

Introductory Bioinformatics: Computer Skills for Bioinformatics

(A Theory and Practice)

Dr. Umesh Prasad Shrivastava, Ph.D.

*Associate Professor,
Tribhuvan University, TRM Campus, Birganj, Nepal*



JYOTI SHRIVASTAVA
Mithila Nagar, Birganj-16

Publisher

JYOTI SHRIVASTAVA
Mithila Nagar, Birganj-16

Author

Dr. Umesh Prasad Shrivastava, Ph.D.
Associate Professor,
Tribhuvan University,
TRM Campus, Birganj, Nepal
Email: upshrivastava@gmail.com
Mobile: +977-9845311392

First Edition: July 2018

Price: Rs. 350.00 only

Computer:

Mr. Vikash Thakur
Aadarsh Nagar, Birganj-13
Mobile: +977-9807263490

Printed at:

Photocopy world Service Center
Aadarsh Nagar, Birganj-13

ISBN: 978-9937-0-5054-8

Copyright © 2018 by Author

All right reserved. No part of this book may be reproduced, stored in a retrieval system or transmitted in any form or by any means: electronic, electrostatic, magnetic, tape, mechanical photocopying, recording or otherwise without the written permission of the author.

Contents

Preface

Chapter 1. Introduction to Bioinformatics **1-16**

- 1.1. Introduction--1
- 1.2. Definition of Bioinformatics--2
- 1.3. Role of Computer in Bioinformatics--4
- 1.4. Goals of Bioinformatic Analysis--5
- 1.5. Bioinformatics Technical Toolbox--6
- 1.6. A Chronological History of Bioinformatics--7
- 1.7. Key Areas of Bioinformatics--11
 - 1.7.1. Computational Biology--11
 - 1.7.2. Genomics--12
 - 1.7.3. Proteomics--13
 - 1.7.4. Pharmacogenomics--14
 - 1.7.5. Pharmacogenetics--14
 - 1.7.6. Cheminformatics--14
 - 1.7.7. Medical Informatics--15

Chapter 2. Bioinformatics Resources **17-36**

- 2.1. Introduction to GenBank --17
- 2.2. Access to GenBank --19
- 2.3. GenBank Data Usage--22
- 2.4. Depositing Data into the Public Databases--23
 - 2.4.1. GenBank Deposition--23
 - 2.4.2. Confidentiality--23
- 2.5. Genome databases--24
 - 2.5.1. NCBI Genome Resources--24
 - 2.5.2. TIGR Genome Resources--26
 - 2.5.3. Ensembl--27
 - 2.5.4. Other Sequencing Centers--28
 - 2.5.5. Organism-Specific Resources--28
- 2.6. Annotating and Analyzing Whole Genome Sequences--29
 - 2.6.1. Genome Annotation--29
 - 2.6.1.1. MAGPIE--30

- 2.6.1.2. Genome Comparison--30
 - 2.6.1.2.1. PipMaker--31
 - 2.6.1.2.2. MUMmer--31
- 2.7 Other major databases--32
 - 2.7.1. PDB--32
 - 2.7.2. Depositing Data into the PDB--35

Chapter 3. Scopes and Applications of Bioinformatics 37-46

- 3.1. Scopes of Bioinformatics--37
 - 3.1.1 What are the career options?--38
- 3.2 Applications of Bioinformatics--38
 - 3.2.1 Bioinformatics in medicine and medical science--38
 - 3.2.1.1 Antibiotic resistance--39
 - 3.2.1.2 Gene therapy--39
 - 3.2.1.3 Health and Drug discovery--39
 - 3.2.1.4 Molecular medicine--39
 - 3.2.1.5 Personalized medicine--40
 - 3.2.1.6 Preventative medicine--41
 - 3.2.2 Bioinformatics in Industries & Environment--41
 - 3.2.2.1 Alternative energy sources--41
 - 3.2.2.2 Climate change Studies--41
 - 3.2.2.3 Development of Drought resistance varieties--41
 - 3.2.2.4 Waste cleanup--41
 - 3.2.3 Bioinformatics in Agriculture & other Applied Fields--42
 - 3.2.3.1 Crop improvement--42
 - 3.2.3.2 Insect resistance--42
 - 3.2.3.3 Improve nutritional quality--42
 - 3.2.3.4 Veterinary Science--42
 - 3.2.3.5 Biotechnology--43
 - 3.2.3.6 Comparative Genomics--44
 - 3.2.3.7 Evolutionary studies--44
 - 3.2.3.8 Genome Annotation--44
 - 3.2.3.9 Microbial genome applications--44
 - 3.2.3.10 Prediction of Protein Structure--45
 - 3.2.3.11 Sequence Analysis--45
 - 3.2.4 Bioinformatics in Forensic Science--45

3.2.4.1 Forensic analysis of microbes--45

3.2.4.2 Bio-weapon creation--46

Chapter 4. Web Applications

47-64

4.1. Introduction to Internet--47

4.1.1 World Wide Web--49

4.1.2 File Fetching--53

4.1.3 Internet versus Intranet --54

4.2. Internet Resources of Biochemical/Biological Interest--54

4.3 Navigation on the World Wide Web--58

4.4 Browsers --59

4.4.1 Google Chrome--60

4.4.2 Mozilla Firefox--60

4.4.3 Microsoft Edge--61

4.4.4 Opera--61

4.4.5 Torch Browser--61

4.5 Finding Information on the World Wide Web--62

Chapter 5. Bioinformatics Tools

65-90

5.1 Sequence analysis tools and techniques--65

5.2 Gene finders and Feature Detection in DNA--66

5.3 Bioinformatics-Programs & Tools--68

5.4 Major categories of Bioinformatics Tools--69

5.4.1 Homology and Similarity Tools--69

5.4.1.1 Basic Local Alignment Search Tool (BLAST)--69

5.4.1.2 COBALT--72

5.4.1.3 Concise Microbial Protein BLAST--72

5.4.1.4 Conserved Domain Search Service (CD Search) -74

5.4.1.5 Genome BLAST --74

5.4.1.6 Genome ProtMap --75

5.4.2 Protein Function Analysis --76

5.4.3 Structural Analysis --77

5.4.4 Sequence Analysis --77

5.5 Multiple Sequence Alignment Tools --77

5.5.1 Clustalw --77

5.5.2 Clustal Omega --77

5.5.3 Kalign --77

5.5.4 MAFFT --78

- 5.5.5 MUSCLE --78
- 5.5.6 MView --78
- 5.5.7 T-Coffee --78
- 5.6. Some examples of Bioinformatics Tools--79
 - 5.6.1. EMBOSS --79
 - 5.6.2. FASTA --82
 - 5.6.3. RasMol --82
 - 5.6.4. PROSPECT --82
 - 5.6.5. PatternHunter --83
 - 5.6.6. COPIA--83
- 5.7. DNA Translation --83

Chapter 6. Biological Data and Information Retrieval from Database 91-128

- 6.1. Introduction to biological database systems using appropriate software --92
 - 6.1.1 Genomic Data --92
 - 6.1.2 Sequence Data Formats --93
 - 6.1.2.1 FASTA Format --93
 - 6.1.2.2 PHYLIP Format --94
- 6.2 Conversion of sequence formats using readseq --95
- 6.3 Primary Sequence Databases- GenBank, EMBL-Bank, and DDBJ - --96
 - 6.3.1 History-96
 - 6.3.2 Sequence Submission to the Databases --98
 - 6.3.2.1 Submission to NCBI/GenBank--98
 - 6.3.2.2 Submission to ENA/EMBL-Bank --99
 - 6.3.2.3 Submission to DDBJ --100
 - 6.3.3 Availability of the Submitted Sequence to the Public --101
 - 6.3.4 Sequence Flatfile Format --101
 - 6.3.4.1 GenBank Sequence Flatfile Format --103
 - 6.3.4.2 EMBL-Bank Sequence Flatfile Format --105
 - 6.3.5 Divisions of the NCBI Primary Sequence Database --108
- 6.4 Secondary and Specialized Databases Available to Public --109
 - 6.4.1 A Special Note on Various NCBI Databases --116
- 6.5 Data retrieval Tools --120
 - 6.5.1 Search and Retrieval using Entrez/GQuery --120
 - 6.5.2 Search and Retrieval using PubMed --122
 - 6.5.3 Search and Retrieval using OMIM --123

- 6.5.4 Search and Retrieval for Genomic DNA, mRNA & Amino Acids using Nucleotide Sequences Databases-124
- 6.5.5 Search and Retrieval Using Sequence Retrieval System (SRS)--124

Chapter 7. Computational Approach for Phylogenetic Analysis **129-154**

- 7.1 Phylogeny & Phylogenetic Analysis --129
 - 7.1.1 Key points for Phylogenetic trees --130
- 7.2 Fundamental Elements of Phylogenetic Models --132
- 7.3. Cladograms, Dendrograms, Phylograms, and Phylogenetic Trees --132
- 7.4. Types of Trees: Tree construction Methods --133
 - 7.4.1. Maximum Parsimony Method –138
 - 7.4.2. Distance Analysis method --140
 - 7.4.2.1. Unweighted Pair Group Method with Arithmetic Mean (UPGMA) --142
 - 7.4.2.2. Neighbor Joining (NJ) --143
 - 7.4.2.3. Fitch-Margoliash (FM) --143
 - 7.4.2.4. Minimum Evolution (ME) --145
 - 7.4.2.5. Which Distance-Based Tree-Building Procedure Is Best? --145
 - 7.4.3. Maximum likelihood Method --146
- 7.5. Some tips for reading phylogenetic trees --147
- 7.6. Phylogenetic Analysis Tools --148
- 7.7. Elements of phylogenetic models--150
- 7.8. Principles of Phylogenetic tree Construction --151
- 7.9. Application of Phylogenetic Analysis --151

Chapter 8. Genomics **155-166**

- 8.1 What is Genome? --155
- 8.2 Advances in Genomics --158
- 8.3 Structural Genomics --158
- 8.4 Functional genomics --160
- 8.5 Comparative genomics --161

- 8.5.1 Cereal comparative genomics --163
- 8.5.2 Comparative mapping of cereal genomes --163

Chapter 9. RNA Structure Analysis **167-174**

- 9.1. Some features of RNA --167
 - 9.1.1. Instability of mRNA --167
 - 9.1.2. 5'- and 3'-Untranslated Regions of mRNA --168
 - 9.1.3. Secondary Structures in RNA --169
- 9.2. Coding Versus Noncoding RNA --170
 - 9.2.1 Small Noncoding RNA, Long Noncoding RNA
Competing Endogenous RNA, and Circular RNA --171

Chapter 10. Microarray Bioinformatics **175-196**

- 10.1. Microarray --175
- 10.2. Microarray technologies --177
- 10.3 Principles of the application of microarray technology --178
- 10.4 The experimental process of microarray analysis --180
- 10.5 Secure storage of microarray data 181
- 10.6 Transforming microarray data --181
- 10.7 Annotation of microarray data --182
- 10.8 Analysis and visualization of microarray data --182
- 10.9 Sharing and publication of microarray data --183
- 10.10 Descriptions of freely available research database packages
 - 10.10.1 BASE --185
 - 10.10.2 Gecko --187
 - 10.10.3 RNA Abundance Database (RAD) --188
 - 10.10.4 Stanford Microarray Database (SMD) --190
 - 10.10.5 TM4 --192
- 10.11 Potential of Microarray analysis --194
- 10.12 Microarray Products --195
- 10.13 Microarray-Identifying interactions --195

Chapter 11. Programming Language (Perl CGI) 197-214

11.1 What is Perl like? --197

11.2 What is Genetic Algorithms? --199

11.3 History of Genetic Algorithms --199

11.4 Why Genetic Algorithms? --200

11.5 Genetic Algorithms Overview --200

11.6. Genetic Algorithms Terminology --201

11.6.1 Fitness Functions --201

11.6.2 Individuals --201

11.6.3 Populations and Generations --202

11.6.4 Diversity --202

11.6.5 Fitness Values and Best Fitness Values --203

11.6.6 Parents and Children --203

11.6 How the Genetic Algorithm Works --203

References

215-218

Syllabi of various courses

219-224

Index

225-228