2.1 Reading raw data (.CEL files)	636
19.2.2 Intensity distributions	637
19.2.3 Box plots	639
19.2.4 Probe intensity images	639
19.2.5 Quality control metrics	641
19.2.6 RNA degradation curves	647
19.2.7 Quality control plots	649
19.2.8 Probe-level model (PLM) fitting. RLE and NUSE plots	654
19.3 Quality control of Illumina data	660
19.3.1 Reading Illumina data	660
19.3.2 Bead-summary data	663
19.3.2.1 Raw probe data import, visualization, and quality assessment using "beadarray"	663
19.3.2.2 Raw probe data import, visualization, and quality assessment using "lumi"	665
19.3.3 Bead-level data	669
19.3.3.1 Raw bead data import and assessment	669
19.3.3.2 Summarizing from bead-level to probe-level data	690
19.4 Summary	691
20 Data preprocessing and normalization	693
20.1 Introduction	693
20.2 General preprocessing techniques	694
20.2.1 The log transform	694
20.2.2 Combining replicates and eliminating outliers	696
20.2.3 Array normalization	698
20.2.3.1 Dividing by the array mean	701
20.2.3.2 Subtracting the mean	701
20.2.3.3 Using control spots/genes	703
20.2.3.4 Iterative linear regression	703
20.2.3.5 Other aspects of array normalization	704
20.3 Normalization issues specific to cDNA data	704
20.3.1 Background correction	704
20.3.1.1 Local background correction	704
20.3.1.2 Sub-grid background correction	705
20.3.1.3 Group background correction	705
20.3.1.4 Background correction using blank spots	705

20.3.1.5	Background correction using control spots	705
20.3.2	Other spot level preprocessing	706
20.3.3	Color normalization	706
20.3.3.1	Curve fitting and correction	708
20.3.3.2	LOWESS/LOESS normalization	710
20.3.3.3	Piece-wise normalization	713
20.3.3.4	Other approaches to cDNA data normalization	715
20.4	Normalization issues specific to Affymetrix data	715
20.4.1	Background correction	715
20.4.2	Signal calculation	718
20.4.2.1	Ideal mismatch	718
20.4.2.2	Probe values	719
20.4.2.3	Scaled probe values	720
20.4.3	Detection calls	721
20.4.4	Relative expression values	722
20.5	Other approaches to the normalization of Affymetrix data	722
20.5.1	Cyclic Loess	722
20.5.2	The model-based dChip approach	723
20.5.3	The Robust Multi-Array Analysis (RMA)	724
20.5.4	Quantile normalization	724
20.6	Useful preprocessing and normalization sequences	727
20.7	Normalization procedures in R	728
20.7.1	Normalization functions and procedures for Affymetrix data	728
20.7.2	Background adjustment and various types of normalization	734
20.7.3	Summarization	735
20.8	Batch preprocessing	738
20.9	Normalization functions and procedures for Illumina data	739
20.10	Summary	743
20.11	Appendix: A short primer on logarithms	746
21	Methods for selecting differentially expressed genes	749
21.1	Introduction	749
21.2	Criteria	751
21.3	Fold change	753
21.3.1	Description	753
21.3.2	Characteristics	754
21.4	Unusual ratio	756
21.4.1	Description	756
21.4.2	Characteristics	757
21.5	Hypothesis testing, corrections for multiple comparisons, and resampling	758

21.5.1 Description	758
21.5.2 Characteristics	760
21.6 ANOVA	760
21.6.1 Description	760
21.6.2 Characteristics	761
21.7 Noise sampling	761
21.7.1 Description	761
21.7.2 Characteristics	763
21.8 Model-based maximum likelihood estimation methods	764
21.8.1 Description	764
21.8.2 Characteristics	768
21.9 Affymetrix comparison calls	768
21.10 Significance Analysis of Microarrays (SAM)	769
21.11 A moderated t-statistic	770
21.12 Other methods	771
21.13 Reproducibility	772
21.14 Selecting differentially expressed (DE) genes in R	774
21.14.1 Data import and preprocessing	774
21.14.2 Fold change	775
21.14.3 Unusual ratio	777
21.14.4 Hypothesis testing, corrections for multiple comparisons, and resampling	779
21.15 Summary	791
21.16 Appendix	792
21.16.1 A comparison of the noise sampling method with the full-blown ANOVA approach	792
22 The Gene Ontology (GO)	795
22.1 Introduction	795
22.2 The need for an ontology	796
22.3 What is the Gene Ontology (GO)?	797
22.4 What does GO contain?	798
22.4.1 GO structure and data representation	798
22.4.2 Levels of abstraction, traversing the DAG, the "True Path Rule"	800
22.4.3 Evidence codes	801
22.4.4 GO coverage	803
22.5 Access to GO	803
22.6 Other related resources	806
22.7 Summary	807

23 Functional analysis and biological interpretation of microarray data	809
23.1 Over-representation analysis (ORA)	809
23.1.1 Statistical approaches	811
23.2 Onto-Express	814
23.2.1 Implementation	814
23.2.2 Graphical input interface description	815
23.2.3 Some real data analyses	820
23.2.4 Interpretation of the functional analysis results	825
23.3 Functional class scoring	826
23.4 The Gene Set Enrichment Analysis (GSEA)	826
23.5 Summary	827
24 Uses, misuses, and abuses in GO profiling	829
24.1 Introduction	830
24.2 "Known unknowns"	830
24.3 Which way is up?	831
24.4 Negative annotations	834
24.5 Common mistakes in functional profiling	834
24.5.1 Enrichment versus <i>p</i> -values	834
24.5.2 One-sided versus two-sided testing	835
24.5.3 Reference set	836
24.5.4 Correction for multiple comparisons	837
24.6 Using a custom level of abstraction through the GO hierarchy	837
24.7 Correlation between GO terms	840
24.8 GO slims and subsets	845
24.9 Summary	846
25 A comparison of several tools for ontological analysis	849
25.1 Introduction	849
25.2 Existing tools for ontological analysis	850
25.3 Comparison of existing functional profiling tools	863
25.3.1 The statistical model.	864
25.3.2 The set of reference genes	865
25.3.3 Correction for multiple experiments	865
25.3.4 The scope of the analysis	867
25.3.5 Performance issues	867
25.3.6 Visualization capabilities	867
25.3.7 Custom level of abstraction.	871
25.3.8 Prerequisites and installation issues	872
25.3.9 Data sources	874
25.3.10 Supported input IDs	874

25.4 Drawbacks and limitations of the current approach	876
25.5 Summary	878
26 Focused microarrays – comparison and selection	881
26.1 Introduction	881
26.2 Criteria for array selection	883
26.3 Onto-Compare	884
26.4 Some comparisons	885
26.5 Summary	891
27 ID Mapping issues	893
27.1 Introduction	893
27.2 Name space issues in annotation databases	894
27.3 A comparison of some ID mapping tools	898
27.4 Summary	901
28 Pathway analysis	903
28.1 Introduction	903
28.2 Terms and problem definition	904
28.3 Over-representation and functional class scoring approaches in pathway analysis	910
28.3.1 Limitations of the ORA and FCS approaches in pathway analysis	911
28.4 An approach for the analysis of metabolic pathways	914
28.5 An impact analysis of signaling pathways	914
28.5.1 Method description	914
28.5.2 An intuitive perspective on the impact analysis	915
28.5.3 A statistical perspective on the impact analysis	917
28.5.4 Calculating the gene perturbations	920
28.5.5 Impact analysis as a generalization of ORA and FCS	921
28.5.5.1 Gene perturbations for genes with no upstream activity	921
28.5.5.2 Pathway impact analysis when the expression changes are ignored	921
28.5.5.3 Impact analysis involving genes with no measured expression change	923
28.5.5.4 Impact analysis in the absence of perturbation propagation	924
28.5.5.5 Adding a new dimension to the classical approaches	924
28.5.6 Some results on real data sets	926
28.6 Variations on the impact analysis theme	935

28.6.1	Using the perturbation accumulation	935
28.6.2	Calculating the perturbation p -value using bootstrap- ping	935
28.6.3	Combining the two types of evidence with a normal in- version approach	936
28.6.4	Correcting for multiple comparisons	938
28.6.5	Other extensions	941
28.7	Pathway Guide	941
28.7.1	Data visualization capabilities	943
28.7.2	Portability	944
28.7.3	Export capabilities	944
28.7.4	Custom pathways support	944
28.7.5	Reliability benchmarks: speed, number of FPs, distribu- tion of p -values under the null distribution	946
28.8	Kinetic models versus impact analysis	946
28.9	Conclusions	949
28.10	Data sets and software availability	950
29	Machine learning techniques	951
29.1	Introduction	951
29.2	Main concepts and definitions	952
29.3	Supervised learning	955
29.3.1	General concepts	955
29.3.2	Error estimation and validation	956
29.3.3	Some types of classifiers	959
29.3.3.1	Quadratic and linear discriminants	959
29.3.3.2	k -Nearest neighbor classifier	960
29.3.3.3	Decision trees	961
29.3.3.4	Neural Networks	962
29.3.3.5	Support vector machines	964
29.3.4	Feature selection	968
29.4	Practicalities using R	970
29.4.1	A leukemia dataset	970
29.4.2	Supervised methods	971
29.4.3	Variable importance displays	973
29.4.4	Summary	973
30	The road ahead	977
30.1	What next?	977
	Bibliography	981
	Index	1027