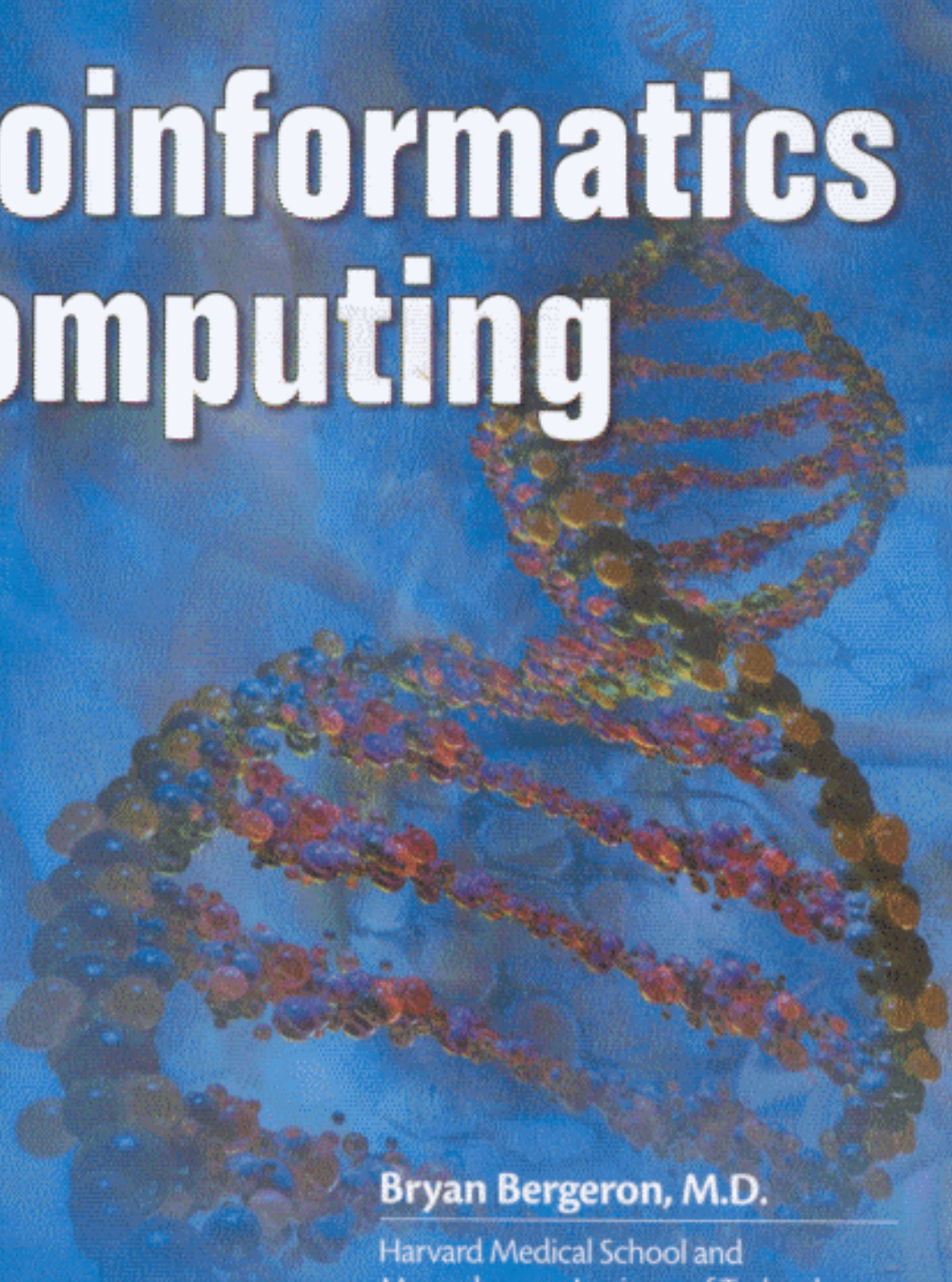


The complete, practical guide to bioinformatics for life scientists

Bioinformatics Computing



Bryan Bergeron, M.D.

Harvard Medical School and
Massachusetts Institute of Technology

CONTENTS

Preface	xiii
Acknowledgments	xxii

Chapter 1 • THE CENTRAL DOGMA 2

THE KILLER APPLICATION	4
PARALLEL UNIVERSES	8
WATSON'S DEFINITION	15
TOP-DOWN VERSUS BOTTOM-UP	19
INFORMATION FLOW	23
From Data to Knowledge	25
CONVERGENCE	28
Control	30
Archiving	34
Numerical Processing	35
Communications	39
ENDNOTE	40

Chapter 2 • DATABASES 42

DEFINITIONS	48
DATA MANAGEMENT	51
Pharmacogenomics and Aggression	54
Complexity	64
DATA LIFE CYCLE	66
Data Creation and Acquisition	66
Use	68
Data Modification	68
Archiving	70
Repurposing	74
Disposal	75
Managing the Life Cycle	75
DATABASE TECHNOLOGY	75
Database Architecture	80
Database Management Systems	84
INTERFACES	93
IMPLEMENTATION	96
Infrastructure	100
ENDNOTE	102

Chapter 3 • NETWORKS 104

GEOGRAPHICAL SCOPE	112
COMMUNICATIONS MODELS	113
TRANSMISSIONS TECHNOLOGY	114
PROTOCOLS	116
BANDWIDTH	119

TOPOLOGY	120
HARDWARE	123
Media	123
Network Electronics	126
CONTENTS	133
SECURITY	134
Antiviral Utilities	134
Authentication	135
Firewalls	136
Encryption	136
Process	138
OWNERSHIP	139
IMPLEMENTATION	139
MANAGEMENT	141
ON THE HORIZON	141
ENDNOTE	142

Chapter 4 • SEARCH ENGINES 144

THE SEARCH PROCESS	148
The Search	152
Results	155
SEARCH ENGINE TECHNOLOGY	155
Intelligent Agents	156
Portals	158
User Interface Tools	159
Utilities	162
SEARCHING AND INFORMATION THEORY	163
COMPUTATIONAL METHODS	166

Search Algorithms	166
Approximate Searches	168
SEARCH ENGINES AND KNOWLEDGE MANAGEMENT	171
ON THE HORIZON	174
ENDNOTE	175

Chapter 5 • DATA VISUALIZATION 176

SEQUENCE VISUALIZATION	182
Sequence Maps	182
STRUCTURE VISUALIZATION	186
Visualization Tools	187
Rendering Tools	187
USER INTERFACE	198
User Interface Components	199
Alternative Metaphors	203
Display Architecture	205
ANIMATION VERSUS SIMULATION	209
GENERAL-PURPOSE TECHNOLOGIES	209
ON THE HORIZON	212
ENDNOTE	214

Chapter 6 • STATISTICS 216

STATISTICAL CONCEPTS	218
Progress	220
MICROARRAYS	222
IMPERFECT DATA	231

BASICS	237
Randomness	237
Variability Is Cumulative	238
Approximation	238
Interface Noise	239
Assumptions	239
Sampling and Distributions	241
Hypothesis Testing	243
QUANTIFYING RANDOMNESS	243
DATA ANALYSIS	245
TOOL SELECTION	250
STATISTICS OF ALIGNMENT	253
CLUSTERING AND CLASSIFICATION	254
ON THE HORIZON	258
ENDNOTE	258

Chapter 7 • DATA MINING **260**

METHODS	264
Selection and Sampling	266
Preprocessing and Cleaning	266
Transformation and Reduction	270
Data-Mining Methods	270
Evaluation	272
Visualization	272
Designing New Queries	272
TECHNOLOGY OVERVIEW	272
INFRASTRUCTURE	273
PATTERN RECOGNITION AND DISCOVERY	274

MACHINE LEARNING	278
Inductive Logic Programming	280
Genetic Algorithms	280
Neural Networks	283
Statistical Methods	284
Decision Trees	287
Hidden Markov Models	289
TEXT MINING	290
Natural Language Processing	290
Text Summarization	295
TOOLS	296
ON THE HORIZON	298
ENDNOTE	300

Chapter 8 • PATTERN MATCHING **302**

FUNDAMENTALS	305
Pairwise Sequence Alignment	306
Local Versus Global Alignment	308
Multiple Sequence Alignment	308
Computational Methods	310
DOT MATRIX ANALYSIS	311
SUBSTITUTION MATRICES	313
DYNAMIC PROGRAMMING	317
WORD METHODS	324
BAYESIAN METHODS	326
MULTIPLE SEQUENCE ALIGNMENT	327
Dynamic Programming	327
Progressive Strategies	327

Iterative Strategies	328
Other Strategies	329
TOOLS	329
Nucleotide Pattern Matching	330
Polypeptide Pattern Matching	332
Utilities	338
Sequence Databases	338
ON THE HORIZON	339
ENDNOTE	339

Chapter 9 • MODELING AND SIMULATION **342**

DRUG DISCOVERY	348
FUNDAMENTALS	350
Components	350
Process	351
Perspectives	355
Numeric Considerations	358
Algorithms	361
Hardware	364
Issues	365
PROTEIN STRUCTURE	366
Ab Initio Methods	368
Heuristic Methods	377
SYSTEMS BIOLOGY	383
TOOLS	383
ON THE HORIZON	385
ENDNOTE	385

Chapter 10 • COLLABORATION 388

COLLABORATION AND COMMUNICATIONS 390**Asynchronous Communications 393****Synchronous Communications 393****Asynchronous Collaboration 394****Synchronous Collaboration 394****STANDARDS 399****Niche Solutions 400****Tool and Data Accessibility 400****Data Uniformity 402****Interoperability 403****Open Source 408****ISSUES 409****Platform Dependence 409****Security 410****Intellectual Property 410****Economics 411****ON THE HORIZON 413****ENDNOTE 414****BIBLIOGRAPHY 419****INDEX 428**