Index

a

Affymetrix HG-U133A microarrays 67
age-standardized incidence, estimation 30
age-standardized rate 29
allelic heterogeneity 134
analysis, of experiments 8
– linear regression 8–11
– logistic regression (*Y* discrete) 11–13
– survival modeling 13–15
analysis of variance (ANOVA) 194, 195, 203, 207
– Fisher's linear discriminant analysis 194, 195
– *F* test 203
– OVRSVM framework 208
array comparative genomic hybridization (aCGH) 7, 245, 246, 263, 264

b

Bayesian gene selection approach 76
Bayesian model averaging (BMA) 43, 44, 46, 47, 52, 85
– computational assessment 48, 49
– iterative BMA algorithm (iBMA) 47–49
Bayesian SSVS approach 79
Bernoulli prior distribution 79
beta-uniform model (BUM) 19
bias 76, 80, 202, 203
– correction procedures 247
bioconductor 35

c

cancer
– data, applications 275, 276
–– potential 280–282
–– type I experiments 279, 280
–– type II experiments 276–278
– and general disease databases 30, 31, 33
– genes 145, 263
–– index data set 181, 187
–– Reactome FI/Fetch Index 188, 190
– measurement devices 263, 264
– subtypes, and therapies 112, 113
the Cancer Genome Atlas (TCGA)
– data about gene expression 32
– glioblastoma multiforme (GBM) 178
–– data set 178, 180
– mutation data file format (MAF) 182
– ovarian cancer (OV) mutation data set 174, 180, 182, 186
C/C++, programming language 35
chronic myeloid leukemia (CML) 44
– bone marrow/blood, cancer of 44
– progression gene expression data 49, 50
–– resources 54
cluster analysis 96, 193
colon cancer 110
– development 110, 111
– label-free proteomics 120, 121
–– exon-level significance 125, 126
–– genomic width of proteomic data 127
–– peptide-level significance 122–125
–– results 126, 127
–– whole protein-level significance 122
– molecular subsystems 113
–– interpretation 117–119
–– manifold 114–117
–– measurements 113, 114
–– validation 119, 120
– pathway paradigm 111, 112
colorectal cancer 109
– genetic background/environment contribution 110
complex disease gene network (CGN) 139
conditional distribution 81
copy number alteration (CNA) data analysis 178, 194, 196, 206, 241, 244–246, 249, 251, 258

Statistical Diagnostics for Cancer: Analyzing High-Dimensional Data, First Edition.
Edited by Frank Emmert-Streib and Matthias Dehmer.

– and cancer 244, 245
– limitations, need further investigations 259
– monoclonality, expriment results
–– significance 206, 207
–– testing existence 204–206
– for sporadic and recurrent mechanisms 245
– for testing existence of monoclonality 198–200
–– assessing statistical significance, of monoclonality 200, 201
–– preprocessing 200
–– visualization of monoclonality 201
– testing origin of ovarian cancer, expriment results
–– stage 1 207, 208
–– stage 2 208–211
– two-stage analytical method, for testing origin of cancer 201, 202
–– basic assumptions 202, 203
–– feature selection, and classification 203, 204
–– tissue heterogeneity correction 203
–– transcriptional network comparison 204
CRAN 35, 179
cumulative distribution function (CDF) 11

d

data applications
– analyzing DNA copy number data 249, 251
– for cancer and disorder 30
– CML progression data, case study 49, 50
– CNA data analysis for testing existence of monoclonality 198–200
– genomic data 128
– leukemia data 83
– lymphoma data 87
– mesothelioma cancer data set 102–104
–– results, from various procedures 104, 105
– proteomic data 119, 120
– TCGA OV Mutation Data Set 182–189
degree of freedom 12, 13, 65
deviance (DEV) for logistic regression model 12
differential dependency network 197, 198
– workflow 199
DiNAMIC approach, for analyzing DNA copy number data 241, 242, 251
– assessing statistical significance 252, 253
– Bootstrap test-based confidence intervals 257, 258
– confidence intervals for recurrent CNAs 256, 257
– cyclic shifts 251, 252
– limitations, need further investigations 259
– peeling 253–255
disease gene network (DGN) 137–139
disease gene prioritization 145
– diffusion-based methods 147
– disease-module-based methods 146
– linkage methods 145
disease genes, in protein interaction networks 134, 139–142
– complex disease gene networks 140
– intramodular hubs 142
– phenodiv genes 141, 142
– phenosim genes 141, 142
– phenotypically similar complex diseases 140
– topological position 142
disease genes, systems properties of 136
– disease-gene networks 137–139
– disease modules, identification of 143, 144
– network measures 136, 137
– protein interaction networks 139–142
disease modules, identification 143, 144
DNA copy number aberrations (CNAs) 241, 246
– data analysis 249
–– additional preprocessing and summary statistics 249, 250
–– analyzing with DiNAMIC (*See* DiNAMIC approach)
–– assessing statistical significance 250, 251
–– multiple sample methods to detect recurrent CNAs 249
–– multiple testing 250
– detecting recurrent CNAs, features 241
– "loss of heterozygosity" (LOH) approaches 242
– measurement of 245, 246
–– genoCNA and BACOM 246
–– mechanisms of 243, 244
– methods based on Hidden Markov models 248
– notation 247
– "null" distributions, and adjustments for multiple comparisons 242
– profiles of prostate cancer samples 265
– quality control and preprocessing 247
– segmentation algorithms 248
– single sample methods 246, 247
– thresholding 247, 248
– true copy number at each locus 242, 243
–– maternal and paternal alleles 243
DNA methylation arrays 5, 6
DNA microarray gene expression data 75, 95, 97

– characteristics 75
downloadable protein databases 34

e
empirical Bayes (EB)
– components 103
– estimation 59, 60
entropy 208
– conditional/unconditional 273
– data sets 271, 273
– estimation 262, 272
– genomic-transcriptomic, association with 272, 273, 277, 279, 280
–– statistical methodology to analyze data 268–275
– increase 261, 264, 266
– molecular, increases with 280
– mutual information and 272, 283
– normal and cancer transcriptome 269
– "normality-based" 270
– p^{53} pathway entropy increase 282
– simulation study 275
– statistical arguments 266–268
– surge 277
– variation 279
expectation maximization (EM) algorithm 60, 67, 72, 73, 101

f
false discovery rate (FDR) 4, 16, 58, 126, 185, 250
false positive probability (FPP) 66, 67
familial adenomatous polyposis (FAP) 109
Fisher's discriminant analysis 194, 195
Fortran, programming language 35
full unsupervised procedures 97
fully supervised procedures 96, 97

g
gene expression data 43–47, 50, 52, 53, 157, 162, 167, 176, 198, 203, 266, 282, 283
gene expression microarrays 3, 5, 6, 9, 263
Gene Expression Omnibus (GEO) database 30, 67
– available "RNA" and "genomic" samples 31
generalized g-prior (gg-prior) 76, 79, 89
gene regulatory networks, in human cancer 154, 155, 156, 215
– basic theory 154, 155
– Bayesian inference approach for analysis 164, 165
–– ERα transcriptional regulatory dynamics 165–167
– brute-force approach 218
– computational methodology 156
– contents 155
–– gene expression, and function 155
–– gene function, in systematic framework 156
–– structure properties/regulatory relationships, gene networks 155
–– transfer rules of genetic information, during gene expression 155, 156
– discrete genetic regulatory network model 217
– discrete-time, discrete-space Markov chain model 217
– dynamic Bayesian networks 218
– first-order Markov model, characterized by 218, 219
– genetic interventions, main approaches 219
–– heuristic control strategies 221, 222
–– optimal stochastic control 219–221
–– structural intervention strategies 222, 223
– human melanoma (*See* human melanoma gene regulatory network)
– optimal perturbation control 223–226
–– feasibility problem 226
–– minimal-energy perturbation control 226–228
–– robustness 231
–– trade-offs between 228–231
– probabilistic Boolean networks 218, 219
– *in silico* analytical approach 156–158
–– estrogen-dependent breast cancer cell line 158–160
–– genome-wide mapping of TGFβ/SMAD4 targets 160–163
– stochastic models of genetic interactions 216
–– dynamics 217
gene set analysis (GSA) algorithm 20
gene set enrichment analysis (GSEA) 20
genetic cancer network 28, 29
genome-wide association studies (GWAS) 135, 139, 145, 146
gg-SSVS approach 77, 82, 83, 86, 88, 89
Gibbs sampler 80, 81
g-prior distribution, of regression coefficients 76

h
hierarchical clustering 96, 195, 196
hierarchical mixture modeling 59, 60
high-throughput expression data 43
high-throughput platforms 1–21
human disease genes 138

– perspectives 148
human disease network (HDN) 27–29, 138
human diseases, genetic architecture 134–136
human melanoma gene regulatory network 231–235
– Frobenius norm increase with 235
– initial steady-state distribution 233
– melanoma gene regulatory network 232
– optimal perturbed matrix 233, 234
– probability transition matrix 233
– seven-gene probabilistic Boolean network model 233
– SLEM as decreasing function 234
human protein reference database (HPRD) 20, 116

i

interactome 139–147
iterative BMA algorithm (iBMA) 47–54
– perspective 53
– power 50, 51

k

Kaplan Meier survival estimation 94
K-means clustering 96
Kronecker product for covariance 78

l

label-free proteomics 120, 121
– exon-level significance 125, 126
– peptide-level significance 122–125
– results 126, 127
– whole protein-level significance 122
leave one out cross validation (LOOCV) 48, 76, 204. *See also* validation
– predictive probability 82
linear regression 8
– by examining deviation (DEV) 9
– logistic regression (*Y* discrete) 11–13
– multiple regression 11
– simple linear regression 9–11
"loss of heterozygosity" (LOH) approaches 242

m

MAQC Consortium 58
Markov chain Monte Carlo (MCMC) 76
mass spectrometry platforms 6, 7
mass-to-charge (m/z) ratio 6
matrix variate distribution 77
mean squared error (MSE) 10
measurement devices, for DNA copy numbers 263
microarray technology, perspective 72
model-based clustering methods 96
molecular subsystems 109, 113
– interpretation 117–119
– manifold 114–117
– measurements 113, 114
– validation 119, 120
Monte Carlo estimation 82
mortality (Mor) rates, estimation 30, 31
multinomial probit model 78
multiple testing type I errors 15, 16
– adjusted bonferroni method 17
– family wise error rate (FWER) method 17
– generalized Hochberg procedure 18
– generalized Šidàk procedure 18, 19
– Holm procedure 17, 18
– *k*-FWER method 17
– minP and maxT procedures 18, 19
multivariate Bayesian model, using *g*-prior 76
multivariate normal linear regression model 78

n

nearest-mean classifier 197
network-based approaches, network motif 173
network measures
– betweenness centrality 137
– closeness centrality 136
– clustering coefficient 137
– degree 136
network modules
– approaches in cancer studies, applications based on 179
–– cancer driver gene search based on network modules 179, 180
–– network modules and cancer prognostic signatures 179
–– using network patterns to identify cancer mechanisms 180
– containing functionally similar genes/proteins 174, 175
– searching methods 175
–– community search algorithms 177, 178
–– greedy search algorithms 175, 176
–– mutual exclusivity-based search algorithms 178
–– network clustering algorithms 176, 177
–– objective function guided search 176
–– weighted gene expression network 178, 179
next-generation sequencing (NGS) 27, 32, 37, 246
nondifferential genes 57, 58, 63, 66, 67

null hypothesis 10, 13, 15, 16, 18, 125, 206, 210, 250, 274, 276, 280

o

one-*versus*-all support vector machine 196, 197
one-*versus*-rest SVM (OVRSVM) classifier 196
Online Mendelian Inheritance in Man (OMIM) 28, 115, 134, 137, 141
optimal discovery procedure (ODP) 63, 64
ordinary differential equation (ODE) methods 153
OVR committee classifier solution 197

p

partial deviance 12
position weight matrices (PWMs) 156
posterior probability of differentially expressed (PPDE) 63, 65–71
– forest plots of θ_js forgenes based on 70
– selected top-ranked genes based on PPDE ranking 69
preprocessing HT platforms 7, 8
principal components analysis (PCA) 20
prior distribution 58, 59, 77, 79, 86, 89, 165
protein–protein interactions 32, 34, 119, 136, 143, 174, 175, 181
– and integration of gene in AML 143
proto-onco genes 263
P-values 4, 10–12, 18, 19, 68, 114, 120, 126, 176, 251, 274, 279, 282

r

ranking and selection methods 60
– based on effect sizes 60
–– posterior mean 61
–– rank posterior mean 61, 62
–– tail-area posterior probability 62, 63
– based on selection accuracy of differential genes 63
–– evaluating selection accuracy 64, 65
–– posterior probability of differentially expressed 63, 64
rank posterior mean (RPM) 61, 62, 65, 66, 68, 70, 71
– forest plots of θ_js for genes based on 70
– top-ranked genes based on RPM ranking 68
R-based visualization and analysis packages 34–36
R-centric implementation of data integration 36
Reactome FI Cytoscape plug-in 174, 180
– analyzing TCGA OV mutation data set 182
–– cancer gene index data overlay analysis 187–189
–– dialog to select file format and set parameters 182
–– FI subnetwork containing genes mutated from 183
–– Kaplan–Meier survival analysis for Module 6 189
–– loading the mutation file 182–184
–– module-based CoxPH survival analysis results 188
–– module-based survival analysis 186, 187
–– network clustering, and network module functional analysis 184–186
–– pathway annotations 185
–– properties of network modules 185
–– reactome pathway "Integrin cell surface interactions" 186
–– subnetwork, zoomed-in to show FI annotations 184
– cancer gene index data set 181
– construction of functional interaction network 181
– network clustering algorithm 181
real data analysis 83
– computational time 89
– leukemia data 83–87
– lymphoma data 87, 88
RNA-Seq platforms 3, 5, 6
R packages 35

s

semisupervised clustering (SS-Clust) algorithm 93, 102
semisupervised procedures 97, 98
– semisupervised clustering 99, 100
– semisupervised RPMM 100, 101
–– limitation 101
–– *vs. K*-means clustering 101
semisupervised recursively partitioned mixture models (SS-RPMM) 93, 102
– components 103
signal-to-noise ratio 45, 96
simulations 65–67
– results 66, 67
single-nucleotide polymorphism (SNP) arrays 245, 246
stochastic search variable selection (SSVS)
– method for gene selection 76, 89
support vector machine (SVM) 196
surface-enhanced laser desorption and ionization–time-of-flight (SELDI-TOF) 7
survival modeling 13
– Kaplan–Meier analysis 13–15

t
tail-area posterior probability (TPP) 62
t distribution 11, 81, 123, 124
transcription factors (TFs) 154
tumor-suppressor genes 263
type I error 4
– multiple testing 15, 16

u
unsupervised learning 96

v
validation 4, 21, 43, 48–51, 98, 113, 181, 204, 219
– laboratory 51, 52
– molecular subsystems 119, 120
variable selection, on gene expression data 44–46

w
web-based software 34
web interfaces 35
Wilcoxon test statistic 76