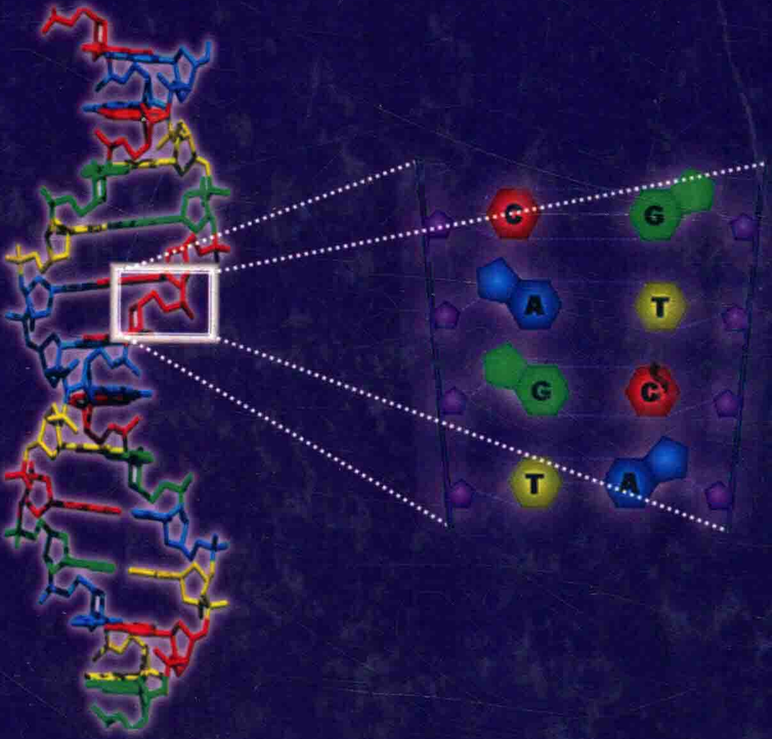
# Elementary Bioinformatics



Imtiyaz Alam Khan

PharmaMed
PRESS
An Imprint of
Pharma Book Syndicate

# Elementary Bioinformatics

**Imtiyaz Alam Khan**

*M.Sc., PGADB*

Bioinformatics tutor, Microbiologist.
Ex-Lecturer, Dept. of Biotechnology,
Rai University, Hyderabad.

**PharmaMed Press**
*An imprint of Pharma Book Syndicate*

**A unit of BSP Books Pvt. Ltd.**

4-4-316, Giriraj Lane,
Sultan Bazar, Hyderabad - 500 095.

***Published by :***

**PharmaMed Press**
*An imprint of Pharma Book Syndicate*
**A unit of BSP Books Pvt. Ltd.**
4-4-316, Giriraj Lane, Sultan Bazar, Hyderabad - 500 095.
Phone: 040-23445605, 23445688; Fax: 91+40-23445611
E-mail: info@pharmamedpress.com

**ISBN : 978-93-52300-09-9 (HB)**

# Elementary Bioinformatics

# Preface

This book presents good coverage of Bioinformatics concepts, concise coverage of multiple sequence alignment practicals, gene identification practicals, domain identification practicals. More illustrations have been included in most of the chapters to emphasize the importance of topics under study. This particular edition is prepared to facilitate students, academicians and research workers. Genetic Engineering which is a part of Bioinformatics, is also explained very clearly. This book is designed keeping in view of all Universitie's Syllabus; basically meant for JNTU B.Tech Students. Anybody interested in learning basic Bioinformatics can use it.

This book will serve well as a supplement to either a formal Bioinfomatics Course or an integrated curriculum. We hope that senior students, graduates, postgraduates, academicians and researchers will use this book as a reference for Bioinformatics information.

## TO THE READER

"Elementary Bioinformatics" is a very handy book, which is helpful in understanding Bioinformatics from its grass root levels. As we all know, bioinformatics is a highly emerging field. It is now in its primordial state and within few years it will reach climax. Understanding Bioinformatics at this level that too in a very conceptual, concise and vivid manner is what the requirement of present scenario. Every topic is explained very clearly, especially sequence alignment, multiple sequence alignments, phylogenetics predictions, gene predictions and so on.. After reading them the reader will get clear concept of those topics. Few practicals are also included in order to make them comfortable with online Bioinformatics tools.

"Elementary Bioinformatics" is designed to provide a logical framework for organizing, learning, reviving and applying the conceptual and factual information required.

# Acknowledgement

I am extremely thankful to Azher Khan, Asst. Prof., Dept of Computer Science DCET for his contributory work on Internet - Elementary Commands and Protocols in Chapter 1 of this book.

# Contents

1 Internet

1.1 Scope of Bioinformatics ........ 1
1.2 Introduction (Elementary Commands and Protocols) ........ 2
1.3 How Different Internet Services are Used ........ 3
1.4 Telnet ........ 3
1.5 FTP ........ 4
1.6 E-mail ........ 5
1.7 Using the Web ........ 9
1.8 Problems on the Web ........ 10
1.9 Some Practical Considerations for Using The Web ........ 11
1.10 Finding Web Resources ........ 13
1.11 Search Engines ........ 13

2 Sequence Analysis
(Introduction to Bioinformatics/Sequence Analysis)

2.1 Concept of DNA Protein Sequence Alignment and their Importance ........ 15
2.2 Strings ........ 15
2.3 Sequence Alignment and Alignment Programs ........ 17
2.4 Comparison of Two Sequences Schematic Comparison ........ 17

2.5 Dot –Matrix Analysis ........ 22
2.6 Scoring Schemes and Substitution Matrixes ........ 29
2.7 Dynamic Programming ........ 32
Needleman-Wunch and Smithwaterman Algorithm ........ 32
2.8 Heuristic Methods Blast and Fasta ........ 35
2.9 Searching Patterns ........ 41
2.10 The Statistics of Sequence Similarity Scores ........ 45

## 3 Calculation of Sequence Alignments

3.1 Principle of Multiple Sequence Alignment ........ 55
3.2 Global Alignments ........ 56
3.3 Local Alignments ........ 57
3.4 What is a Multiple Alignment? ........ 58
3.5 Pair Wise Sequence Alignment Verses Multiple Sequence Alignment ........ 64
3.6 Phylogenetic Trees ........ 68
3.7 Tools Used Such as Clustal X ........ 68
3.8 Multiple Sequence Alignments Using ClustalX (PRACTICAL) ........ 69
3.9 What is the Use of Sequence Patterns ........ 72
3.10 PSI –BLAST ........ 74
3.11 Structure Based Sequence Alignment-COMPARER ........ 74
3.12 Blast3 ........ 76
3.13 Multiple Sequence Alignments and Patterns ........ 76
3.14 Global Multiple Sequence Alignments ........ 76
3.15 Dynamic Programming Approach ........ 76
3.16 Progressive Pairwise Approach ........ 78
3.17 Modified Dynamic Programming Approach ........ 79
3.18 Patterns ........ 81
3.19 Patterns Derived from Unaligned Sequences ........ 82
3.20 Optimal and Sub Optimal Alignments ........ 82

## 4 Computational Biology (Biological Database Management System)

4.1 Introduction to Databases ..... 83

4.2 Introduction to Molecular Biology Databases ..... 87

4.3 Biological System Databases ..... 96

4.4 DNA Databanks ..... 98

4.5 The EMBL Nucleotide Sequence Database ..... 99

4.6 Data Bank of Japan ..... 99

4.7 SWISS – PROT ..... 100

4.8 Protien Information Resource ..... 101

4.9 PFAM ..... 102

4.10 INTERPRO ..... 103

4.11 PROSITE DATABASE : Database of Protein Families and Domains ..... 105

4.12 Protein Data Bank ..... 107

4.13 Structural SCOP : Classification of Proteins ..... 108

4.14 Class, Architecture, Topology And Homologous Superfamily ..... 110

## 5 Computational Biology : Information Retrieval From Databases

5.1 Contents of an Entry ..... 113

5.2 Organization Classification ..... 113

5.3 Search Concepts ..... 113

5.4 Database Searching Basic Considerations ..... 114

5.5 HOW to Perform Database Searching ..... 115

5.6 Tools for Searching ..... 116

5.7 Homology Searches ..... 119

5.8 Finding Domain And Functional Site Homologies With Database Search ..... 119

5.9 Programme & Website for Database Searches ..... 120

# 6 Phylogenetic Analysis

6.1 Introduction .......... 122
6.2 What Is Phylogenetic Systematics ? .......... 124
6.3 Understanding the Evolutionary Process .......... 124
6.4 What is Sampling .......... 127
6.5 Founder Effect .......... 128
6.6 Phylogenetic Trees : Presenting Evolutionary Relationships .......... 128
6.7 Methods of Phylogenetic Analysis .......... 130
6.8 The Origins of Molecular Phylogenetics .......... 131
6.9 Gene Trees Versus Species Trees – Why are They Different ? .... 135
6.10 Molecular Phylogenetics Terminology .......... 135
6.11 Molecular Clock Hypothesis .......... 136
6.12 The Importance of Molecular Phylogenetics .......... 136
6.13 A Brief Review of the Common Tree-building Methods Used in Phylogenetic Inference. .......... 137

# 7 Phylogenetic Prediction

7.1 Tree of Life .......... 143
7.2 Evolutionary Trees .......... 143
7.3 Rooted Trees .......... 144
7.4 Star Topology (Unrooted Trees) .......... 145
7.5 Methods for Determining Evolutionary Trees .......... 145
7.6 Neighbour-Joining Algorithm .......... 154
7.7 Unweighted Pair Group Method with Arithmetic Mean (UPGMA) . 155
7.8 Difficulties With Phylogenetic Analysis .......... 158
7.9 How to Make a Phylogenetic Tree .......... 160
7.10 Exercise: Sequence Alignments and Tree Building .......... 168

## 8 Genetic Engineering

8.1 Genetic Engineering .......... 178
8.2 Techniques of Genetic Engineering .......... 178
8.3 Restriction Enzymes .......... 179
8.4 Vectors .......... 180
8.5 Plasmids .......... 181
8.6 Gene Transfer .......... 182
8.7 Genetic Markers .......... 184
8.8 Polymerase Chain Reaction .......... 186
8.9 DNA Probes .......... 189
8.10 Electrophorises .......... 190
8.11 DNA Sequencing .......... 192
8.12 DNA Sequencing Problem .......... 194
8.13 Similarity of the Human Genome .......... 195
8.14 Tools of DNA Sequencing .......... 196
8.15 Gel Electrophoresis .......... 197
8.16 Shotgun Sequencing .......... 199
8.17 Sequencing Strategies .......... 201
8.18 Applications of Genetic Engineering .......... 204
8.19 Gene Products .......... 205
8.20 New Phenotypes .......... 210
8.21 The Future of Gene Therapy .......... 211

## 9 DNA Sequencing

9.1 Fragment and Map Assembly and Combinational Approaches to Sequencing .......... 213
9.2 cDNA sequencing .......... 215
9.3 Assembly .......... 216
9.4 Genome Sequence Assembly Using Trace Signals and Additional Sequence Information .......... 218
9.5 Genome Sequencing .......... 228

## 10 Genome Mapping

10.1 Cognate Modification Enzymes ............ 236
10.2 Gel Electrophoresis of DNA Fragments ............ 237
10.3 Physical Map of Genome ............ 238
10.4 Physical Mapping using Hybridization ............ 238
10.5 DNA Hybridization - Melting Curves ............ 239
10.6 Variants of Southern Gels ............ 243
10.7 Genome Fragment Identification : DNA Microarrays ............ 243
10.8 Genome Fragment Isolation : Cloning and High Resolution Physical Mapping of Genomes ............ 246
10.9 Genome Mapping : Genetic -> Physical ............ 251

## 11 Gene Identification

11.1 Sequence Feature Extraction / Annotation ............ 257
11.2 Gene Annotation ............ 258
11.3 Analysis of Gene Functions and Metabolic Pathways From Databases ............ 264
11.4 Computer Assisted Analysis of Transcription Control Regions ............ 265
11.5 Translational Signals ............ 265
11.6 Splicing Signals ............ 266
11.7 Comparative Sequence Analysis ............ 267
11.8 Eukaryotic Gene Prediction ............ 268

**Appendix A**
List of Bioinformatics Course Web Pages ............ 277

**Appendix B**
Downloadable Resources to Power Up Your Web Browser ............ 285

**Appendix C**
Background/Introductory material on Molecuar Biololgy ............ 287

# 1
# Internet

## 1.1 Scope of Bioinformatics

Bioinformatics has emerged as a distinct discipline that straddles the interface between the traditional biological sciences and the computer sciences and advance computational methodologies. It is rapidly becoming a powerful new approach to understanding life, and it may well reverse the reductionist paradigm that has held sway in molecular biology ever since Erwin Schrodinger turned on a generation of physicists to biology with the publication of "What is life ? more than 50 years ago.

The new discipline of bioinformatics promises to provide the tools needed to attack the complexity of conducting holistic biological research, "because of this complexities, biology will eventually become the most computational science, surpassing physics," said Delici, who predicts that within the next 10 to 15 years bioinformatics will become the integral part of biology.

Modern day biotech and drug discovery industry has witnessed the development of high through put automated equipment, which enables amassing of data faster than it can be analyzed and the utilized. Pharmaceutical research will clearly be the one major benefactor of development in bioinformatics. 100 of new drug targets have been identified by using computational techniques, which involves searching for genes similar to known protein encoding genes. In the future, virtual toxicology screening may be the first setup in predicting the effects of new chemical or complex metabolic pathways. In addition, bioinformatics will likely provide the methodology, to make highly accurate predictions about protein tertiary structure based on amino acid sequences and a viable means to design drugs based on computer simulation of the docking of small molecules to the predicted protein architecture.

According to figures from Framingham mass based market analysis firm IDC Bio-IT market will reach $38 billion in 2006. The market includes pharmaceutical, healthcare research and biotechnology companies, as well as Govt. linked institutions. A number of recent work force studies have shown that there is a high current and unmet demand for people trained to various levels of expertise in bioinformatics, serve the upcoming biotech and biopharmaceutical industry which has, observed significant growth in genomic era.

## 1.2 Introduction (Elementary Commands and Protocols)

The Internet has become an important tool for biological and biomedical research scientists. Using the Internet, it is possible to perform a number of kinds of analysis on research data and to search for and obtain information. Over the last several years, the number of tools and the amount of information relevant to biologists available on the Internet has grown and the ease of use of these tools has grown as well. As a result of both of these trends, the value of Internet resources for biologists now significantly outweighs the costs in time and money of using it. The overall goal of this chapter is to help biologists use the Internet effectively and to illustrate to computer scientists, how biologists are currently using the Internet.

***This chapter has three specific goals :***

1. To provide background information which will help demystify computer network usage.
2. To provide an introduction to the resources available to biologists over the Internet in sufficient detail to allow the students in this course to explore and learn how to use these resources on their own.
3. To provide practical instruction to these students on using the specific network resources needed during the remainder of the course.

It is assumed that the students can use a Web Browser (e.g. Internet Explorer or Netscape Navigator) to access the contents of the course.

## 1.3 How Different Internet Services are Used

As noted above, different Internet services are characterized by client software, server software, a set of capabilities they agree upon (e.g. text, pictures in the GIF format, etc.), a protocol by which they communicate (i.e. how the data is encoded in a stream of bytes), and a port on which the client contacts the server to begin communication. The distinction between different services can be blurred because different services can perform similar functions, because different services can share capabilities, and because of the existence of multifunction clients (and servers). Specifically, many modern Web clients are highly multifunctional, being gopher, ftp, e-mail and net news clients in addition to being Web clients. Finally, note that although this chapter deals with services delivered via the Internet, some of these same services can be delivered via other, very different kinds of networks, Internets, or dedicated connections.

What is presented here is a very superficial overview of a few of the available Internet services.

## 1.4 Telnet

Telnet is one of the oldest of the network services and perhaps the easiest to understand. Telnet allows one computer to "log on" to another computer as if it were a terminal. Once logged on, you frequently will have all the privileges of a local user; you can run programs, create and delete files. This is probably the most common way that users with accounts will use a computer.

Although "full service logins" as is described above are perhaps the most common use of the telnet protocol, in fact as much control as the host's system administrator desires may be imposed on a telnet connection. Thus, a telnet service may be advertised with a public login name and password. Login with this name, however, is likely to be restricted to a limited number of commands. The National Institutes of Health in the United States used, at one point, such a telnet login to disseminate information as to the membership of study sections. Such specialized telnet services have become much less common since the rise in popularity of the Web.

A telnet session can negotiate a range of different protocols, but this almost always includes ASCII text. Because many protocols for other services (e.g. SMTP, HTTP) are encoded as ASCII text, a telnet client can sometimes be used to connect to a server for these other protocols. Most people will use a telnet client the first time connecting to a MOO, and some people will continue to use telnet as their client, although most of us find dedicated clients to be significantly more convenient. Similarly, it is possible to connect to a Web server with a telnet client if you understand the syntax of HTTP. This is almost never done to *use* a Web server, but is occasionally done when debugging.

From a practical point of view, every telnet host will be different, and thus you will need to learn about each one as you have occasion to use it.

## 1.5 Ftp

Telnet is useful for interactive computer access, but is much less useful for transferring files. Ftp is an older service designed specifically for file transfer. Originally like telnet, it was intended for account owners. However, as it became apparent that it was useful to make files available to the world at large without giving all those wanting the files an account, the variant of "anonymous ftp" developed. In this variant, logging in with a "magic" user name (most commonly "anonymous" or "ftp") eliminates the requirement for a password.

In 1996, "To a large extent, use of the World Wide Web has rendered (direct) ftp access obsolete." Although there was and is some truth to that statement (especially given that files on an ftp server can be retrieved by a Web client), the need for ftp clients has not vanished. Some users will choose to avoid them, preferring the simplicity of dealing with a single piece of software, but within their domain ftp clients are more versatile than Web browsers, in some cases one has more control with an Ftp client, and for simple file transfer they are quicker and more convenient.