精要速览系列

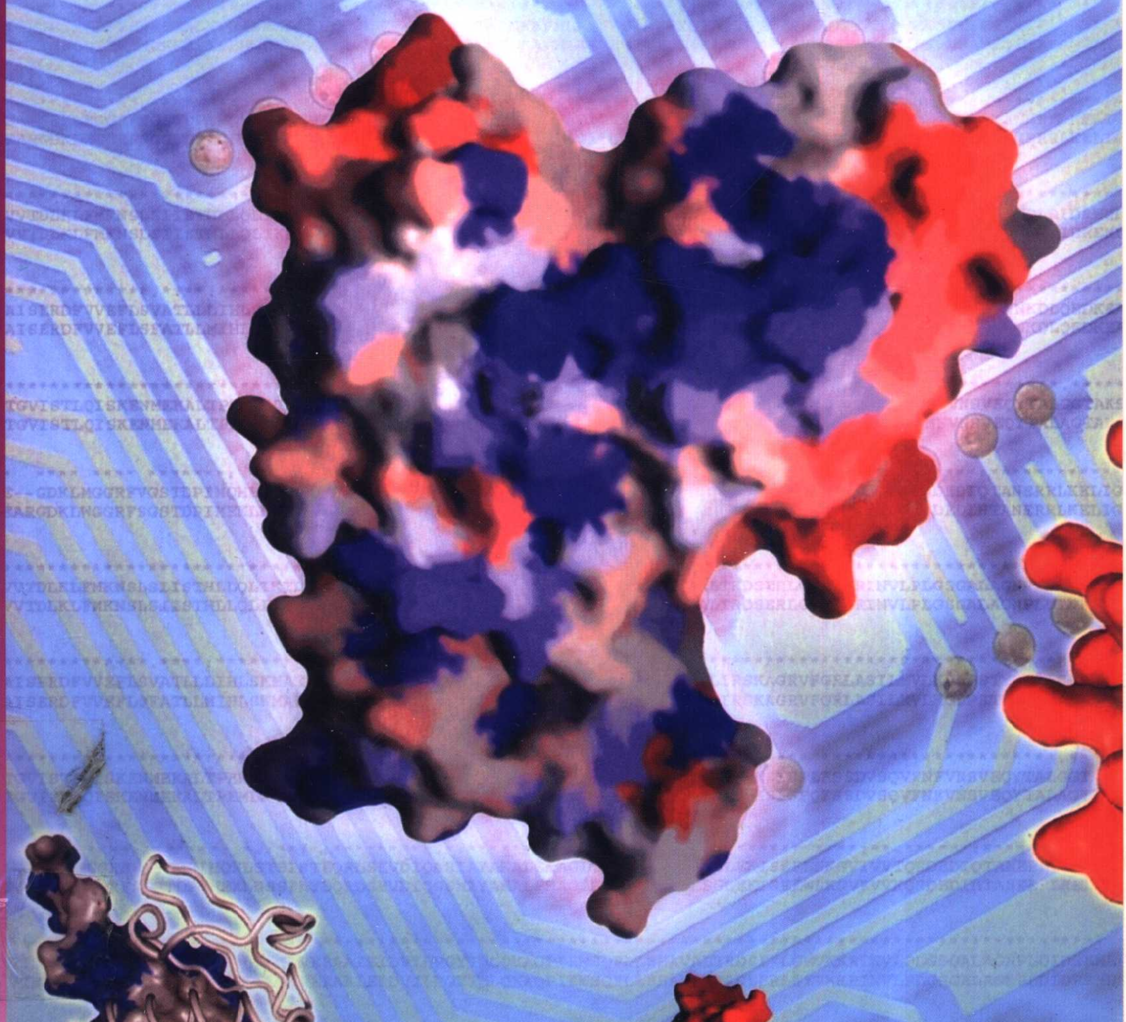
Instant Notes 先·锋·版

D.R. Westhead, J.H. Parish & R.M. Twyman

Bioinformatics

生物信息学

影印本



科学出版社
www.sciencep.com

精要速览系列——先锋版

Instant Notes in

Bioinformatics

生物信息学

（影印版）

D. R. Westhead, J. H. Parish & R. M. Twyman

科 学 出 版 社
北 京

图字:01-2002-6154

内 容 简 介

“精要速览系列(*Instant Notes Series*)”是国外教材“Best Seller”榜的上榜教材。该系列结构新颖,视角独特;重点明确,脉络分明;图表简明清晰;英文自然易懂,被国内多所重点院校选用作为双语教材。先锋版是继“现代生物学精要速览”之后推出的跨学科的升级版本。

本书是该系列中的《生物信息学》分册,全书共15章,全面系统地概括了该学科的核心内容和前沿动态,涉及生物学数据的获得与处理、系统发生学、芯片数据分析、蛋白质组数据分析、生物信息学的应用等主要内容。

本书是指导大学生快速掌握生物信息学基础知识的优秀教材,也是辅助教师授课的极佳教学参考书,同时可供生命科学相关专业的研究生参考。

图书在版编目(CIP)数据

生物信息学/(英)韦斯特海德(Westhead,D.R.),帕里什(Parish,J.H.),特怀曼(Twyman,R.M.)著.—影印本.—北京:科学出版社,2003.4
(精要速览系列——先锋版)

ISBN 7-03-011025-0

Ⅰ.生… Ⅱ.①韦… ②帕… ③特… Ⅲ.生物信息论-英文
Ⅳ.Q811.4

中国版本图书馆CIP数据核字(2002)第097604号

责任编辑:周 辉 黄 方
责任印制:刘士平/封面设计:王 浩

科学出版社 出版
北京东黄城根北街16号
邮政编码:100717
http://www.sciencep.com
双青印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2003年4月第 一 版 开本:B5(720×1000)
2003年4月第一次印刷 印张:17
印数:1—4 000 字数:330 120

定价:32.00元

(如有印装质量问题,我社负责调换〈环伟〉)

ABBREVIATIONS

2D-PAGE	two-dimensional polyacrylamide gel electrophoresis
AC	approximate correlation
ACeDB	A *C. elegans* DataBase
ADIT	AutoDep Input Tool
AE	annotated exon
AN	actual negative
AP	actual positive
AQL	ACeDB query language
BDGP	Berkeley *Drosophila* Genome Project
BIND	Biomolecular Interaction Network Database
BIOS	Basic Input–Output System
BRET	bioluminescent resonance energy transfer
CASP	Critical Assessment of Structure Prediction
CD	circular dichroism
CD	candidate drug
CDE	common desktop environment
cDNA	copy DNA
CDS	coding sequence
CGI	common gateway interface
CIP	Cahn–Inglod–Prelog
CORBA	Common Object Request Brokering Architecture
DBMS	database management system
DDBJ	DNA Databank of Japan
DEC	Digital Equipment Corporation
DIP	Database of Interacting Proteins
DNS	Domain Name Server
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
ENU	ethylnitrosourea
EP	Expression Profiler
ES cells	Embryonic stem cells
ESI	electrospray ionization
EST	expressed sequence tag
ExPASy	Export Protein Analysis System (Switzerland)
FE	false exon
FN	false negative
FP	false positive
FRET	fluorescent resonance energy transfer
FTP	file transfer protocol
GASP	Gene Annotation aSsessment Project
GEO	Gene Expression Omnibus
GFP	green fluorescent protein
GGTC	German Gene Trap Consortium
GNOME	GNU network object model environment
GOLD	Genomes Online Database
GOR	Garnier–Osguthorpe–Robson
GRAIL	Gene Recognition and Assembly Internet Link
GSS	genome survey sequence
GST	glutathione S-transferase
GUI	graphical user interface
HIV	human immunodeficiency virus
HMM	hidden Markov model
HSP	high-scoring segment pair
HTG	high-throughput genomic sequence
HTML	hypertext markup language
HTS	high-throughput screening
http	hypertext transfer protocol
IP	Internet Protocol
ISP	Internet Service Provider
KDE	K desktop environment
KEGG	Kyoto Encyclopedia of Genes and Genomes
LCA	last common ancestor
LOG	Laplacian of Gaussian
m/e or m/z	mass/charge ratio
MAD	multiwavelength anomalous diffraction
MAGE	microarray and gene expression
MAGE-ML	microarray gene expression markup language
MAGE-OM	microarray gene expression object model
MALDI	matrix-assisted laser desorption/ionization
ME	missing exon
MGED	Microarray Gene Expression Database
MIAME	minimum information about a microarray experiment

MIME	Multiple Internet Mail Extensions
MIR	multiple isomorphous replacement
MMDB	Molecular Modeling Database
mRNA	messenger RNA
MS	mass spectrometry
MSD	Macromolecular Structure Database
MS-DOS	Microsoft Disk Operating System
MSF	multiple sequence format
NBRF	National Biomedical Research Foundation
NCBI	National Center for Biotechnology Information
NDB	Nucleic Acid Data Bank
NJ	neighbor joining
NMR	nuclear magnetic resonance
NNSSP	Nearest Neighbour Secondary Structure Prediction
NOE	nuclear Overhauser effect
OMIM	OnLine Mendelian Inheritance in Man
ORF	open reading frame
PAGE	polyacrylamide gel electrophoresis
PAM	accepted point mutations
PAUP	phylogenetic analysis using parsimony
PCNA	proliferating cell nuclear antigen
PCR	polymerase chain reaction
PDB	Protein Data Bank
PE	predicted exon
PERL	Practical Extraction and Reporting Language
PH	pleckstrin homology
PHYLIP	phylogenetic inference package
pI	isoelectric point
PIR	Protein Information Resource
PN	predicted negative
PP	predicted positive
RCSB	Research Collaboratory for Structural Bioinformatics
RMSD	root mean square deviation
rRNA	ribosomal RNA
RT	reverse transcription
SAGE	serial analysis of gene expression
SDS	sodium dodecyl sulfate
SELDI	surface-enhance laser desorption/ionization
SH2, SH3	Src – homology domain
SMART	Simple Modular Architecture Research Tool
SMILES	Simplified Molecular Input Line Entry Specification
SNP	single nucleotide polymorphism
SOM	self-organizing map
SPR	surface plasmon resonance
SQL	symbolic query language
SRS	sequence retrieval system
SSE	secondary structure element
STS	sequence tagged site
T_C	Tanimoto coefficient
TCP	Transmission Control Protocol
TE	true exon
TN	true negative
TP	true positive
TrEMBL	Translated EMBL
tRNA	transfer RNA
UML	Unified Modeling Language
UPGMA	unweighted pair group method using arithmetic mean
UPGMC	unweighted pair group method using centroid value
URL	uniform resource locator
WE	wrong exon
WPGMA	weighted pair group method using arithmetic mean
WPGMC	weighted pair group method using centroid value
WST	watershed transformation
WWW	World Wide Web
XML	eXtensible Markup Language
Y2H	yeast two-hybrid

PREFACE

Computational analysis of biological sequences has been practised for a long time, but its importance to the practising biochemist or molecular biologist has only emerged in the last ten years or so as high-throughput sequencing techniques have brought a flood of valuable sequence data of relevance to many research projects. Thus the discipline of bioinformatics has emerged and grown enormously in importance. Center stage in these developments has been the sequencing of whole genomes. The human genome was finished more than a year ago and convenient access to the annotated data for the non-bioinformatician is just becoming available.

Following DNA sequencing, other experimental techniques have been developed to the point where they can be described as high-throughput. Gene- and protein-expression patterns can be studied with microarrays on the scale of whole genomes, protein interactions can be studied on the same scale with the yeast two hybrid system, and there is even talk of high-throughput structure determination. Where there is large-scale data generation there is need for computational data handling and analysis, and so the subject of bioinformatics grows.

It is now crucial that the practising experimental biologist has a knowledge of bioinformatics, and so the subject is beginning to appear and grow in undergraduate curricula. This book is essentially aimed at the experimental biologist, perhaps as an undergraduate or maybe further along a career path, who needs a working knowledge of the subject. It would also be usful to computer scientists and others looking to move into a biological domain. We begin by describing data generation and databases, and then move to the 'classical' bioinformatics question, "What can I do with my sequence?". Following this we look at the newer bioinformatics problems associated with structure, expression, proteomics, interactions and pathways.

Throughout the book we aim to describe what is possible, and the strengths, limitations and potential pitfalls of methods and analyses. We will tell you how to do things, but this is not a software manual for commonly used packages. They have their own manuals that are (mostly) much better than anything we could provide. Many of the methods we describe rely on quite complex mathematical, statistical or computational techniques. Often we choose not to describe these at all, but where we do we have aimed for a simple conceptual understanding. It is often said that it is not necessary to understand in detail how the internal combustion engine works in order to drive a car. The same applies to understanding of underlying methods in bioinformatics, but where a little understanding is possible and helpful, we have tried to provide it. We hope that it is useful.

Acknowledgments

The authors would like to thank Jonathan Ray, Sarah Carlson and David Hames for their help, understanding and support during the preparation of the book, and Chris Hodgson for advice and discussion on Section O.

David Westhead dedicates this book to his parents Robert and Mavis,
his wife Andrea, and children Elizabeth and Francis.

Howard Parish dedicates this book to his grand-daughter Scarlett.

Richard Twyman dedicates this book to his parents, Peter and Irene,
his children, Emily and Lucy, and to Hannah, Joshua and Dylan.

Contents

Abbreviations vii
Preface ix

Section A – Overview of bioinformatics **1**
- A1 The scope of bioinformatics 1
- A2 Bioinformatics and the Internet 4
- A3 Useful bioinformatics sites on the WWW 7

Section B – Data acquisition **9**
- B1 Sequencing DNA, RNA and proteins 9
- B2 Determination of protein structure 15
- B3 Gene and protein expression data 17
- B4 Protein interaction data 22

Section C – Databases – contents, structure and annotation **29**
- C1 File formats 29
- C2 Annotated sequence databases 35
- C3 Genome and organism-specific databases 41
- C4 Miscellaneous databases 45

Section D – Retrieval of biological data **51**
- D1 Data retrieval with Entrez and DBGET/LinkDB 51
- D2 Data retrieval with SRS (sequence retrieval system) 55

Section E – Searching sequence databases by sequence similarity criteria **59**
- E1 Sequence similarity searches 59
- E2 Amino acid substitution matrices 64
- E3 Database searches: FASTA and BLAST 69
- E4 Sequence filters 78
- E5 Iterative database searches and PSI-BLAST 80

Section F – Multiple sequence alignment: gene and protein families **83**
- F1 Multiple sequence alignment and family relationships 83
- F2 Protein families and pattern databases 86
- F3 Protein domain families 91

Section G – Phylogenetics **97**
- G1 Phylogenetics, cladistics and ontology 97
- G2 Building phylogenetic trees 103
- G3 Evolution of macromolecular sequences 109

Section H – Sequence annotation **113**
- H1 Principles of genome annotation 113
- H2 Annotation tools and resources 118

Section I – Structural bioinformatics **123**
I1 Conceptual models of protein structure 123
I2 The relationship of protein three-dimensional structure to protein function 129
I3 The evolution of protein structure and function 131
I4 Obtaining, viewing and analyzing structural data 136
I5 Structural alignment 143
I6 Classification of proteins of known three-dimensional structure: CATH and SCOP 146
I7 Introduction to protein structure prediction 149
I8 Structure prediction by comparative modeling 151
I9 Secondary structure prediction 156
I10 Advanced protein structure prediction and prediction strategies 161

Section J – Microarray data analysis **165**
J1 Microarray data: analysis methods 165
J2 Microarray data: tools and resources 170
J3 Sequence sampling and SAGE 175

Section K – Proteomic data analysis **179**
K1 Analyzing data from 2D-PAGE gels 179
K2 Analyzing protein mass spectrometry data 185

Section L – Higher-order systems **189**
L1 Modeling and reconstructing molecular pathways 189
L2 Protein interaction informatics 194
L3 Higher-order models 199

Section M – Cheminformatics in biology **203**
M1 Conventions for representing molecules 203
M2 Cheminformatics resources 209

Section N – Bioinformatics in the pharmaceutical industry **215**
N1 Bioinformatics and drug discovery 215
N2 Pharmainformatics resources 219

Section O – Basic principles of computing in bioinformatics **227**
O1 Running computer software 227
O2 Computer operating systems 230
O3 Software downloading and installation 232
O4 Database management 236

Further reading 239
Glossary 243
Index 255

A1 THE SCOPE OF BIOINFORMATICS

Key Notes

What is bioinformatics?

Bioinformatics is the combination of biology and information technology. It is the branch of science that deals with the computer-based analysis of large biological data sets. Bioinformatics incorporates the development of databases to store and search data, and of statistical tools and algorithms to analyze and determine relationships between biological data sets, such as macromolecular sequences, structures, expression profiles and biochemical pathways.

The role of computers in bioinformatics

Computers are required in bioinformatics for their processing speed (allowing repetitive tasks to be carried out quickly and systematically) and for their problem-solving power. However, many problems addressed by bioinformatics still require expert human input, and the integrity and quality of the source data are also critical.

Scope of this book

This book is designed to provide the newcomer to bioinformatics with enough information to understand the principles of bioinformatic applications and to gain some practice in their use. The text covers basic introductory subjects such as the role of the Internet as well as the key areas of bioinformatics: the use of databases, sequence and structural analysis tools, and tools for annotation, expression analysis and the analysis of biochemical and molecular pathways. Section O is an appendix of peripheral information on computer operating systems and software.

Instant Notes Bioinformatics WWW site

Throughout this book there are references to various WWW sites that house information resources, databases and bioinformatic tools. Since the WWW is constantly evolving, the addresses of these sites are likely to change. For convenience, links to these sites are listed on an accompanying WWW site (http://www.bios.co.uk/inbioinformatics), which will be regularly updated by authors. If any of the URLs in this book do not work, the accompanying WWW site is probably the best way to reach the required site.

Related topics

Bioinformatics and the Internet (A2)

Useful bioinformatics sites on the WWW (A3)

What is bioinformatics?

Bioinformatics is the marriage of biology and information technology. The discipline encompasses any computational tools and methods used to manage, analyze and manipulate large sets of biological data. Essentially, bioinformatics has three components:

- The creation of **databases** allowing the storage and management of large biological data sets.
- The development of **algorithms** and **statistics** to determine relationships among members of large data sets.
- The use of these tools for the analysis and interpretation of various types of biological data, including DNA, RNA and protein sequences, protein structures, gene expression profiles and biochemical pathways.

The term bioinformatics first came into use in the 1990s and was originally synonymous with the management and analysis of DNA, RNA and protein sequence data. Computational tools for sequence analysis had been available since the 1960s but this was a minority interest until advances in sequencing technology (Topic B1) led to a rapid expansion in the number of stored sequences in databases such as GenBank (Topic C2). Now the term has expanded to incorporate many other types of biological data, for example protein structures, gene expression profiles and protein interactions. Each of these areas requires its own set of databases, algorithms and statistical methods, some of which are discussed in this book.

The role of computers in bioinformatics

Bioinformatics is largely although not exclusively a computer-based discipline. Computers are important in bioinformatics for two reasons. First, many bioinformatics problems require the same task to be repeated millions of times. For example, comparing a new sequence to every other sequence stored in a database (Topic E3) or comparing a group of sequences systematically to determine evolutionary relationships (Topic G2). In such cases, the ability of computers to process information and test alternative solutions rapidly is indispensable. Second, computers are required for their problem-solving power. Typical problems that might be addressed using bioinformatics could include solving the folding pathway of a protein given its amino acid sequence, or deducing a biochemical pathway given a collection of RNA expression profiles. Computers can help with such problems, but it is important to note that expert input and robust original data are also required.

Scope of this book

This book is based on the authors' experience in teaching bioinformatics at undergraduate and postgraduate level. A common starting point for those new to bioinformatics is 'What can I do with this sequence?' This book is designed to give the reader an informed background to understanding methods used in bioinformatics and sufficient examples and technical details to enable him or her to answer real problems. We describe the role of the Internet in bioinformatics (Section A), how data used in bioinformatics is generated (Section B), the importance of databases (Section C) and how these are accessed and searched (Section D). We discuss sequence analysis (Sections E, F and G), sequence annotation (Section H), structural analysis and prediction (Section I), gene and protein expression analysis (Sections J and K) the bioinformatics of protein interactions (Sections L and M) and some applications of bioinformatics in the pharmaceutical industry (Section N). Section O comprises a series of appendices providing background information on file formats, computer operating systems and software.

There are topics in computational biology that we have intentionally omitted. Software designed specifically for structure refinement, automated instrumentation (including robotics) and other types of data collection are omitted. We include methods for molecular graphics but, otherwise, we omit graphical and other aids to document presentation.

Instant Notes Bioinformatics WWW site

This book makes reference to many databases and computer software tools that are available on the World Wide Web as well as various informative web sites. Although the addresses for many of these resources are listed in the book, the Internet is constantly evolving and such addresses are subject to change on a regular basis. For convenience, links to all the sites discussed in the text can be found on an accompanying WWW site (http://www.bios.co.uk/inbioinformatics). The WWW site also contains further information, updates and links that are not found in this book.

A2 BIOINFORMATICS AND THE INTERNET

Key Notes

The Internet

The Internet is an international computer network that uses a particular protocol, known as TCP/IP, to package and route data. Most academic, government and commercial institutions have access to the Internet, and individuals may also gain access by subscribing to an ISP. The Internet provides a versatile system for exchanging biological data.

The World Wide Web (WWW)

The WWW is an Internet-based system for information exchange using a protocol called http. Programs called browsers can access hypertext documents on the Internet by searching for the relevant address, known as a URL. Hypertext documents contain text and other multimedia objects (images, audio files, etc.) but their most important property is the presence of hyperlinks that allow the browser direct access to other hypertext documents. These may be hosted by any computer on the Internet, and in this way, the user can rapidly jump from computer to computer around the world viewing and downloading information.

Browsing, working and downloading

The WWW is a rich source of biological data and bioinformatics resources. Many bioinformatic tools and databases can be accessed and used over the Internet. However, due to constraints in the speed of data transfer or access to the Internet, it may also be appropriate to download databases and bioinformatic software and use them on a local computer.

Related topics

Useful bioinformatics sites on the WWW (A3)

Installing bioinformatic software locally (O4)

The Internet

Biological information is stored on many different computers around the world. The easiest way to access this information is for the computers to be joined together in a network. A **computer network** is a group of computers that can communicate, for example over a telephone system, therefore allowing data to be exchanged between remote users. A typical computer network is shown in *Fig. 1.* For transfer, data are first broken into small **packets** (units of information), which are sent independently and reassembled when they arrive at their destination. If information is sent from computer A to computer C, it can travel

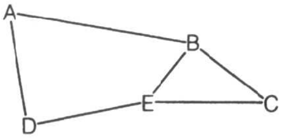


Fig. 1. A simple computer network.

via two different routes. In one case computer B acts as a **router**, and in the other case computers D and E both act as routers. The availability of different routes through the network means that communications can be maintained between computers A and C even if part of the network is unavailable, for example if computer B ceases to function.

The **Internet** is an international network of computers derived from an earlier system, **ARPAnet**, developed by the US military. The Internet as we know it began in 1969, when four American universities were connected together for the first time allowing the rapid exchange of scientific data. The number of computers linked to the Internet has grown exponentially over the last 30 years and it is now estimated that over 20 million computers have access, many of them personal computers in people's homes. Information transfer over the Internet is governed by a set of **protocols** (procedures for handling data packages) called **TCP/IP**. TCP is the **Transmission Control Protocol**, which determines how data is broken into packages and reassembled. IP is the **Internet Protocol**, which determines how the packets of information are addressed and routed over the network. To access the Internet, a computer must have the correct hardware (generally a network card and/or a modem), the appropriate software and permission for network access. Many institutions have automatic access to the Internet, but private users must subscribe to an **Internet Service Provider (ISP)**.

The World Wide Web (WWW)

The **World Wide Web (WWW)** is a way of exchanging information over the Internet using a program called a **browser**. A number of browsers are available for working on the WWW, the most widely used of which are **Internet Explorer** and **Netscape Navigator**. Most computers are sold nowadays 'Internet ready' with the appropriate hardware and one or both of these browser programs installed as standard. The WWW was developed in 1992 and allows the display of information pages containing **multimedia objects** (e.g. text, images, audio and video) in a special format called **hypertext**. In a hypertext document, text is displayed normally and can be read and manipulated like any other text document, but some words and objects are highlighted in a different color and these are known as **hypertext links** (or simply **hyperlinks**). Clicking on a hyperlink directs the browser to access another hypertext document, which might be on the same computer or might be on any other computer linked to the Internet. The new document may have its own hyperlinks and thus the process can be repeated allowing the user to move rapidly from computer to computer around the world downloading information as he or she goes (this is commonly known as **surfing the web** or **surfing the net**).

The WWW works on the basis that each hypertext document has a unique address known as a **uniform resource locator (URL)**. URLs take the format http://restofaddress, where 'http://' identifies the protocol for communication over the WWW (**hypertext transfer protocol**) and 'restofaddress' provides a location for the hypertext document on the Internet. Every computer on the Internet has an **IP address**, which is in the form of four integers conventionally separated by dots. Associated with this is a text version of the address, for example http://www.bios.co.uk, which is the publisher's address. The equivalent IP address for the publisher is 195.172.6.15. If a local user tries to contact http://www.bios.co.uk, how does the browser find the correct site? The local computer first contacts Internet computers called **Domain Name Servers (DNSs)** that try to understand parts of the address starting with the most significant (right hand) part. For example, most text addresses have a country abbreviation, in this case 'uk' for United Kingdom, but American addresses do not since the Internet was an American invention. If the computer one is trying to access is providing a service on the WWW, it is known as a **web server**. This means there are numerous files available for browsing, and each can be identified by a unique URL. Such files are specified by extra characters separated from the main Internet address by a solidus (/). For example, the URL http://www.bios.co.uk/bioinformatics refers to a subdirectory on the publisher's web server that corresponds to the web site accompanying this book. Once the DNS has found the Internet name for the server, it is for the server itself to work out what do about any extensions to the URL such as '/bioinformatics'.

Browsing, working and downloading

Browsing the Internet is simply a case of clicking on the desired hyperlinks and allowing the associated pages to download. Some pages download faster than others, which may be due to content (pages with many images and other large files will take longer to download than pages that contain text alone) or due to the speed of connection (there are bottlenecks in many parts of the Internet, and it is advisable to find a local web server to minimize the number of routers the information has to pass through). It is also notable that the Internet will be busier at certain times of the day, and during the weekends when recreational use increases. Many bioinformatics sites are hosted by several web servers in different locations around the world to reduce such bottlenecks. Different web servers providing the same service are called **mirrors**.

To access a particular web site, it may first be necessary to type in the URL in the **address bar** of the browser. Once a page has been accessed, however, it should not be necessary to type in the URL again. Browser programs maintain a list of URLs that have been visited (the **History file**) and any URL can be added to a list of **Favorites** (in Internet Explorer) or **Bookmarks** (in Netscape Navigator) to allow easy access in the future. Where does one start on the Internet? A number of public **search engines** are available allowing the user to search for sites of interest using particular keywords, but it may be better to start with some dedicated bioinformatics sites. For the absolute beginner, a selection of useful bioinformatics web sites is listed in Topic A3.

Having got the feel of bioinformatics on the WWW, what are the merits and demerits of installing software locally (Topic O4), rather than using a WWW site? Although locally installed software will usually run faster than the same application used over the Internet, some software is difficult to install and might need expert help. There are advantages in having local copies of simple sequence alignment and other software if you are working 'at home', that is, limited by rates of data transmission on telephone lines. However, the use of locally installed databases can be disadvantageous because updates will be published less frequently than the WWW-based versions. Many academic institutions have an **Intranet**, that is, a local network that can be accessed only from computers within the institution. Such local networks may provide a number of bioinformatics tools and applications, which will usually run just as fast as locally installed software.

A3 USEFUL BIOINFORMATICS SITES ON THE WWW

Key Notes

Useful bioinformatics sites

For the absolute beginner, the Internet can be daunting and intimidating. Nine good starting points (gateways) for bioinformatics on the Internet are listed. Each of the sites is well maintained, simple to use and provides a wealth of resources such as links, databases and bioinformatics software.

Searching the Internet

Once the user has gained some experience using the Internet, information can be found relatively easily. If the bioinformatic gateways do not provide the information required, general-purpose search engines can be used to locate pages containing specific key words or phrases. Otherwise, the home pages of academic institutions and biotechnology companies can be useful starting points.

Pitfalls and hints

When using a search engine it is important to refine the search and to avoid ambiguous words and phrases, as these can pull out numerous irrelevant pages. Literature databases can help by providing useful keywords and phrases to use as search terms.

Related topics

Bioinformatics and the Internet (A2)

Annotated sequence databases (C2)

Genome and organism-specific (C3)

Miscellaneous databases (C4)

Useful bioinformatics sites

For absolute beginners, we have listed nine good starting points for bioinformatics on the WWW (*Table 1*). Each of these **gateway sites** is comprehensive, has many useful links and is well maintained and stable. Time spent browsing and using these sites will provide an accurate feeling for the bioinformatics resources available on the Internet.

Searching the Internet

Although the nine web sites listed in *Table 1* provide some of the best starting points for bioinformatics on the WWW, there is a great deal of specialist biological data that cannot be accessed directly from these sites. Finding relevant data on the Internet is made simpler by the availability of general-purpose **search engines,** such as Google, Yahoo, Lycos, AltaVista and Hotbot. These tools search the entire Internet for pages that contain particular keywords or phrases, and they can also be used to search for files of a particular type, such as image files or video files. For example, one might search the Internet using the phrase 'alcohol dehydrogenase' to find pages containing information about that enzyme. Alternatively, one might look for image files of a particular insect or flower, or video files of frog development. Relevant sites are displayed as a list

Table 1. Nine good starting points for bioinformatics on the WWW

URL	Note
General bioinformatics 'gateways'	
http://www.ncbi.nlm.nih.gov/	National Center for Biotechnology Information homepage. A resource for public databases, bioinformatics tools and applications. Links to many useful sites and resources for bioinformatics software.
http://www.ebi.ac.uk/	The EMBL European Bioinformatics Institute outstation. A resource for biological databases and software, much of which has excellent tutorial support.
http://www.expasy.ch/	The ExPASy (Expert Protein Analysis System) Molecular Biology Server. Maintained by the Swiss Institute of Bioinformatics (SIB). Provides links, databases and software resources for the analysis of protein sequences, structures and expression.
http://www.embl-heidelberg.de/	European Molecular Biology Laboratory homepage
http://www.gmd.de/Welcome.en.html	German National Centre for Information Technology homepage
http://links.bmn.com/	The BioMedNet gateway to thousands of biological websites, includes a search facility and provides descriptions of each of the web sites listed.
Genome projects	
http://wit.integratedgenomics.com/GOLD/	Genomes On Line Database, with links to genomic databases and progress reports on genome projects.
http://www.genome.ad.jp/kegg/	Kyoto Encyclopaedia of Genes and Genomes. A very comprehensive Japanese site including metabolic maps.
Computing notes	
http://foldoc.doc.ic.ac.uk/foldoc/index.html	FOLD (Free Online Dictionary of Computing). A good place to look up meanings of computer jargon.

of hits, with hyperlinks allowing direct access to the page of interest. The problem with general-purpose search engines is that they have not been developed specifically with molecular biology in mind, and the information they provide can be irrelevant or misleading, especially if the search term used has other connotations.

As an alternative to search engines, the home pages of academic institutions or biotechnology companies can also be a good place to start. Many universities, for example, maintain comprehensive web sites with pages for staff to describe research projects and display data, and such sites often contain hyperlinks to sites of related interest.

Pitfalls and hints

On a general-purpose search engine it is probably better to start with a set of key words that is very restrictive and then remove some of the words if no hits are generated. If a search term is too broad (e.g. 'biochemistry') it will produce a ridiculous number of hits and it will be impossible to check all the listed pages. Search terms with known alternative uses are also best avoided. For example, searching the Internet with the word 'steroid' will likely hit more pages on body-building than molecular biology! A positive suggestion is to use a literature database on the WWW such as PubMed (Topic C4) to look for useful and appropriate keywords and phrases to use as search terms.