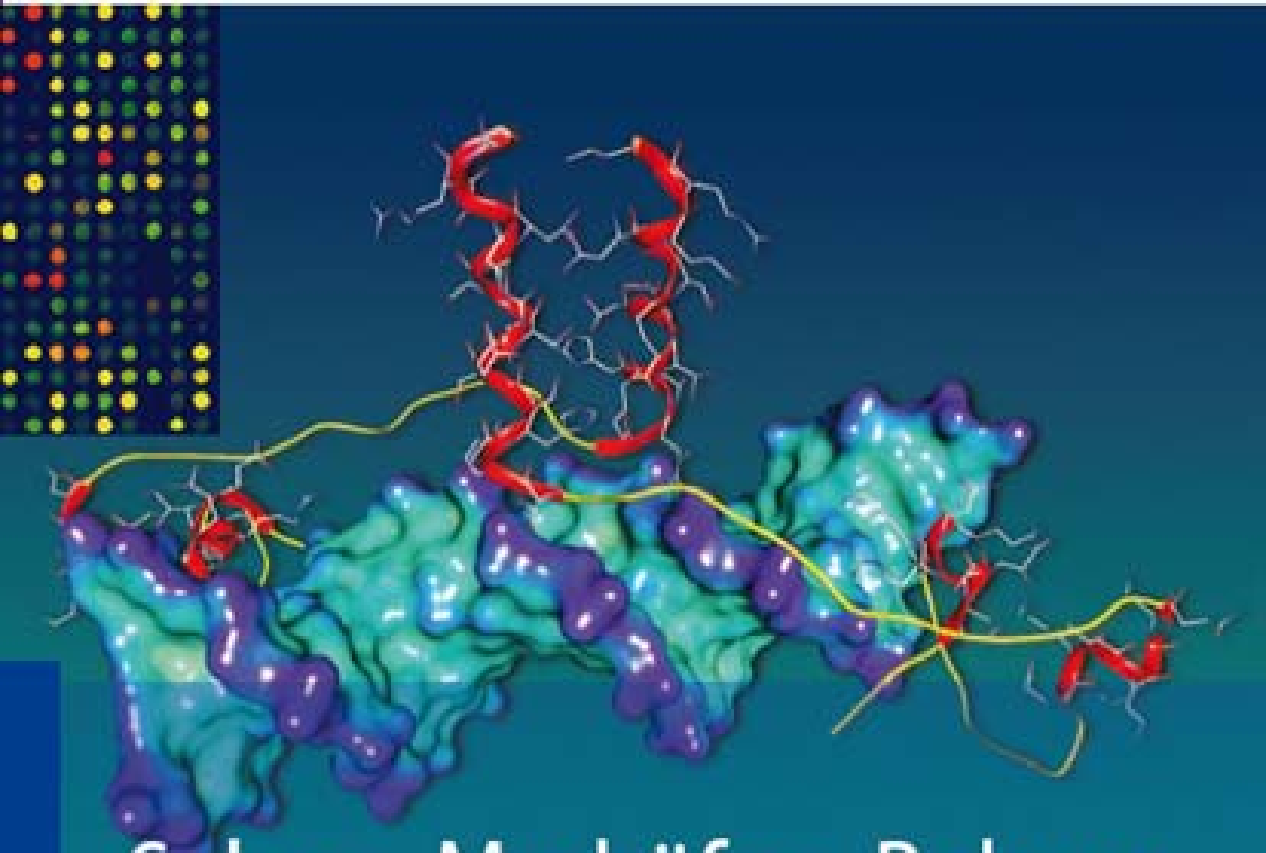


Selzer · Marhöfer · Rohwer

Applied Bioinformatics

An Introduction

Including Exercises and Solutions

Springer

Applied Bioinformatics

P.M. Selzer • R.J. Marhöfer • A. Rohwer

Applied Bioinformatics

An Introduction

With 69 Figures, 61 in color and 6 Tables

Springer

Prof. Paul M. Selzer, Ph.D.
Intervet Innovation GmbH
BioChemInformatics
Zur Propstei
55270 Schwabenheim
Germany
paul.selzer@intervet.com

Andreas Rohwer, Ph.D.
Intervet Innovation GmbH
BioChemInformatics
Zur Propstei
55270 Schwabenheim
Germany
andreas.rohwer@intervet.com

Richard J. Marhöfer, Ph.D.
Intervet Innovation GmbH
BioChemInformatics
Zur Propstei
55270 Schwabenheim
Germany
richard.marhoefer@intervet.com

Translators:

PD Frank Seeber, Ph.D.
Dept. Biology/Parasitology
Philipps University Marburg
Karl-von-Frisch-Str. 8
35042 Marburg
Germany
seeber@staff.uni-marburg.de

Conor Caffrey, Ph.D.
Director of Biochemistry and Molecular Parasitology
Sandler Center for Basic Research in Parasitic Diseases
Byers Hall N508
University of California, San Francisco
1700 4th Street
San Francisco, CA 94158-2330
USA
caffrey@cgl.ucsf.edu

ISBN: 978-3-540-72799-6 e-ISBN: 978-3-540-72800-9

Library of Congress Control Number: 2007937094

First published in German under the title: Angewandte Bioinformatik 2004

Cover design: WMXDesign GmbH, Heidelberg, Germany

Printed on acid-free paper 5 4 3 2 1 0

springer.com

Preface

Though a relatively young discipline, bioinformatics is finding increasing importance in many life science disciplines, including biology, biochemistry, medicine, and chemistry. Since its beginnings in the late 1980s, the success of bioinformatics has been associated with rapid developments in computer science, not least in the relevant hardware and software. In addition, biotechnological advances such as those witnessed in the fields of genome sequencing, microarrays and proteomics have contributed enormously to the bioinformatics boom. Finally, the simultaneous breakthrough and success of the World-Wide Web has facilitated the worldwide distribution of and easy access to bioinformatics tools.

Today, bioinformatics techniques such as the BLAST algorithm, pairwise and multiple sequence comparisons, queries of biological databases, and phylogenetic analyses have become familiar tools to the natural scientist. Many of the software products that were initially unintuitive and cryptic have matured into relatively simple and user-friendly products being easily accessible over the Internet. One no longer needs to be a computer scientist to proficiently operate bioinformatics tools with respect to complex scientific questions. Nevertheless, what remains important is an understanding of fundamental biological principles together with a knowledge of the appropriate bioinformatics tools available and how to access them. Also, and not least, is the confidence to apply these tools correctly to generate meaningful results.

The present book is based on a lecture series of Paul M. Selzer, professor of biochemistry at the Interfaculty Institute for Biochemistry, Eberhard-Karls-University, Tübingen, Germany, and was first published in German in 2004. The book is unique in that it includes both exercises and their solutions, thereby making it suitable for the classroom. To access a wider international audience, the authors have prepared a revised second edition in English. Each of the three authors has many years of accumulated expertise in research and development within the pharmaceutical industry, specifically in the area of bioinformatics and cheminformatics at Intervet Innovation GmbH, a worldwide leading animal health company. The aim of this book is both to introduce the daily application of a variety of bioinformatics tools and to provide an overview of a complex field. However, the intent is neither to describe nor even derive formulae or algorithms, but rather to facilitate rapid and structured access to applied bioinformatics by interested students and scientists. Therefore, even though detailed knowledge in computer programming is not required to understand or apply this book, proficiency in operating a PC, its standard software, and the Internet are prerequisites.

Each of the eight chapters describes important fields in applied bioinformatics and provides both references and WWW links. Detailed exercises and solutions are meant to encourage the reader to practice and learn the topic, and become proficient in the relevant software. If possible, the exercises are chosen in such a way as that examples, such as protein or nucleotide sequences, are interchangeable. This allows the reader to choose examples that are closer to his/her scientific interest based on a sound understanding of the underlying principles. Direct input required by the user either through text or by pressing buttons is indicated in Courier font. Finally, the book concludes with a detailed glossary of common definitions and terminologies used in applied bioinformatics.

Schwabenheim, June 2007

We thank Intervet Innovation GmbH, Schwabenheim, Germany for its continuous support in preparing this textbook.

The Circulation of Genetic Information

Genetic information is encoded by a four-letter alphabet, which in turn is translated into proteins using a 20-letter alphabet. Proteins fold into three-dimensional structures that perform essential functions in single-celled or multicellular organisms. These organisms are under constant selection pressure, which in turn leads to changes in their genetic information.

Cover picture

The upper left side of the figure shows part of a DNA microarray in which the gene expression patterns of two bacterial strains have been overlaid. Compared to the reference strain, red spots indicate overexpression, green spots - lower expression, yellow spots - similar expression and black spots - a lack of complementary cDNA.

On the right, the three-dimensional molecular structure of a protein-DNA complex is depicted. The transcription activator Gal4 from *Saccharomyces cerevisiae* is shown bound to a DNA oligomer (PDB-ID: 1D66). Gal4 is represented by a ribbon model in which α-helices and loops are drawn in red and yellow, respectively. The side chains of the amino acids in the loops are not shown. For the DNA oligomer, local bending of the molecular surface is color-coded where darker colors represent increased bending [Brickmann J, Exner TE, Keil M, Marhöfer RJ (2000) Molecular graphics - trends and perspectives. J Mol Mod 6:328-340]. The structure was produced on a Silicon Graphics Octane 2 workstation using the software package MOLCAD/Sybyl (Tripos Inc.) [Brickmann J, Goetze T, Heiden W, Moeckel G, Reiling S, Vollhardt H, Zachmann CD (1995) Interactive visualization of molecular scenarios with MOLCAD/Sybyl. In: Bowie JE (Hrsg) Data visualization in molecular science - tools for insight and innovation. Addison-Wesley Publishing Company Inc, Reading, Massachusetts, USA, S 83-97].

Contents

1 Computers, Operating Systems and the Internet

1.1 Computers and Operating Systems

The computer is the most important tool for a scientist in bioinformatics. Until a few years ago large mainframe computers were necessary for the calculations and database searches required in bioinformatics. Because of the ever increasing efficiency of modern personal computers (PC) this is no longer the case and most tasks can be accomplished with today's PC. Large computer systems are only required when very large sets of calculations have to be carried out or where large amounts of data are handled. In these cases, rather than large mainframe or special-purpose computers, clusters of networked computers are sufficient and are based on the same technology as a normal PC (Fig. 1.1). While in the office and home environment the two operating systems Windows and MacOS are mainly used, UNIX is generally required for bioinformatics. However, Microsoft's operating system Windows gains increased importance in this area, e.g. through Microsoft's digital pharma initiative [microsoft lifesciences].

From the emergence of the first PCs in the 1970s, different operating systems have been developed. The CP/M operating system of Digital Research, Inc., introduced in 1974, was one of the first systems operational on many microprocessors. At the beginning of the 1980s, IBM equipped their newly introduced IBM PC with the Microsoft operating system MS DOS, leading to this system's ascendancy over the competition. The operation of MS DOS was exclusively by command lines and thus

Fig. 1.1. High performance cluster on the basis of Intel x86 CPUs. (Printed with permission of Silicon Graphics GmbH)

anything but user-friendly or even intuitive. In parallel, the idea of implementing a graphical user interface (GUI) sprouted. In 1983, the Apple Lisa was the first PC marketed whose operating system had a graphical user interface for interaction with the user. The GUI laid the foundation for the success of the Apple Macintosh computer.

Microsoft had also come to realize the advantages of a GUI and started the development of a corresponding operating system. The operating system Windows from Microsoft hit the market in the middle of the eighties. However, it only became widely accepted upon the publication of Windows version 3.0 in the early nineties. Windows 3.0 was still a pure single user operating system, but Windows 3.11 introduced network functionality. Starting with Windows 95, Microsoft has begun to develop a completely

new Windows operating system, which today contains concepts of UNIX operating systems in the current versions of Windows XP, Windows Server 2003 and Windows Vista.

Apple has also advanced its MacOS operating system and the current version MacOS X offers a UNIX-based operating system with a perfected GUI. However, because Apple had never achieved a world market position like Microsoft, a comparatively small number of bioinformatic applications, especially in the freeware and shareware sector, are available for the MacOS system.

UNIX was developed at Bell Laboratories in 1969 as a multi-user system. A multi-user system simply means that several users can use a computer at the same time. Initially, UNIX was written in the programming language *Assembler*, which depends on the processor type used. A complete new programming of UNIX was carried out in 1973 in the newly developed programming language C, to make porting to different processor types easier. The development of UNIX was later licensed so that different developers endeavored to modify and sell an extended UNIX version of their own. Several UNIX derivatives have arisen from this competition, including, among others, BSD-UNIX of the University of California, Berkeley, as well as System V. Over the years several UNIX derivatives were integrated into one another, resulting in increasingly more efficient operating systems. The current UNIX derivatives are based on the portable system operating interface (POSIX) standard, an international standard which integrates the two main systems, BSD-UNIX and System V. Systems that are POSIX-compliant are System V release 4, Solaris 10 and Linux. Linux has emerged as one of the most modern UNIX systems since the version 0.01 developed in 1991 by Linus Torvalds, a student of the University of Helsinki. One reason for Linux's success is the free distribution of Linux under the *GNU general public license* (GNU GPL). This has allowed hundreds of programmers worldwide to collaborate on the improvement of Linux, resulting in an extremely successful operating system. Because of its efficiency, flexibility and stability, Linux is gaining more and more acceptance in all areas at present, including bioinformatics.

UNIX operating systems are generally module-based, i.e., different functionalities are combined into single modules. For example, if it is not inevitably necessary that a Linux computer access the hard disk of another Linux computer via a network, installation of this functionality (network file system, NFS), which is integrated in a certain network module, can be renounced. Because of this modularity, the installation of the very same UNIX type can look very different on two different computers. The basic organization and instructions are, however, the same.

Almost all UNIX variants meanwhile offer a GUI that permits operations via a mouse. These interfaces are based on the X Window system of the X.Org Foundation. Besides operation via the generally intuitive GUI, the GUI also has an internal shell that can be opened for operation via command lines (Fig. 1.2). The shell represents a window for the input and output of text and corresponds to the DOS window or the command prompt on a Windows PC. Frequently the shell is a more powerful and considerably faster way of operation.

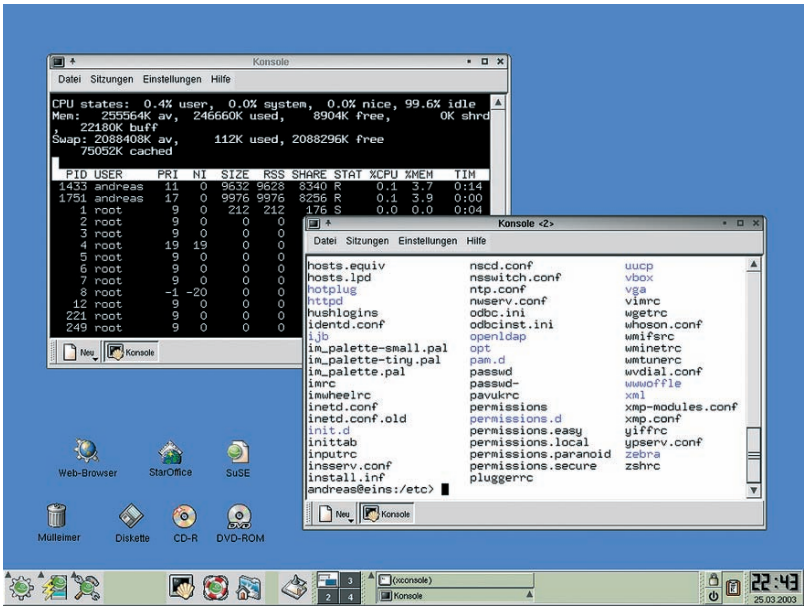


Fig. 1.2. Linux desktop interface with two shell windows

1.2 Internet and WWW

The following sections give a short outline of the development of the Internet and explain some services that play a role on the Internet. For reasons of completeness and to obtain an overview, this section may also be worthwhile for experienced Internet users and WWW surfers.

The Internet as we know it today was originally designed by the US military. Within the US Ministry of Defense, the Advanced Research Projects Agency (ARPA) was founded which, among other operations, was also engaged in the development of a computer communication network. The project was designed so that failure in one part of the network would not lead to a complete shutdown. The result was ARPANET and this can be regarded as the basis of today's Internet. With the establishment of two network protocols, the Transmission Control Protocol (TCP) and the Internet Protocol (IP), also known by the abbreviation TCP/IP, a definition of the term Internet was possible. Since then, the word Internet describes TCP/IP networks, that is networks connected to each other via the communication protocols, TCP/IP.

Tim Berners-Lee, a computer scientist at the European Organization for Nuclear Research (CERN), developed the World Wide Web (WWW) in 1990/1991 (Berners-Lee et al. 1999). The WWW has rapidly become very popular and today, together with e-mail, dominates use of the Internet. The terms WWW and Internet are used interchangeably in general language. However, WWW is actually only a communications protocol placed on top of the Internet, or, more simply, a service on the Internet. Another service besides WWW and e-mail is the File Transfer Protocol (FTP), which plays a large role in bioinformatics. The success of the WWW is certainly due to its simple usability and the combination of different communication protocols. Modern Internet browsers, e.g. Microsoft's Internet Explorer or the Open Source browsers of the Mozilla family, not only master HTTP, the communication protocol of the WWW, but also other protocols such as FTP and thus offer a comfortable communications platform.

Originally, the WWW was developed for the exchange of scientific data and it is, therefore, not surprising that the WWW is also important in bioinformatics: biological databases are provided to scientists over the WWW, bioinformatic analyses are performed, literature is searched for and read and information on bioinformatics and biology is offered. In retrospect, it could be argued that the automation of laboratory processes such as sequencing together with the development and the spread of the Internet and the WWW have made possible the success of bioinformatics in the first place.

1.3 The Physical Connection to the Internet

The first step to using the WWW or one of the other services is to create a physical connection between one's own computer and the Internet. There are several ways to do this. A common method is to use a modem which allows a connection to the Internet over the existing copper telephone line. Modem connections permit transmission rates of at most 56 kilobits per second (kbps) and are thus relatively slow. Besides the conventional analog modem, a number of other technical solutions were developed for data transmission. One of those that has gained a relatively high distribution rate in Europe is the Integrated Services Digital Network (ISDN). ISDN technology has optimized the network components (switches etc.) for the transfer of digital data and thereby permits a transmission rate of 64 kbps per channel. By simultaneously using several channels, the transmission rate can be extended. Since a connection fee arises for the use of every channel, the usage of multiple channels leads to a multiplication of those fees. For this reason bundling of channels has not gained acceptance for private use. ISDN uses a different wiring technology in comparison with analog telephony so that, in principle, new wiring would be necessary. With the help of special technology, however, the use of analog lines is also possible.

Another technology that uses the copper telephone line is the Digital Subscriber Line (DSL). At present, DSL permits

a transmission rate of up to 9 megabits per second (Mbps). The maximal transmission rate depends on the line length, however. Therefore, asynchronous DSL (ADSL) is normally used. Here, the data transmission to the user is carried out at a considerably higher speed than the data transmission from the user to the switching center. Since users usually download more information than they feed into the net, ADSL allows the optimal use of the capacities. The ability to use the available telephone line as well as the simultaneous use of digital (computer) and analog (phone) communications is the greatest advantage of DSL or ADSL, respectively. The latter is achieved by using different frequency ranges of the frequency spectrum for the two sorts of communication. A special splitter separates the incoming information and either passes the data communication on to a computer or the voice communication to the phone. The connection to the computer is via a Twisted Pair Ethernet connection. Ethernet is the standard that is normally used in greater networks (companies, universities etc.). DSL is very popular nowadays. Between March 2005 and March 2006, the number of DSL connections increased worldwide by 39% – up to 150 million [dsl news]. While the distribution of DSL is increasing, an even faster technology called VDSL is on the horizon. It permits transfer rates up to 50 Mbps and, thus, a download speed of 6.5 MB/s. Currently, there are no applications from single users that would require such a data rate. However, with new developments such as IP-TV, the television reception over the Internet, video-on-demand and the digital videotape library, such data rates will gain popularity in the future.

The broadband cable network (cable television) represents another technology for the fast transfer of digital communication. Its maximum transmission rate is approximately 10 Mbps. The connection with the broadband cable network is via a dedicated cable modem that is connected to the computer by Twisted Pair Ethernet. Similar to ADSL, an asynchronous data rate is also maintained in the broadband cable network, i.e., the speed at which data can be downloaded from the Internet onto the user's computer very much exceeds the speed at which the user can upload data.

1.4 The Logical Connection to the Internet

When the computer is physically attached to a communication network via modem, ISDN, DSL or any other method as described in the previous section then the next step is the logical connection to the Internet. This is achieved via an Internet Service Provider (ISP) that is connected to the Internet over a dedicated line and oversees the relaying between its customers and the Internet. Most ISPs offer a flat-rate fee that covers the Internet access costs.

Before one connects to the Internet, one should consider data security and virus protection. Any computer that is registered on the Internet over a service provider is a *de facto* component of the Internet. This means that the computer is visible on the Internet at the moment the connection is established. Depending on the operating system, different security holes can be exploited by potential attackers ("hackers") to obtain rights on a foreign computer. The aim of such attacks is often to delete or change data, to block computers or use them for illegal purposes. As protection, one should consider the installation of a firewall that prevents the unauthorized access of computers behind the firewall from the outside. For PC users, firewalls have been developed that consist of a software solution which is installed on the computer and protects it. The constant updating of active anti-virus software is also very much recommended.

1.5 Internet Services

1.5.1 E-mail

E-mail, besides the WWW, is probably the most frequently used service on the Internet and also might have been the first contact medium for most users of the Internet before the WWW. E-mail offers a number of advantages over conventional communication

by letter, fax or phone. In comparison with conventional mail, e-mail is fast with a very low cost. However, e-mail also has disadvantages. Prior to receipt of an email, the message is routed via many intermediate servers and for operators of these servers it is technically possibly to read the information. Therefore, confidential information should never be sent by e-mail without encryption. Despite this risk, e-mail has gained acceptance as a main communication form in daily life, in particular, in academic and corporate environments.

An e-mail address is usually described as follows: *user@computer.domain*. User is the name of the respective user and *computer.domain* describes the computer in which the corresponding user account is localized, for example, *john.doe@company.com*. In the header of an e-mail are found the addresses of the sender and recipient. A subject line follows as well as some information about the route taken by the e-mail. Most modern e-mail programs conceal this information by default and only show the name of the sender, the recipient and the subject line. Depending on the operating system, there is a wide choice of e-mail programs users to select. The programs vary both in their handling and in functional range.

Providers of e-mail accounts on the WWW, e.g., Lycos, Yahoo, Freenet, Web.de, GMX etc. often allow e-mail access via a normal WWW browser. Thus, it is possible to read and write e-mails on any computer that has access to the Internet. However, e-mail can only be accessed as long as the computer is connected to the Internet – it is not possible to read e-mails offline.

In the past, many computer viruses and worms have spread via e-mail, exploiting security gaps within e-mail programs. One should always be conscious of this risk when using e-mail. Therefore, in addition to the already mentioned personal firewall, an incoming e-mail, especially from an unknown sender, should be handled with care.

Besides personal communication, e-mail also enables discussion within larger groups, e.g., mailing lists. Contributions to a discussion are sent to a special e-mail account which then distributes the news automatically to all registered recipients on the list.

An example of such a mailing list is the Computational Chemistry List (CCL) [ccl]. A similar service is offered by *Usenet Newsgroups* or *News* for short. However, in this case, contributions to a discussion are not sent out to the subscribers by e-mail, but are posted on a black board. Retrieval of the information requires a special program, a news reader. Besides using special news readers which are similar to e-mail programs, it is also possible to read news on the WWW, e.g. at Google [google groups]. Newsgroups are organized hierarchically into categories which are part of the group name. The group name starts with the topmost category, after which, separated by dots, sub-categories follow until the name of the actual group finally appears. As an example, a typical group name *is bionet.molbio.embldatabank. Bionet* refers to the main category; *molbio* stands for molecular biology and corresponds to the first sub-category. The term *embldatabank* indicates that the discussions in the group are concerned with the EMBL DNA database. Some popular main categories are listed in Table 1.1.

Table 1.1. Important and popular main categories of USENET newsgroups

Category	Explanation
alt.	All sorts of subject areas
bionet.	Biological subject areas
biz.	Topics concerning products and services
comp.	Topics concerning computers, hardware, software, etc.
de.	German-speaking *newsgroups*
humanities.	Fine art, literature, etc.
misc.	Miscellaneous
news.	Discussions and information about News
rec.	Topics concerning recreational activities
sci.	Natural sciences, social sciences, etc.
soc.	Social questions, cultural aspects, etc.
talk.	Debates, discussions concerning current issues

Before contributing, the list of Frequently Asked Questions (FAQ) of the corresponding group should be read. All conventions of politeness (*Netiquette*) for that group are posted in the FAQs. Violations of these basic rules are perceived as impolite and may result in the automatic deletion of the news and exclusion of the originator from the group. The FAQs also indicate whether it is a moderated or non-moderated group. In moderated groups, every message is first read by a member of the moderator team and will only be published if it is within the discussion topic of the group and the respective netiquette is not violated.

1.5.2 FTP

The File Transfer Protocol (FTP) is a simple service between two computers for the transfer of data. A connection between the computer of the user and the server is established which lasts for the complete FTP session. To use FTP, accounts on both the server and the user's computer are required. This makes possible the transfer of personal data between two accounts on different computers. Publicly accessible data, e.g., the GenBank database, are frequently available also via FTP service over the Internet. For such data, it is not necessary that each user has an account on the server. Instead, the anonymous FTP system is used. No username and password combinations are required, rather, the terms *anonymous* or *FTP* are typed in. Instead of a password, one's complete e-mail address is entered to allow for statistical analysis by the operator of the server.

FTP was developed in the UNIX environment and is, therefore, text based. For Windows and Macintosh, many programs also use a GUI. These interfaces frequently follow the system components of the respective operating system in their appearance. For example, with Windows the appearance and operation resembles Windows explorer. In addition to special FTP programs, WWW browsers can also be used, at least for anonymous FTP. If one follows the link to an FTP server, the

browser will automatically send anonymous as the user name and the user's e-mail address as the password.

The directory structure of the FTP server is presented as an HTML document by the browser and navigation is performed by clicking on links. Files can be downloaded using the same procedures as on the WWW. The exact steps, however, are dependent on the browser used. Usually, a click with the right mouse button on the file and choosing *Save Target as…* should work.

Since the directory structure of an FTP server can be very confusing, many operators offer a README file in which the directory structure is explained. All FTP servers are similar in that files offered for download are found below the directory. In addition, some servers offer links to important lists or files already in the main directory.

In Fig. 1.3, a text-based FTP session with the FTP server of the National Center for Biotechnology Information (NCBI) [ncbi] is shown as an example. In this session, the current README file for the sequence database GenBank is downloaded from the server. The user's input is highlighted in bold for clarification. With the first input a connection to the FTP server is established. Under Windows this command can be entered using a DOS window or a command shell. `anonymous` is entered as the user name, and one's e-mail address as the password. The command `ls` generates a table of contents of the current directory. With `cd genbank`, one changes to the directory genbank. All files of the sequence database genbank can be found in this list. With the next command (`ascii`) the FTP server is put into ASCII mode. In the FTP protocol there are two different transfer modes, ASCII and binary. ASCII is used to transfer plain text files whereas binary mode allows the transmission of binary files such as executable programs, pictures or video files. Modern servers usually recognize the file type and choose the right transfer mode automatically, but the appropriate mode may be entered manually by choosing `ascii` or `bin`. The command get README.genbank eventually downloads the file *README.genbank* into the current directory of the local computer. With the command `bye` or `quit`, the FTP session is closed.

```
> ftp ftp.ncbi.nih.gov
331-(----GATEWAY CONNECTED TO ftp.ncbi.nih.gov----)
331-(220-)
331-( Warning Notice!)
331-( )
331-( This is a U.S. Government computer system, which may be accessed and used)
331-( only for authorized Government business by authorized personnel.)
331-( Unauthorized access or use of this computer system may subject violators to)
331-( criminal, civil, and/or administrative action.)
331-( )
331-( All information on this computer system may be intercepted, recorded, read,)
331-( copied, and disclosed by and to authorized personnel for official purposes,)
331-( including criminal investigations. Such information includes sensitive data)
331-( encrypted to comply with confidentiality and privacy requirements. Access)
331-( or use of this computer system by any person, whether authorized or)
331-( unauthorized, constitutes consent to these terms. There is no right of)
331-( privacy in this system.)
331-( )
331-( ------------------------------------------------------------------------)
331-( )
331-( Welcome to the NCBI ftp server. The official URL for NCBI ftp server is)
331-( "ftp://ftp.ncbi.nih.gov", please use it.)
331-( )
331-( Public data may be downloaded by logging in as "anonymous" using your E-
mail)
331-( address as a password.)
331-( )
331-(220 FTP Server ready.)
Name: anonymous
331 Anonymous login ok, send your complete email address as your password.
Password: **************************
230 Anonymous access granted, restrictions apply.
Remote system type is UNIX.
Using binary mode to transfer files.
ftp> ls
200 PORT command successful
150 Opening ASCII mode data connection for file list
dr-xr-xr-x  12 ftp      anonymous     4096 Feb  4 18:36 blast
dr-xr-xr-x  11 ftp      anonymous     4096 Nov 19 20:32 entrez
dr-xr-xr-x  11 ftp      anonymous    16384 Feb 14 18:10 genbank
dr-xr-xr-x  19 ftp      anonymous     4096 Dec 26 16:32 genomes
dr-xr-xr-x   8 ftp      anonymous     4096 Dec 17 12:44 mmdb
dr-xr-xr-x  80 ftp      anonymous     4096 Feb  4 15:10 pub
dr-xr-xr-x   3 ftp      anonymous     4096 Feb  8 04:19 pubmed
dr-xr-xr-x  11 ftp      anonymous     4096 Dec  6 21:00 refseq
dr-xr-xr-x  59 ftp      anonymous     4096 May 15  2002 repository
dr-xr-xr-x   7 ftp      anonymous     4096 Feb 12 17:50 sky-cgh
dr-xr-xr-x  20 ftp      anonymous     4096 Dec 12 15:18 snp
dr-xr-xr-x   2 ftp      anonymous     4096 Jan 26  1996 tech-reports
dr-xr-xr-x  11 ftp      anonymous     4096 Dec 27 13:58 toolbox
226 Transfer complete.
ftp> cd genbank
250 CWD command successful.
ftp> ascii
200 Type set to A
ftp> get README.genbank
local: README.genbank remote: README.genbank
200 PORT command successful
150 Opening ASCII mode data connection for README.genbank (14740 bytes)
226 Transfer complete.
15091 bytes received in 0.31 seconds (47.38 Kbytes/s)
ftp> bye
221 Goodbye.
>
```

Fig. 1.3. Anonymous FTP session for the transfer of a file (*README.genbank*) from the NCBI FTP server. The user's input is in bold. (Printed with permission of the NCBI)

The most important commands of the FTP protocol are listed as follows.

! !<command>
Executes the indicated command on the local computer.
Example:
List the contents of the local directory in compact form.
In the Windows operating system:
!dir/w
In the UNIX operating system:
`!ls`

ascii `ascii`
Switches the FTP server to the ASCII transfer mode.

bin `bin`
Switches the FTP server to the binary transport mode.

bye `bye`
Terminates the active FTP session.

cd `cd <directory>`
Changes to the indicated directory on the FTP server. Also see lcd.
Example:
Change on the server to the directory *pub*.
`cd pub`

get `get <file>`
Downloads the indicated file from the FTP server onto the local computer.
Example:
Transfer the file *transporter.seq* from the server to the local computer.
`get transporter.seq`

help `help`
Lists all FTP commands.

lcd `lcd <directory>`
Changes to the indicated directory on the local computer. Also see cd.

Example:
Change to the directory *sequencedata* on the local computer.
`lcd sequencedata`

ls `ls`
Lists the contents of the current directory on the FTP server.

mget `mget <wildcard>`
Downloads several files from the FTP server onto the local computer. In the interactive mode, the user is prompted before every file download. Otherwise, the non-interactive mode must be turned on by typing `prompt` before the `mget` command. Also, see `mput` and `prompt`.
Example:
Transfer all files that carry the ending *.tfa* from the server to the local computer.
`mget *.tfa`

mput `mput <wildcard>`
Uploads several files from the local computer onto the FTP server. In the interactive mode, a prompt is made prior to each file upload. Otherwise, the non-interactive mode must be turned on by typing `prompt` before the `mput` command. Also see `mget` and `prompt`.
Example:
Transfer all files that carry the ending *.txt* to the server.
`mput *.txt`

prompt `prompt`
Switches between the modes *interactively* and *non-interactively* when multiple commands are executed.

put `put <file>`
Uploads the indicated file from the local computer onto the FTP server.
Example:
Transfer the file *sequence.fas* from the local computer to the server.
`put sequence.fas`

pwd `pwd`
Shows the name of the current directory on the FTP server.

quit `quit`
Terminates the active FTP session.

1.5.3 World Wide Web

Despite its utility, FTP has a number of disadvantages. For example, in classic FTP sessions without a WWW browser, a file must first be loaded onto the local computer in order to see its content. Systems were soon developed that permitted viewing of the file contents in a client/server environment without prior download to the local computer (client). One of the better known systems was the text-based *Gopher* system, which could be regarded as a precursor of the WWW.

The special design of the WWW permits movement within the document structure without an understanding of its organization. The connection between different documents is carried out via hyperlinks. Hyperlinks are structures (text, pictures, icons etc.) that are embedded in a document and connected to further information. Upon clicking on a hyperlink the connected information is displayed. The information pertaining to the respective hyperlink or document may not be necessarily localized on the same server, but can be placed on any server. To allow for the unambiguous identification of the respective documents, a Uniform Resource Locator (URL) is used. The URL specifies both the required protocol (http, ftp, gopher) and the name of the server where the desired document is stored. The URL describes the complete path to the desired file. For example, the URL for the sequence database GenBank at NCBI is: http://www.ncbi.nlm.nih.gov/Genbank/index.html. The protocol used (http) is followed by a colon and two slashes, separating the protocol term from the server name. After the server, the complete path to the desired file

follows. The individual directory names are separated by slashes. For some WWW servers, a port number (an expansion of the network address) needs to be specified. The port number is inserted after the server name and is separated by a colon, e.g., http://www.ncbi.nlm.nih.gov:80/Genbank/index.html. The two addresses above only differ in terms of the port number. If the standard port number (80) is used its explicit statement is not necessary.

The ever-increasing number of WWW servers and ranges of products aggravates the search for relevant information. Many providers therefore offer link lists to interesting WWW servers, such as bioinformatik.de [bioinformatik]. Nevertheless, as it is frequently difficult to find desired information, the use of a search engine is recommended. Search engines are programs that search the WWW at regular intervals for new information, index those and put the resulting indices into databases. Table 1.2 lists some of the best-known search engines. The corresponding databases can then be searched very conveniently. Search engines use different strategies for indexing. Some engines index only terms that are

Table 1.2. Popular WWW search engines

Search engine	URL
Alltheweb	http://www.alltheweb.com/
Alta Vista	http://www.altavista.com/
ASK.com	http://www.ask.com/
Excite	http://www.excite.com/
Google	http://www.google.com/
Lycos	http://www.lycos.com/
MetaCrawler	http://www.metacrawler.com/
MSN Search	http://search.msn.com/
Yahoo	http://www.yahoo.com/

recorded in the header of the document, whereas others use all of the terms in the document. Other search engines also assess the frequency with which terms appear within a document, or they weight the words in the title and the remaining document, respectively. The differing strategies for indexing of WWW documents can lead to different search outputs between search engines using the same query. Therefore, it can make sense to use several search engines in parallel if the desired information is not found with the first search. Metasearch engines automatically send the query to several search engines, collect the results and then combine them. Thus, several indexing strategies are covered and a search as comprehensive as possible is carried out. A disadvantage, however, can be the very long list of results produced. Consequently, the search strategy should be adapted to the respective problem. Table 1.2 lists some of the best-known search engines.

1.6 Using UNIX

Although today's UNIX variants, including their functional extensions, possess very large sets of commands, only a small set of commands is required for routine operation. For most commands, the syntax is very similar and usually follows the general scheme:

```
command [options] [<input files>] [<output files>]
```

Which options for a respective command are available and whether an input or output file is expected is explained in the respective electronic manuals, the man pages. Man pages are relatively short instructions for the commands, which can be shown with the UNIX command

```
man <command>
```

Another interesting characteristic of UNIX are the wild cards. These are characters that can be used as placeholders for other characters and thus permit the composition of powerful operations (Table 1.3). Most UNIX commands execute the command in the standard setting without further enquiry, even

Table 1.3. UNIX wild cards

Wild card	Explanation
*	Placeholder for any number of any characters. Example: Delete all files in the current directory. Caution when using this wild card! `rm *`
?	Placeholder for any character. Example: Display all files from the list of the files dat1.fas, dat2.fas, dat6.fas and dat10.fas that contain only one digit. `cat dat?.fas`
[ ]	Placeholder for a character from the set of characters listed in the brackets. Example: Delete dat_a.fas, dat_b.fas, dat_e.fas and dat_z.fas but not dat_k.fas and dat_x.fas. `rm dat_[a-e,z].fas`

for destructive operations such as *rm*, *cp* or *mv*. Therefore, for destructive operations great care should be exercised when using wild cards and one should check first the effect of the chosen wild card with a harmless command, for example, with the *ls* command, to avoid any unintentional surprises.

Below some important UNIX commands are listed in short form, together with a short example of each. Recommended textbooks that deal with UNIX are listed in the references. (Chivers 2001, Reichard and Johnson 1995, Robbins 2005, Taglor 2001).

apropos `apropos <keyword>`

Looks for one or several keywords in the manual pages. Identical with the command `man -k`.

cat `cat [options] <files>`
Displays the content of the indicated files on the screen.
Example:
Show the content of the file sequence.fas on the screen.
```
cat sequence.fas
```
Combine the contents of the files seq1.fas, seq2.fas and seq3.fas in a file with the name sequence.fas.
```
cat seq1.fas seq2.fas seq3.fas >
sequence.fas
```

cd `cd <directory>`
Changes the current directory.
Example:
Change to the directory/usr/people/jdoe.
```
cd /usr/people/jdoe
```

cp `cp [options] <file1> <file2>`
`cp [options] <files> <directory>`
Copy file1 to file2 or copy several files into a directory.
Example:
Copy the file seq.fas from the current directory into a file with the name human.fas in the directory /scratch.
```
cp seq.fas /scratch/human.fas
```

file `file [options] <files>`
Shows the file type according to the data it contains. Possible types are ascii text, c program text, data, empty, directory and others.
Example:
Specify the type of the file /usr/bin/ls.
```
file /usr/bin/ls
```

grep `grep [options] <regular expression> <files>`
Searches in the mentioned files for lines containing the regular expression, in the simplest case a character string.
Example:
Show all lines that contain the character string ATG in the file sequence.fas.
```
grep ATG sequence.fas
```

head `head [options] <files>`
Displays the first N lines of one or several files. Without additional options $N = 10$.
Example:
Display the first 10 lines of the file sequences.fas.
`head sequence.fas`
Display the first 30 lines of the file seq.tfa.
`head -30 seq.tfa`

ls `ls [options] [<names>]`
Lists the contents of directories. If no name is given the content of the current directory is displayed.
Example:
Show the content of the directory/usr.
`ls /usr`
Displays the content of the current directory in detailed form.
`ls -l`
Show the content of the current directory and indicate executable files with an asterisk (*) and directories with a slash (/).
`ls -F`

man `man [options] <command>`
Displays the content of the manual page for the desired command.
Example:
Display the manual page for the command *man*.
`man man`
Display the manual page for the command *ls*.
`man ls`
Display the manual page for the command *cp*.
`man cp`
Display the manual page for the command *mv*.
`man mv`
Search for the keyword *mv* in all available manual pages.
`man -k mv` (identical with `apropos mv`)

mkdir `mkdir [options] <directory>`

Generates one or more new directories.
Example:
In the current directory generate another directory with the name *work*.
```
mkdir work
```
In the current directory generate the directories *work*, *bin* and *junk*.
```
mkdir work bin junk
```

more `more [options] <files>`
Displays the contents of the indicated files page-by-page on the screen.
Example:
Display the content of the file sequences.fas page-by-page.
```
more sequence.fas
```

mv `move [options] <source> <destination>`
Moves files or directories, or renames files or directories, respectively.
Example:
Move the file sequence.fas from the directory /usr/people/jdoe into the directory /usr/people/dduck
```
mv /usr/people/jdoe/sequence.fas
/usr/people/dduck/
```
Rename the file seq.fas to human.fas.
```
mv seq.fas human.fas
```

pwd `pwd`
Displays the complete pathname of the current directory.

rm `rm [options] <files>`
Deletes one or more files.
Example:
Delete the file seq.fas
```
rm seq.fas
```
Delete all files in the current directory that have the extension.fas.
```
rm *.fas
```

rmdir `rmdir [options] <directories>`
Deletes directories. Directories can only be deleted if they are empty. To delete whole directories including

their content the command `rm -r <directories>` can be used.
Example:
Delete the directory junk.
`rmdir junk`

tail `tail [options] <files>`
Displays the last N lines of one or several files. Without additional options $N = 10$.
Example:
Display the last 10 lines of the file sequences.fas.
`tail sequence.fas`
Display the last 50 lines of the file seq.tfa.
`tail -50 seq.tfa`

telnet `telnet [options] <computer name>`
Opens a connection with another computer via the Telnet protocol. The computer name can be given either as a name or as a numeric Internet address.
Example:
Open a connection with a computer named sever1.company.com
`telnet server1.company.com`

wc `wc [options] [<files>]`
Displays the number of characters, words and lines contained in the mentioned files.
Example:
Count the number of characters in the file sequence.fas.
`wc -c sequence.fas`
Count the number of files in the current directory.
`ls -1 | wc -l`

1.7 The History of Bioinformatics

The first algorithm for comparison of protein or DNA sequences was published by Needleman and Wunsch in 1970 (see Chap. 4). Bioinformatics is thus only one year younger than the Internet

progenitor ARPANET and one year older than e-mail which was invented by Ray Thomlinson in 1971. However, the term bioinformatics was only coined in 1978 (Hogeweg 1978) and was defined as the "studies of informatic processes in biotic systems". The Brookhaven Protein Data Bank (PDB) was also founded in 1971. The PDB is a database for the storage of crystallographic data of proteins (see Chap. 3). The development of bioinformatics proceeded very slowly at first. However, the publication of the complete gene sequence of the bacteriophage virus FX174 in 1980 coincided with the first use of the IntelliGenetics Suite – the first software package for the analysis of DNA and protein sequences. In the following year, Smith and Waterman published another algorithm for sequence comparison and IBM marketed the first personal computer (see Chap. 4). A year later, the University of Wisconsin founded a company, the Genetics Computer Group whose main product was the Wisconsin Suite, a software package for molecular biology. At first, both the IntelliGenetics and the Wisconsin Suites were packages of single, relatively small programs that were controlled via the command line. A GUI was later developed for the Wisconsin Suite, which made for more convenient operation of the programs. The IntelliGenetics Suite has since disappeared from the market, but the Wisconsin Suite is now available under the name, Accelrys GCG.

The publication of the polymerase chain reaction (PCR) process by Mullis and colleagues in 1986 represented a milestone in molecular biology and, concurrently, bioinformatics. In the same year, the SWISS-PROT database was founded and Thomas Roderick coined the term "genomics", describing the scientific discipline of sequencing and description of whole genomes. Two years later, the National Center for Biotechnology Information (NCBI) was established and, today, operates one of the most important primary databases (see Chap. 3). The same year also saw the start of the Human Genome Initiative and the publication of the FASTA algorithm (see Chap. 4). In 1991, CERN released the protocols that made possible the World Wide Web (WWW), and Greg Venter published the use of Expressed Sequence Tags (ESTs; see Chap. 5). By the next year, Venter and

his wife, Claire Fraser, had founded The Institute for Genomics Research (TIGR). With the publication of GeneQuiz in 1994, a fully integrated sequence analysis tool appeared that, in 1996, was used in the GeneCrunch project for the first automatic analysis of over 6,000 proteins of the baker's yeast, *Saccharomyces cerevisiae*, the genome of which had just been sequenced. In the same year, Bairoch and colleagues announced the start of the Prosite database (see Chap. 3). One year after the successful implementation of the GeneQuiz package for automatic sequence analysis, LION Biosciences AG was founded in Heidelberg, Germany. The root of one of LION's main products, the integrated sequence analysis package, termed bioSCOUT, was GeneQuiz. Together with other products of the Sequence-Retrieval System (SRS) package, LION Biosciences AG quickly became a very successful bioinformatics company worldwide. This did not last for long, however, and in 2006 the bioinformatics division was sold to BioWisdom, which continues to modify and sell SRS.

Twenty years after the term bioinformatics had been coined, another term, chemoinformatics, was published (Brown 1998). Until then, the terms chemometrics, computer chemistry and computational chemistry were common and are still in use today (Chen 2006). In November 2000, Metalife AG was founded in Winden near Freiburg, Germany. The company pursues a very modern concept through its bioinformatics product Metalife Trinity [metalife] (Fig. 1.4). Trinity offers three modules that cover the fields of database searching and knowledge management, nucleotide and protein sequence analysis, in addition to literature and patent data mining. The applications are set up in a two-step client server model in which the client application, administration and database server applications are written in the programming language C# making use of the *Microsoft .Net* framework. Behind the administration and database server is found a compute farm (consisting of a number of Linux computers) for bioinformatics applications. In contrast to the aforementioned packages, Metalife offers an unusual feature, an integrated semantic database named BioExplorer. This database combines the most important primary and secondary databases

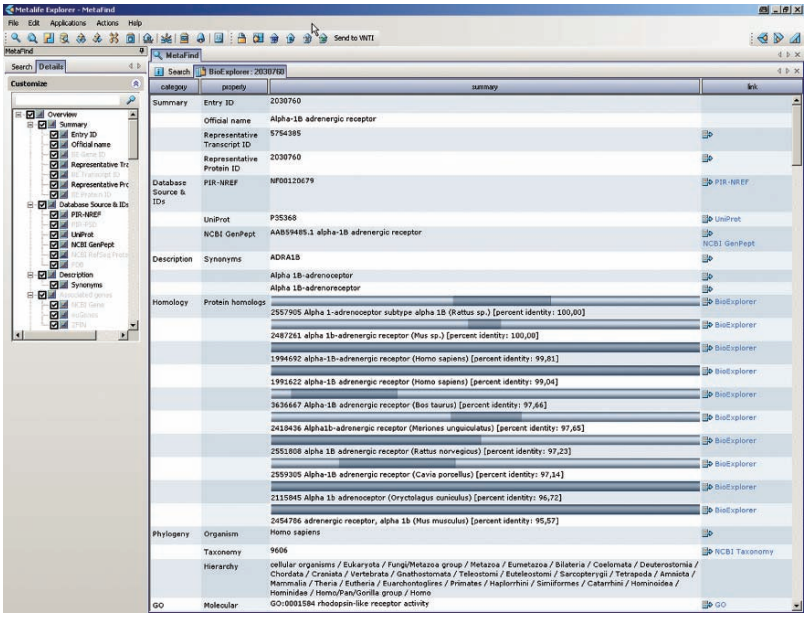


Fig. 1.4. Screenshot of the Metalife Trinity Software Summary View. (Printed with permission of the Metailfe AG)

(see Chap. 3) so that the user can see all of the relevant information, e.g. of a protein sequence, at a glance. It is thus no longer necessary to compare the entries of several databases and filter out redundant information. The BioExplorer database, all primary and secondary databases and the results of the analyses are saved by Metalife in a relational database (e.g. the Microsoft SQL server) on the database server.

The years 1997 and 1998 can be considered as milestones in bioinformatics and molecular biology because the genomes of two important model organisms, *Caenorhabditis elegans* and *S. cerevisiae*, were published. Also, the company Celera was founded by Greg Venter. The same is true for the year 2000 when the genomes of two other model organisms, *Arabidopsis thaliana* and *Drosophila melanogaster*, were completed. The next year saw the publication of the first draft of the human genome, which was eventually finished in 2004 (Fig. 1.5). Further important genomes that have been sequenced and published are those of the mouse, *Mus musculus* (2002); the causative agent of human malaria, *Plasmodium falciparum* (2002),

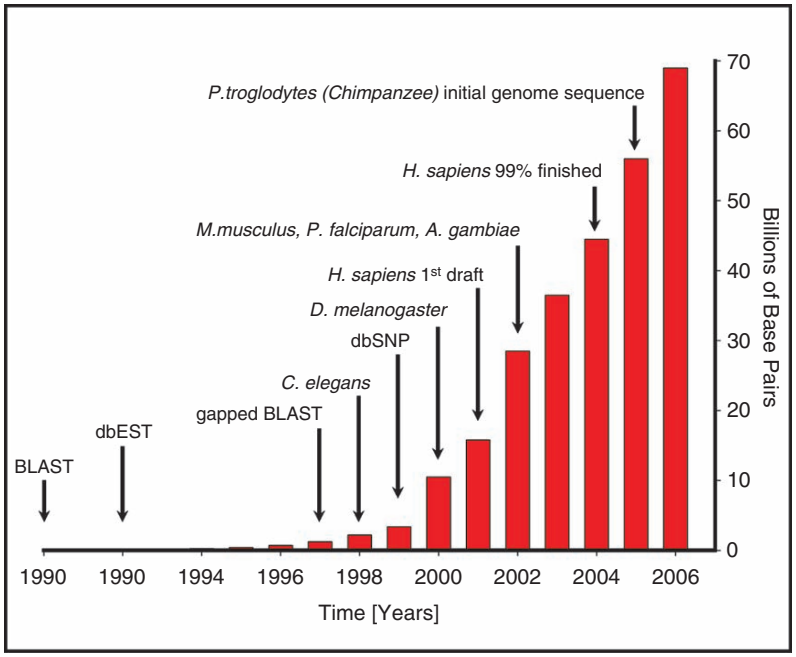


Fig. 1.5. Development of NCBI's GenBank database in connection with some milestones of bioinformatics

and its mosquito vector, *Anopheles gambiae* (2002); the brown rat, *Rattus norvegicus* (2004) and the chimpanzee, *Pan troglodytes*, in 2005. The sequencing of other genomes is an ongoing process and to list them all would go beyond the scope of this short survey. An overview of the completed and ongoing genome projects can be found in the Genomes Online [gold] and Genamics GenomeSeek databases [genamics].

1.8 Exercises

The exercises in the following chapters require Internet access as well as a valid e-mail address. In the following exercises two e-mail accounts are set up. If you already have an e-mail address you can skip exercises 1 and 2 and start directly with the exercises for using UNIX

1. You require a valid e-mail address for some of the exercises. Set up two e-mail accounts with two different providers.
2. Send an e-mail from one of the e-mail accounts to the other address, then go to the second e-mail account and check the receipt of the e-mail. Reply to this e-mail.

   The following exercises explain how to work with the operating system UNIX. A simple Telnet access to a UNIX computer is sufficient for the completion of the exercises. Many universities have a computer pool that also encompasses UNIX systems. If you already have a Linux system running on your computer you can skip these exercises.
3. Log onto the UNIX computer with your personal account and open a UNIX shell, if necessary.
4. List the table of contents of your home directory. The home directory is the current directory immediately after the login process.
5. Check the path of the current directory.
6. Copy the file */usr/motd* into your home directory.
7. Which option prevents the copy command (*cp*) from overwriting existing files of the same name in the target directory?
8. Rename the file *motd* in your home directory to *current*.
9. In your home directory create a directory with the name *message-of-today*.
10. Move the file *current* into the newly created directory.
11. Change to the new directory and display the contents of the file *current* on the screen.
12. Create another directory in your home directory with the name *ftp-download* and change to this directory.
13. Download three arbitrary files that have the extension *.dat* from the directory */pub/databases/embl/release* via *ftp* from the EBI FTP server (ftp.ebi.ac.uk). Note: Many FTP servers

store *ASCII* files in a compressed form, e.g. the *gnu zip* format (extension *.gz*). If the file extension of the compression program (in this case *.gz*) is left out in the file name, the files are expanded automatically before the actual download. Before you start with the download make sure that this is allowed on your computer.

14. Display the first 35 lines of one of the three files on the screen.
15. Display the lines of the three files which contain the term *contig*.
16. How many lines do the three files possess and how many lines contain the term *sequence*?
17. Change to the home directory and delete the directory *ftp-download*.

WWW Links

bioinformatics: http://www.bioinformatik.de/
ccl: http://www.ccl.net/
dsl news: http://www.dsl-news.de/dsl-meldung-2157.htm
faq: http://www.faqs.org/faqs/
genamics: http://genamics.com/genomes/index.htm
gold: http://www.genomesonline.org/
google groups: http://groups.google.com/
metalife: http://www.metalife.com/
microsoft lifesciences: http://www.microsoft.com/lifesciences/
ncbi: http://www.ncbi.nlm.nih.gov/

Literature

Berners-Lee T, Fischetti M, Dertouzos ML (1999) Weaving the web: the original design and ultimate destiny of the World Wide Web by its inventor. Harper, San Francisco, CA

Brown F (1998) Chemoinformatics: what is it and how does it impact drug discovery. Annu Rep Med Chem 33:375–384

Chen WL (2006) Chemoinformatics: past, present, and future. J Chem Inf Model 48:2230–2255

Chivers I (2001) Essential Linux fast, Springer, Berlin
Hogeweg P (1978) Simulation of cellular forms. In: Zeigler BP (ed) Frontiers in System Modelling, Simulation Councils, 90–95
Reichard K, Johnson EF (1995) Teach yourself… UNIX, MIS Press, New York
Robbins A (2005) Unix in a Nutshell, O'Reilly, Sabastopol, CA
Taylor D (2001) Teach yourself Unix in 24 hours. SAMS, Indianapolis, IN

2 The Biological Foundations of Bioinformatics

2.1 Nucleic Acids and Proteins

Nucleic acids and proteins are two important classes of macromolecules that play crucial roles in nature. Deoxyribonucleic acid (DNA) is the carrier of genetic information and ribonucleic acid (RNA) is involved in the biosynthesis of proteins that control the cellular processes of life. The basic constituents of nucleic acids are nucleotides while those of proteins are amino acids.

2.2 Structure of the Nucleic Acids DNA and RNA

The structure of the nucleotides is alike in DNA and RNA (Alberts et al. 2007). The nucleotides consist of a pentose, a phosphoric acid residue, and a heterocyclic base. Nucleotides are linked via chemical bonds between the pentose sugar of one nucleotide and the phosphoric acid residue of the next (see Fig. 2.1). Accordingly, the basic framework of nucleic acid is a polynucleotide where the phosphoric acid forms an ester bond between the 3′ OH group of the sugar residue of one nucleotide and the 5′ OH group of the sugar of the next nucleotide. At one end of the polynucleotide chain, therefore, a phosphate group is connected to the 5′ oxygen of a pentose sugar, whereas at the other end, a free 3′ hydroxyl group is present (see Fig. 2.1).

Each unit of the basic ribose/phosphoric acid residue structure carries a heterocyclic nucleobase that is connected

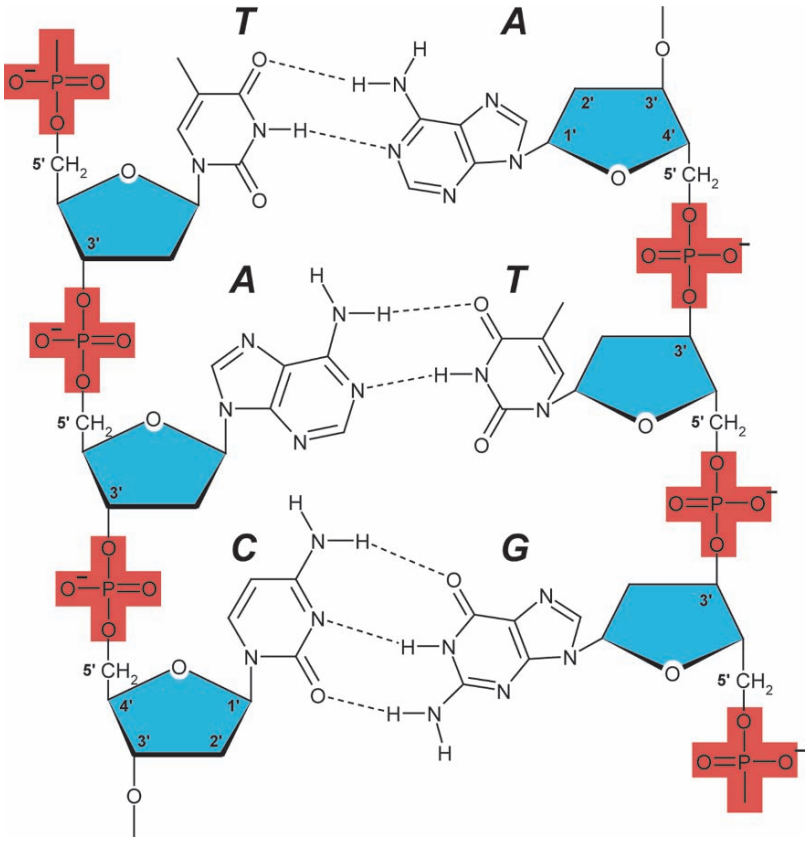


Fig. 2.1. The composition of nucleic acids

to the sugar residue via an *N*-glycosidic linkage. The nucleic acids consist of five different bases (cytosine, uracil, thymine, adenine, and guanine); however, uracil and thymine are found only in RNA and DNA, respectively. Nucleotides may be abbreviated using the first letter of the respective base, and their succession indicates the nucleotide sequence of the nucleic acid strand. DNA and RNA not only differ in their bases, but their respective sugar residues also differ in chemical composition. In RNA, the sugar is a ribose whereas in DNA, 2-deoxyribose is incorporated. DNA consists of two nucleotide strands that

combine in an antiparallel orientation so that hydrogen bonds are formed between the bases of each strand, thus, resulting in a ladder-like structure.

The bases are paired so that a purine ring on one strand interacts with a pyrimidine ring on the opposite strand. Two hydrogen bonds exist between A and T and three between G and C. The two nucleotide strands making up DNA are "complementary" to one another. Therefore, the sequential succession of bases on one strand determines the base sequence on the other strand. Under physiological conditions, DNA exists as a double helix in which the two polynucleotide strands wind right-handedly around a common axis. The diameter of the double helix is 2 nm. Along the double helix, opposing bases are 0.34 nm apart and rotated at an angle of 36° to one other. The helical structure recurs every 3.4 nm and corresponds to 10 base pairs (Watson and Crick 1953a,b).

2.3 The Storage of Genetic Information

The base sequence is the only variable element on the nucleotide strand and, therefore, encodes the necessary information to generate proteins. Proteins comprise varying amounts of up to 20 amino acids and each amino acid is encoded by a triplet of bases, termed codons. If doublet codons were to be used to encode proteins, the resulting $4^2 = 16$ possible combinations would be insufficient to generate 20 amino acids. On the other hand, triplet codons give $4^3 = 64$ possibilities, allowing for more combinations than necessary to encode 20 amino acids. From these theoretical calculations, one can infer that an individual amino acid may be encoded by more than one codon. Therefore, the resulting genetic code is described as being degenerate. The genetic code shown in Fig. 2.2 applies universally to all living organisms; however, some exceptions can be found in mitochondria and ciliates.

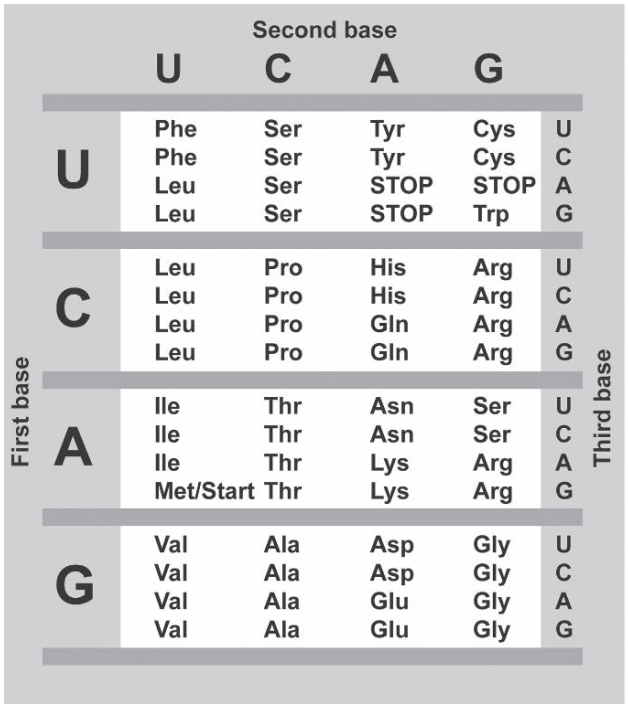


Fig. 2.2. The genetic code

The relationship between DNA, RNA, and proteins is described as the central dogma of molecular biology (Crick 1970) (Fig. 2.3). The entirety of genomic DNA in any organism is described as a genome, whereby the genetic information is encoded in the DNA as the sequence of its bases. This information is transferred to the messenger RNA (mRNA) during the process of transcription. The unambiguous transfer of information is guaranteed by the pairing of complementary bases. The total pool of mRNA in any organism is described as a transcriptome. The process of building proteins from mRNA is called translation. Analogous to the terms genome and transcriptome, the entire pool of proteins in any organism is referred to as the proteome. Overall, therefore, the amino acid composition of proteins is determined by the DNA sequence and this is usually the flow of information.

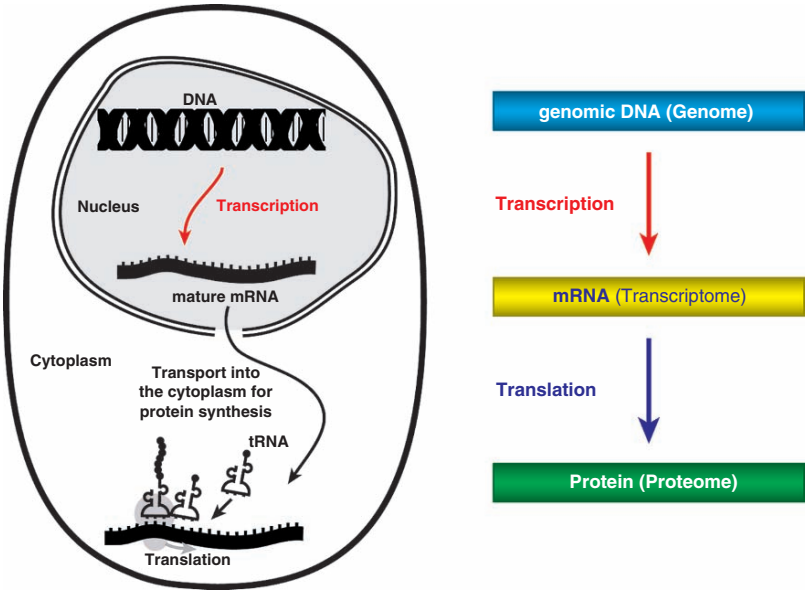


Fig. 2.3. The central dogma of molecular biology. The flow of information always proceeds from the genome to the proteome and not vice versa. Exceptions are reactions that are catalyzed by the reverse transcriptase and replicase of RNA viruses

However, RNA viruses are an exception. They can transcribe their RNA into DNA with the help of a reverse transcriptase, and replicate RNA by means of a replicase.

Thus, a genome comprises genes that contain the information to build proteins. The organization of a gene region, however, is different in prokaryotes from eukaryotes (Fig. 2.4). The most striking difference is that prokaryotic gene information is encoded on a continuous DNA stretch, whereas in eukaryotes, coding exons are interrupted by noncoding introns (Lewin 2007). Eukaryotic transcription of DNA to mature mRNA (containing information derived only from exons) requires several steps. The introns are removed during the process of splicing. Through alternative splicing (removing and joining different introns and exons) different mRNAs and, consequently, different proteins can result from one gene. Alternative splicing, among other mechanisms, explains why

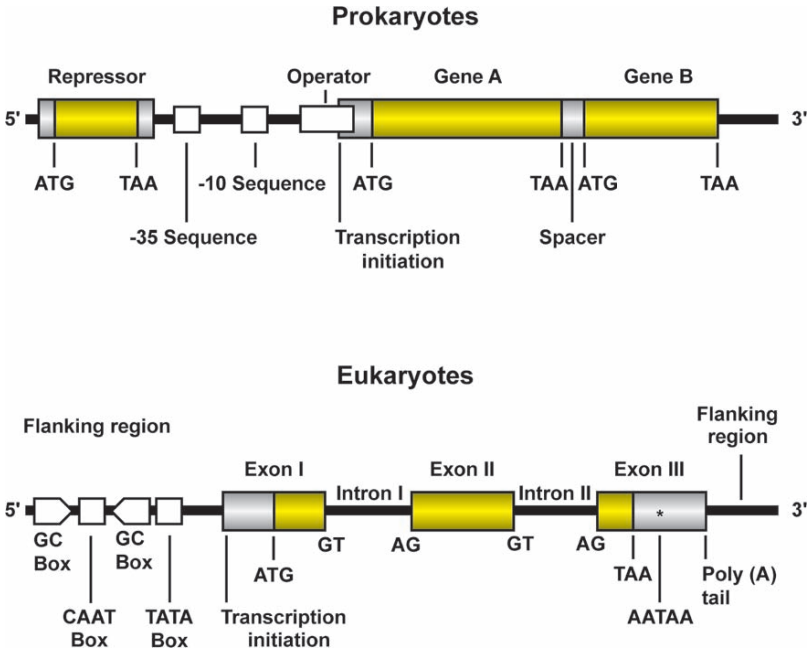


Fig. 2.4. The structure of gene regions of prokaryotes and eukaryotes

a relatively low number of genes are found in the human genome compared to the greater number of proteins actually produced (Claverie 2001, Venter et al. 2001).

2.4 The Structure of Proteins

2.4.1 Primary Structure

As mentioned, proteins are macromolecules comprising the 20 naturally occurring amino acids. Under physiological conditions, proteins fold into characteristic three-dimensional structures

that dictate their biological properties (Berg et al. 2006). The common configuration of natural amino acids is characterized by an amino and a carboxyl group around a central α carbon atom. The respective side chain of each amino acid determines the chemical properties, such as hydrophobic, polar, acidic, or basic (Fig. 2.5). If the characteristics of a protein were to depend solely on the unfolded amino acid sequence (frequently referred to as the primary structure), similar properties would be expected due to the limitation of just 20 amino acids. Indeed, denatured (unfolded) proteins have very similar properties that correspond essentially to a homogeneous cross-section of randomly distributed side chains. Nevertheless, the primary structure is essential for determining secondary and tertiary structure and with that, the three-dimensional conformation of the protein.

Peptide bonds connect individual amino acids in a polypeptide chain. Each amino acid is linked via the acid amide bond of its α carboxyl group to the α amino group of the next. Consequently, polypeptides have free N- and C-termini.

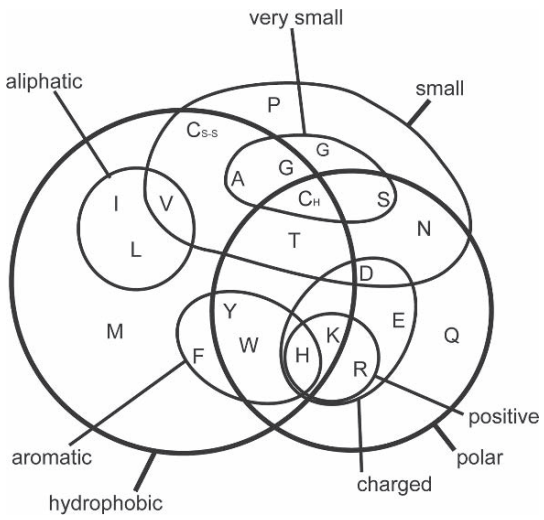


Fig. 2.5. Venn diagram of the properties of the amino acids

Table 2.1. The amino acids

Amino acid	3-Letter code	1-Letter code
Alanine	Ala	A
Cysteine	Cys	C
Aspartic acid	Asp	D
Glutamic acid	Glu	E
Phenylalanine	Phe	F
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Lysine	Lys	K
Leucine	Leu	L
Methionine	Met	M
Asparagine	Asn	N
Proline	Pro	P
Glutamine	Gln	Q
Arginine	Arg	R
Serine	Ser	S
Threonine	Thr	T
Valine	Val	V
Tryptophan	Trp	W
Tyrosine	Tyr	Y

Polypeptide primary structure, i.e., the amino acid sequence from the N- to the C-terminus, can contain between three and several hundred amino acids. Each amino acid in the polypeptide chain is abbreviated either with a three letter or one letter code (see Table 2.1).

2.4.2 Secondary Structure

The term secondary structure defines the local conformation of the backbone of any polymer. In the case of proteins, secondary structure describes the ordered folding patterns of polypeptide chain into helices (α-helix), sheet structures (β-strand), and loops. The key to understanding these more complex structures lies in the geometric properties of the peptide group. Linus Pauling and Robert Corey demonstrated in the 1930s and 1940s that the peptide bond is a rigid, planar structure, which can be attributed to the 40% double bond character of the peptide bond. Accordingly, a polypeptide chain can be regarded as a sequentially linked chain of rigid and planar peptide groups. The chain conformation of a polypeptide can therefore be determined by the torsion angles around the C_α–N binding (ϕ) and the C_α–C binding (ψ) of the constituent amino acid residues. In the planar and fully stretched (all *trans*) conformation, all angles are 180°. Viewed from the C_α atom the angles increase with a clockwise rotation. Not all conceivable values for ϕ and ψ are possible, however, mainly due to steric hindrance caused by the side chains of the amino acids. A Ramachandran plot is a conformation chart of those values that are sterically possible for ϕ and ψ (Fig. 2.6). Areas in the Ramachandran plot that correspond to sterically possible values of angles ϕ and ψ are called permissible areas; those corresponding to values that are not possible are called forbidden areas.

As already mentioned, three components in the secondary structure of proteins can be distinguished, the α-helix, the β-strand, and loops. The polypeptide chain of an α-helix displays a pitch of 0.54 nm with 3.6 residues per turn. As for α helices, β-strands are stabilized by hydrogen bonds. However, they are not found within a polypeptide chain as in the case of a helix, but between neighboring strands. Such β-strands exist in both parallel and antiparallel forms due to the direction of the polypeptide chain. In β-strands, each successive side chain is on the opposite side of the plane of the sheet, with a repetition unit of two residues and at a distance of 0.7 nm. On

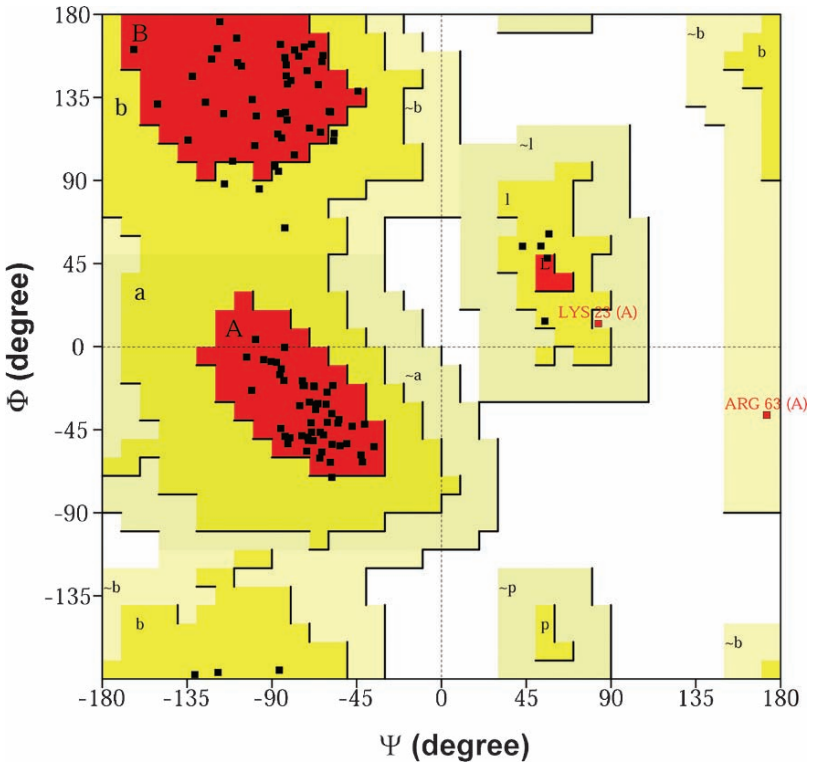


Fig. 2.6. Ramachandran plot of the transcription regulator protein GAL4 from *Saccharomyces cerevisiae*. The amino acids are represented as small black squares. Evidently almost all amino acids lie in preferred, permissible areas (*red* and *yellow*). Two amino acids (LYS23 and ARG63) are found in slightly forbidden areas of the Ramachandran plot. This means that the combination of the values for ψ and ϕ would theoretically not be possible due to steric hindrance of the neighboring side chains. However, in practice it can be observed. The plot was generated with the program PROCHECK (Laskowski et al. 1993, Rullmann 1996); plot statistics were deleted for clarity

average, a globular protein consists of approximately a half each of α-helices and β-sheets. The rest of the protein contains nonrepetitive loops. These frequently change their direction abruptly and so serve as connections between α-helix and/or β-strand elements (Fig. 2.7).

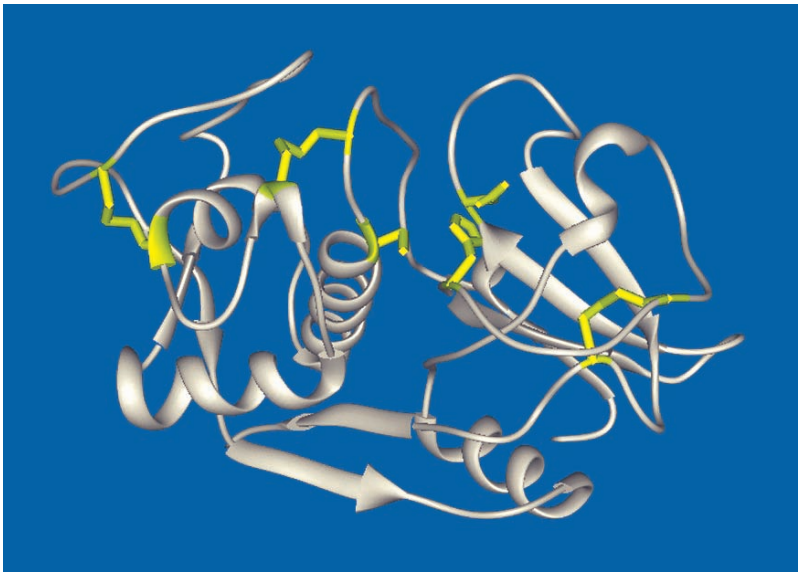


Fig. 2.7. Secondary structure representation (ribbon model) of a cysteine protease from the parasitic organism, *Leishmania major*. *Coils* symbolize α-helices and *arrows* β-strands. Disulfide bridges, which stabilize the three-dimensional structure of the protein, and the catalytically active amino acids (catalytic triad) are represented in *yellow*

2.5 Tertiary and Quaternary Structure

The tertiary structure describes the three-dimensional arrangement and placement of secondary structure elements. Large polypeptide chains (>200 amino acids) frequently fold themselves into several units termed domains. Normally such domains comprise 100–200 amino acids with a diameter of approx. 2.5 nm. The tertiary structure specifies the protein properties, e.g., whether a protein functions as an enzyme or as structural protein. Through the compaction of secondary structural elements and interactions between the amino acids of those elements, the structure of the protein is stabilized. The amino acid interactions include hydrogen bonds between

peptide groups, disulfide bonds between cysteine residues, ionic bonds between charged groups of amino acid side chains, and hydrophobic interactions. Quaternary structure is the arrangement of several polypeptide subunits. These are associated in a specific geometry so that a symmetrical complex is formed. The assembly of the individual subunits is carried out through noncovalent interactions.

2.6 Exercises

1. What is the difference between the two polynucleotides DNA and RNA?
2. DNA consists of two complementary nucleotide strands. Which base pairings are observed between these two nucleotide strands?
3. What is the meaning of the terms genome, transcriptome, and proteome?
4. The 20 naturally occurring amino acids are encoded by base triplets (codons) of the genetic code. Which consideration led to the discovery of the triplet codon organization of the genetic code?
5. Build the genetic code of your name. Should this not be possible, use the name *CRICK*.
6. What is meant by the central dogma of molecular biology?
7. What is meant by the term splicing and how does this process contribute to the discrepancy between the relatively low number of genes in the human genome but the larger number of proteins actually produced?
8. Which amino acids show the following properties: (A) hydrophobic, polar and small, and (B) hydrophobic and aliphatic?
9. In which direction is the primary structure of proteins read?
10. Which structural elements can be found in the secondary structure of proteins?

WWW Links

amino-acids: http://en.wikipedia.org/wiki/Amino_acid
biochemistry: http://en.wikipedia.org/wiki/Biochemistry
ncbi-books: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db= Books

Literature

Alberts B, Johnson A, Walter P, Lewis J (2007) Molecular biology of the cell. Garland Publishing, U.S.

Claverie JM (2001) What if there are only 30000 human genes? Science 291:1255–1256

Crick F (1970) Central dogma of molecular biology. Nature 227:561–563

Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 26:283–291

Lewin B (2007) Genes IX. Jones & Bartlett Publishers, U.S.

Rullmann JAC (1996) AQUA, computer program. Department of NMR Spectroscopy, Bijvoet Center for Biomolecular Research, Utrecht University, The Netherlands

Berg JM, Tymoczko JL, Stryer L (2006) Biochemistry. W.H. Freeman and Co, New York

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al (2001) The sequence of the human genome. Science 291:1304–1351

Watson JD, Crick FHC (1953a) Molecular structure of nucleic acids. Nature 171:737–738

Watson JD, Crick FHC (1953b) Genetical implications of the structure of deoxyribonucleic acid. Nature 171:964–967

3 Biological Databases

3.1 Biological Knowledge is Stored in Global Databases

The most important basis for applied bioinformatics is the collection of sequence data and its associated biological information. For example, with genome sequencing projects such data are generated daily in very large quantities worldwide. To use these data appropriately, a structured filing system of these data is necessary, yet the data should also be accessible to those interested. Annually, the journal Nucleic Acids Research [nar] dedicates an entire issue (first issue in January) to all available biological databases, which are recorded in tabular form with the respective URLs. Furthermore, for a number of databases, original articles describe their functions. This database issue, which is freely accessible also on the WWW, is a good starting point for working with biological databases. Depending on the kind of data included, different categories of biological databases can be distinguished. Primary databases contain primary sequence information (nucleotide or protein) and accompanying annotation information regarding function, bibliographies, cross-references to other databases, etc. Secondary biological databases, however, summarize the results from analyses of primary protein sequence databases. The aim of these analyses is to derive common features for sequence classes, which in turn can be used for the classification of unknown sequences (annotation). In addition, all other databases that save biological or medical information,

e.g., literature databases, are frequently classified as secondary databases.

The use of relational database systems (e.g., Oracle, MS Access, Informax, DB2, etc.), and their ability to manage large data sets, would seem ideal for the structured filing of data, yet these systems have not gained acceptance so far in the field of biological databases. Rather, sequence data and their accompanying information are usually filed in the form of flat file databases, i.e., structured ASCII text files. This is for historical reasons and because ASCII text files offer the advantageous ability to manipulate data without requiring an expensive and complicated database system. ASCII text files also make data exchange between scientists relatively simple. One drawback, however, is that searching for certain keywords within a dataset is both laborious and time-consuming. To minimize this disadvantage, various systems have been developed that can index flat file-based databases, i.e., they are provided with an index register similar to that of a book, thus accelerating keyword-based searches.

3.2 Primary Databases

3.2.1 Nucleotide Sequence Databases

GenBank

The GenBank database [genbank] is perhaps the best-known nucleotide sequence databases available at the US National Center for Biotechnology Information (NCBI) [ncbi]. GenBank is a public sequence database, which in its present version (161.00; August 2007) contains roughly 76 million sequence entries. The entry of sequences into GenBank can be performed by anyone via a WWW page [bankit], or by e-mail [sequin], when working with larger sequence sets. Prior entry of sequence data into either GenBank or one of its associated databases, for example, EMBL or DDBJ, is a prerequisite for the publication of new sequences

in any scientific journal. Each single database entry is provided with an unique identification tag, the accession number (AN). The AN is a permanent record remaining unchanged even if subsequent changes are made to the database record. In some cases, a new AN can be assigned to an existing number if, for example, an author adds a new database record that combines existing sequences. Even then the old AN is retained as a secondary number. The AN is the only way to absolutely verify the identity of a sequence or database entry.

Figure 3.1 shows a GenBank entry. The entry has been shortened at some points and these are indicated by [...]. The

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      baker's yeast.
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales;
            Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
[..]
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGT[..]RQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /function="required for axial budding pattern of S.
                     cerevisiae"
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /db_xref="GI:1293615"
                     /translation="MTQLQISLLLTATISLLH[..]PEML"
[..]
BASE COUNT     1510 a   1074 c    835 g   1609 t
ORIGIN
        1 gatcctccat atacaacggt atctccac[..]agctgttctc tcagctcctc atatttttct
     4981 tgccatgact cagattctaa ttttaagcta ttcaatttct ctttgatc
//
```

Fig. 3.1. Database record of the GenBank database. The entry was shortened at some points as indicated by [...]

required structuring of the database record is performed via defined keywords. Each entry starts with the keyword LOCUS followed by a locus name. Similar to the AN, the locus name is also unique; however, unlike the AN, it may change after revisions of the database. The locus name consists of eight characters including the first letter of the genus and species names, in addition to a six-digit AN. Newer entries have an eight-digit AN. In such cases, the locus name is identical to the AN. On the same line following the locus name, the length of the sequence is stated. A sequence must have at least 50 base pairs to be entered into GenBank. This requirement was introduced only relatively recently and, therefore, some older entries do not fulfill this criterion. Column 3 denotes the type of molecule of the sequence entry. Every GenBank entry must contain coherent sequence information of a single molecule type, i.e., an entry cannot contain sequence information of both genomic DNA and RNA. The last column in the LOCUS line gives the date of the last entry modification. The end of the database record starts with the keyword ORIGIN. In newer entries this field remains empty. The actual sequence information begins on the following line and may contain many lines. A detailed description of all keywords is found on the GenBank sample page [gb-sample].

Entrez

Query of the GenBank database is carried out via the NCBI Entrez system [entrez], which is used for the query of all NCBI-associated databases (Wheeler et al. 2007). Because search terms can be combined by means of logical operators (AND, OR, NOT) and single search terms restricted to certain database fields, Entrez is an important and effective tool for the execution of both simple and complicated searches. The restriction of search terms to single database fields is generally performed by a field ID placed after the term: `search term[field-id]`. For example, the search for a sequence from *Saccharomyces cerevisiae* with length between 3,260 and 3,270 base pairs would require the following search syntax: `Saccharomyces cerevisiae[ORGN] AND`

Table 3.1. Field IDs to restrict search terms to certain database fields in the Entrez system

Field ID	Database field
ACC	Accession number
AU	Author name
DP	Publication date
GENES	Gene name
ORGN	Scientific and common name of the organism
PT	Publication type, e.g. review, letter, technical publication
TA	Journal name, official abbreviation or ISSN number

`3260:3270[SLEN]`. Representative field IDs for performing searches in GenBank are listed in Table 3.1. Complete instructions for the use of Entrez are found on the Entrez help page [entrez-help].

EMBL and DDBJ

The European counterpart to GenBank is the European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL) [embl] located at the European Bioinformatics Institute (EBI) [ebi]. Another primary nucleotide sequence database, the DNA Database of Japan (DDBJ) [ddbj], is operated by the Center for Information Biology (CIB) [cib] in Japan and is the primary nucleotide sequence database for Asia. The three database operators NCBI, EBI, and CIB comprise the International Nucleotide Sequence Database Collaboration and synchronize their databases every 24 h. A query of all three individual databases is therefore not necessary, nor is it required to enter a new nucleotide sequence into all three databases.

While the database format of DDBJ is identical to that of NCBI, that of EMBL differs somewhat. Figure 3.2 shows an entry

```
ID   U49845; SV 1; linear; genomic DNA; STD; FUN; 5028 BP.
XX
AC   U49845;
XX
DT   07-MAY-1996 (Rel. 47, Created)
DT   17-APR-2005 (Rel. 83, Last updated, Version 4)
XX
DE   Saccharomyces cerevisiae TCP1-beta gene, partial cds; and Axl2p (AXL2) and
DE   Rev7p (REV7) genes, complete cds.
XX
KW   .
XX
OS   Saccharomyces cerevisiae (baker's yeast)
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
XX
RN   [1]
RP   1-5028
RX   PUBMED; 7871890.
RA   Torpey L.E., Gibbs P.E., Nelson J., Lawrence C.W.;
RT   "Cloning and sequence of REV7, a gene whose function is required for DNA
RT   damage-induced mutagenesis in Saccharomyces cerevisiae";
RL   Yeast 10(11):1503-1509(1994).
XX
[..]
FH   Key             Location/Qualifiers
FH
FT   source          1..5028
FT                   /organism="Saccharomyces cerevisiae"
FT                   /chromosome="IX"
FT                   /map="9"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:4932"
FT   CDS             <1..206
FT                   /codon_start=3
FT                   /product="TCP1-beta"
FT                   /db_xref="GOA:P39076"
FT                   /db_xref="HSSP:P48424"
FT                   /db_xref="UniProtKB/Swiss-Prot:P39076"
FT                   /protein_id="AAA98665.1"
FT                   /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEAA
FT                   EVLLRVDNIIRARPRTANRQHM"
[..]
XX
SQ   Sequence 5028 BP; 1510 A; 1074 C; 835 G; 1609 T; 0 other;
     gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg        60
[..]
     tgccatgact cagattctaa ttttaagcta ttcaatttct ctttgatc                   5028
//
```

Fig. 3.2. Database record of the EMBL database. The entry has been shortened at some points as indicated by […]

in the EMBL database. The most obvious difference is the use of two-letter codes instead of full keywords. Furthermore, there are small changes in the organization of the individual data fields. For example, the date of the last modification is not listed in the field ID (corresponding to the LOCUS field in GenBank), but appears in the field DT. A complete description of the EMBL format can be found on the EMBL manual page [ebi-manual].

The Sequence Retrieval System

Both DDBJ and EMBL offer two simple WWW forms for basic queries [ddbj-srs, ebi-srs] based on the Sequence Retrieval System (SRS). SRS was developed at EBI to manage primary and secondary biological databases (Etzold et al. 1996). SRS can also facilitate complex queries. Operation of SRS is the same at either DDBJ or EBI and the following section describes the system at EBI.

Searches of different databases can be initiated at the start page of the EBI-SRS. In the pull-down menu of the section *Quick Text Search,* select the database desired and then enter the search term in the text field on the right. In this case, all database fields will be searched without possibility of restriction. Several search words can be combined using the symbols of the logical operators *AND (&)*, *OR (|)*, and *NOT (!)*. On clicking the folder tab, *Library Page,* more complex database queries are possible. First, select and highlight the databases needed for the search. To select all databases of a subsection, click the *All* button (Fig. 3.3a). Simple search queries can then be entered into the text entry field in the upper part of the page. Click *Quick Search* to start the search. The SRS system automatically adds a wildcard at the end of the search term, e.g., a search query with *U4984* would not only find AN U49845, but also other terms that start with U4984. To have a wildcard inserted at the beginning of the search term, an asterisk (*) is required, e.g., *49845. Several search strings entered together are connected automatically with the logical *AND*; however, the remaining logical operators, *OR* and *NOT*, must be explicitly selected. For more complex queries, SRS offers two further query pages, the *Standard Query Form* and the *Extended Query Form*. The standard query form (Fig. 3.3b) has four text entry fields that can be restricted to certain database fields using the corresponding pull-down menus on the left side next to the text fields. The combination of the individual text entry fields is performed automatically via a logical *AND,* but this can be changed with the pull-down menu *combine searches with* on the left side of the page (highlighted in gray). Further possible

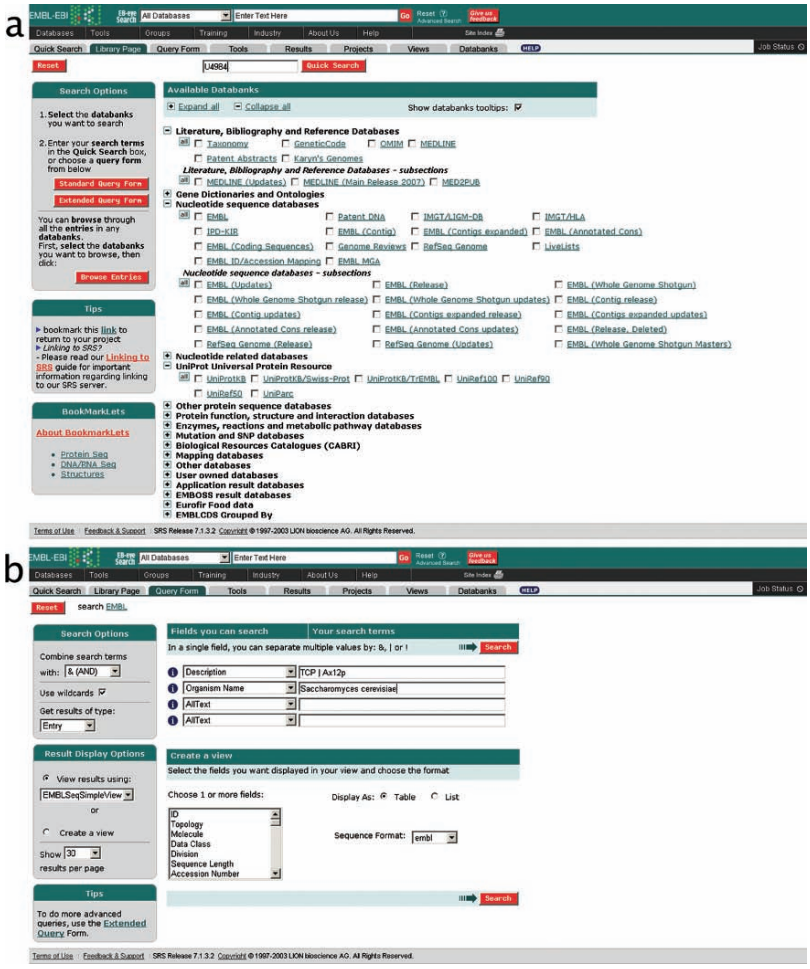


Fig. 3.3. Start page (**a**) and *Standard Query Page* (**b**) of the EBI-SRS server. http://srs.ebi.ac.uk/srs7bin/cgi-bin/wgetz?-page+top+-newId. (Printed with permission of the European Bioinformatics Institute.)

combinations are *OR* and *BUTNOT*. Also, it is possible to combine different terms within a text entry field with the logical operators *AND (&)*, *OR (|)*, and *NOT (!)*, thereby creating very complex queries.

The EBI-SRS also allows the establishment of a permanent project whereby performed searches can be saved. This allows one to interrupt a search at any time and continue it later. To

do this, select the link *Start A Permanent Project* on the EBI-SRS main page. For a *SRS user name,* any name can be used to save the search queries and continue them later. To return to the options page of the database, click on the folder tab *Library Page*. Its use is similar to that of a temporary project, but queries and their results are saved and can be accessed with the folder tab *Projects*. Detailed instructions on how to use the SRS can be found on the help page of the EBI-SRS [ebi-srs-help]. It can be also reached with a mouse click on the *Help* button in the upper right corner of the SRS page.

3.2.2 Protein Sequence Databases

SWISSPROT

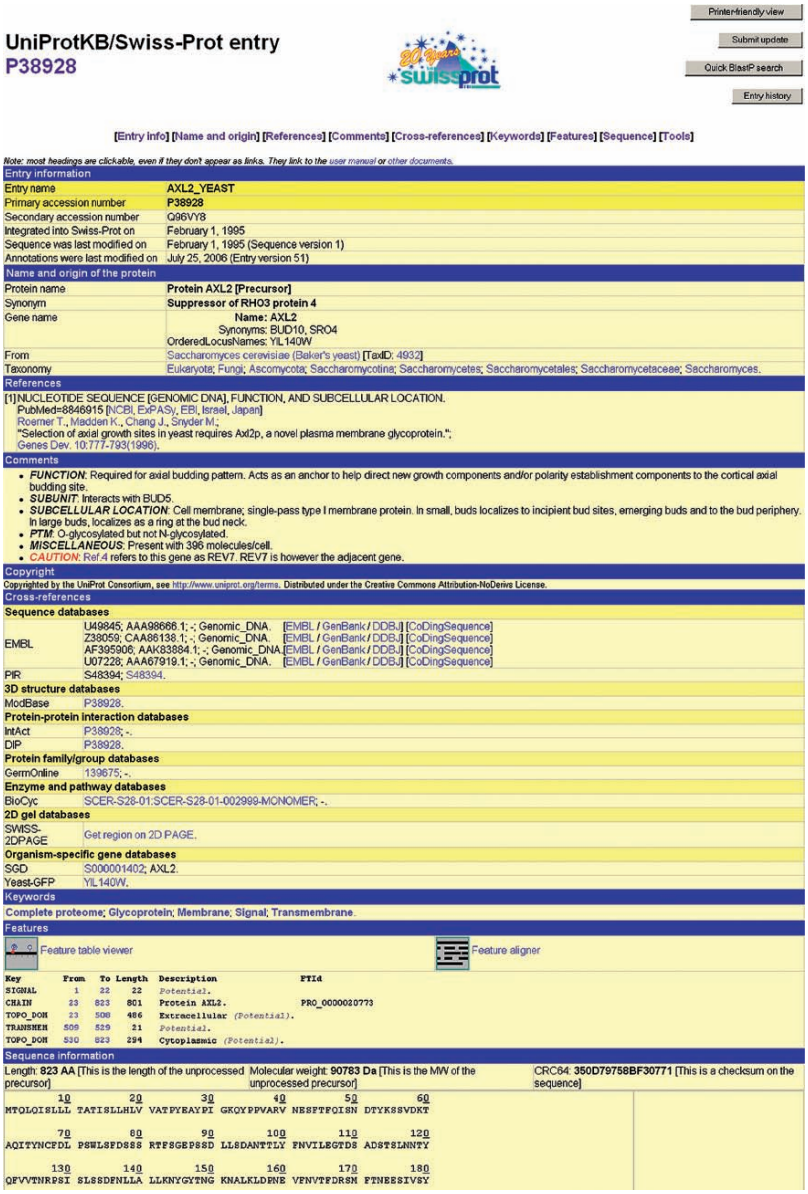
One of the most important collections of annotated protein sequences is the Swissprot database [swissprot] of the *Swiss Institute of Bioinformatics* (SIB), which also operates the *Expert Protein Analysis System* (Expasy) server [expasy]. The Swissprot database is manually curated, i.e., every database record is verified by a specialist and, if necessary, compared with literature data. This results in a high quality database that is considered a gold standard for protein annotation. Furthermore, Swissprot is part of the UniProt databases (see Sect. 3.2.2 – Uniprot) collectively known as the UniProt Knowledgebase (UniProtKB).

Figure 3.4 shows a database record in the Swissprot database. At first sight the entry seems to be similar to that of the EMBL *Nucleotide Sequence Database*. Indeed, the two database formats are related to one other in that the individual data fields in Swissprot start with a two-letter code. Most codes are identical to those of the EMBL database; however, some have been changed or new codes added.

In Swissprot, Expasy offers a special view of the database record, the NiceProt view (Fig. 3.5). Here, the individual parts of the database record are graphically edited for better readability and references to the original literature, and to other databases, are hyperlinked, thus simplifying its use.

```
ID   AXL2_YEAST     STANDARD;      PRT;   823 AA.
AC   P38928; Q96VY8;
DT   01-FEB-1995, integrated into UniProtKB/Swiss-Prot.
DT   01-FEB-1995, sequence version 1.
DT   25-JUL-2006, entry version 51.
DE   Protein AXL2 precursor (Suppressor of RHO3 protein 4).
GN   Name=AXL2; Synonyms=BUD10, SRO4; OrderedLocusNames=YIL140W;
OS   Saccharomyces cerevisiae (Baker's yeast).
OC   Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
OC   Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=4932;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA], FUNCTION, AND SUBCELLULAR LOCATION.
RX   PubMed=8846915;
RA   Roemer T., Madden K., Chang J., Snyder M.;
RT   "Selection of axial growth sites in yeast requires Axl2p, a novel
RT   plasma membrane glycoprotein.";
RL   Genes Dev. 10:777-793(1996).
[..]
RN   [8]
RP   LEVEL OF PROTEIN EXPRESSION [LARGE SCALE ANALYSIS].
RX   MEDLINE=22923965; PubMed=14562106; DOI=10.1038/nature02046;
RA   Ghaemmaghami S., Huh W.-K., Bower K., Howson R.W., Belle A.,
RA   Dephoure N., O'Shea E.K., Weissman J.S.;
RT   "Global analysis of protein expression in yeast.";
RL   Nature 425:737-741(2003).
CC   -!- FUNCTION: Required for axial budding pattern. Acts as an anchor to
CC       help direct new growth components and/or polarity establishment
CC       components to the cortical axial budding site.
CC   -!- SUBUNIT: Interacts with BUD5.
[..]
CC   ---------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC   Distributed under the Creative Commons Attribution-NoDerivs License
CC   ---------------------------------------------------------------------
DR   EMBL; U49845; AAA98666.1; -; Genomic_DNA.
DR   EMBL; Z38059; CAA86138.1; -; Genomic_DNA.
[..]
DR   GO; GO:0005935; C:bud neck; IDA.
DR   GO; GO:0005887; C:integral to plasma membrane; TAS.
DR   GO; GO:0005940; C:septin ring; TAS.
[..]
KW   Complete proteome; Glycoprotein; Membrane; Signal; Transmembrane.
FT   SIGNAL        1     22       Potential.
FT   CHAIN        23    823       Protein AXL2.
FT                              /FTId=PRO_0000020773.
FT   TOPO_DOM     23    508       Extracellular (Potential).
FT   TRANSMEM    509    529       Potential.
FT   TOPO_DOM    530    823       Cytoplasmic (Potential).
SQ   SEQUENCE   823 AA;  90783 MW;  350D79758BF30771 CRC64;
     MTQLQISLLL TATISLLHLV VATPYEAYPI GKQYPPVARV NESFTFQISN DTYKSSVDKT
     AQITYNCFDL PSWLSFDSSS RTFSGEPSSD LLSDANTTLY FNVILEGTDS ADSTSLNNTY
     QFVVTNRPSI SLSSDFNLLA LLKNYGYTNG KNALKLDPNE VFNVTFDRSM FTNEESIVSY
     YGRSQLYNAP LPNWLFFDSG ELKFTGTAPV INSAIAPETS YSFVIIATDI EGFSAVEVEF
     ELVIGAHQLT TSIQNSLIIN VTDTGNVSYD LPLNYVYLDD DPISSDKLGS INLLDAPDWV
     ALDNATISGS VPDELLGKNS NPANFSVSIY DTYGDVIYFN FEVVSTTDLF AISSLPNINA
     TRGEWFSYYF LPSQFTDYVN TNVSLEFTNS SQDHDWVKFQ SSNLTLAGEV PKNFDKLSLG
     LKANQGSQSQ ELYFNIIGMD SKITHSNHSA NATSTRSSHH STSTSSYTSS TYTAKISSTS
     AAATSSAPAA LPAANKTSSH NKKAVAIACG VAIPLGVILV ALICFLIFWR RRRENPDDEN
     LPHAISGPDL NNPANKPNQE NATPLNNPFD DDASSYDDTS IARRLAALNT LKLDNHSATE
     SDISSVDEKR DSLSGMNTYN DQFQSQSKEE LLAKPPVQPP ESPFFDPQNR SSSVYMDSEP
     AVNKSWRYTG NLSPVSDIVR DSYGSQKTVD TEKLFDLEAP EKEKRTSRDV TMSSLDPWNS
     NISPSPVRKS VTPSPYNVTK HRNRHLQNIQ DSQSGKNGIT PTTMSTSSSD DFVPVKDGEN
     FCWVHSMEPD RRPSKKRLVD FSNKSNVNVG QVKDIHGRIP EML
//
```

Fig. 3.4. Database record of Swissprot. The entry is shortened at some points, indicated by [...]



UniProtKB/Swiss-Prot entry
P38928

Printer-friendly view
Submit update
Quick BlastP search
Entry history

[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

Entry information

Entry name	AXL2_YEAST
Primary accession number	P38928
Secondary accession number	Q96VY8
Integrated into Swiss-Prot on	February 1, 1995
Sequence was last modified on	February 1, 1995 (Sequence version 1)
Annotations were last modified on	July 25, 2006 (Entry version 51)

Name and origin of the protein

Protein name	Protein AXL2 [Precursor]
Synonym	Suppressor of RHO3 protein 4
Gene name	Name: AXL2 Synonyms: BUD10, SRO4 OrderedLocusNames: YIL140W
From	Saccharomyces cerevisiae (Baker's yeast) [TaxID: 4932]
Taxonomy	Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.

References

[1] NUCLEOTIDE SEQUENCE [GENOMIC DNA], FUNCTION, AND SUBCELLULAR LOCATION.
PubMed=8846915 [NCBI, ExPASy, EBI, Israel, Japan]
Roemer T., Madden K., Chang J., Snyder M.;
"Selection of axial growth sites in yeast requires Axl2p, a novel plasma membrane glycoprotein.";
Genes Dev. 10:777-793(1996).

Comments

- ***FUNCTION***: Required for axial budding pattern. Acts as an anchor to help direct new growth components and/or polarity establishment components to the cortical axial budding site.
- ***SUBUNIT***: Interacts with BUD5.
- ***SUBCELLULAR LOCATION***: Cell membrane; single-pass type I membrane protein. In small, buds localizes to incipient bud sites, emerging buds and to the bud periphery. In large buds, localizes as a ring at the bud neck.
- ***PTM***: O-glycosylated but not N-glycosylated.
- ***MISCELLANEOUS***: Present with 396 molecules/cell.
- ***CAUTION***: Ref.4 refers to this gene as REV7. REV7 is however the adjacent gene.

Copyright

Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms. Distributed under the Creative Commons Attribution-NoDerivs License.

Cross-references

Sequence databases	
EMBL	U49845; AAA98666.1; -; Genomic_DNA. [EMBL / GenBank / DDBJ] [CoDingSequence] Z38059; CAA86138.1; -; Genomic_DNA. [EMBL / GenBank / DDBJ] [CoDingSequence] AF395906; AAK83884.1; -; Genomic_DNA. [EMBL / GenBank / DDBJ] [CoDingSequence] U07228; AAA67919.1; -; Genomic_DNA. [EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	S48394; S48394.
3D structure databases	
ModBase	P38928.
Protein-protein interaction databases	
IntAct	P38928; -.
DIP	P38928.
Protein family/group databases	
GermOnline	139675; -.
Enzyme and pathway databases	
BioCyc	SCER-S28-01:SCER-S28-01-002999-MONOMER; -.
2D gel databases	
SWISS-2DPAGE	Get region on 2D PAGE.
Organism-specific gene databases	
SGD	S000001402; AXL2.
Yeast-GFP	YIL140W.

Keywords

Complete proteome; Glycoprotein; Membrane; Signal; Transmembrane.

Features

Feature table viewer — Feature aligner

Key	From	To	Length	Description	FTId
SIGNAL	1	22	22	*Potential.*	
CHAIN	23	823	801	Protein AXL2.	PRO_0000020773
TOPO_DOM	23	508	486	Extracellular *(Potential)*.	
TRANSMEM	509	529	21	*Potential.*	
TOPO_DOM	530	823	294	Cytoplasmic *(Potential)*.	

Sequence information

Length: **823 AA** [This is the length of the unprocessed precursor] Molecular weight: **90783 Da** [This is the MW of the unprocessed precursor] CRC64: **350D79758BF30771** [This is a checksum on the sequence]

```
        10         20         30         40         50         60
MTQLQISLLL TATISLLHLV VATPYEAYPI GKQYPPVARV NESFTFQISN DTYKSSVDKT

        70         80         90        100        110        120
AQITYNCPDL PSWLSFDSSS RTFSGEPSSD LLSDANTTLY FNVILEGTDS ADSTSLNNTY

       130        140        150        160        170        180
QFVVTNRPSI SLSSDFNLLA LLKNYGYTNG KNALKLDPNE VFNVTFDRSM FTNEESIVSY
```

Fig. 3.5. Graphically edited Swissprot database record (NiceProt view). (Printed with permission of the Swiss Institute for Bioinformatics)

Because SIB specialists can not keep pace with the growing number of new entries, a supplement to Swissprot has been developed, the TrEMBL database. TrEMBL stands for *Translated EMBL* and contains all nucleic acid to protein translations of the EMBL database that have not yet been included in Swissprot. All entries are annotated automatically, and so their quality is less than those curated.

Both databases can be accessed via the Swissprot main page [swissprot]. Simple queries can be entered into the text field in the upper border of the main page. An SRS, similar to the EBI-SRS in operation, is available for more complex queries (see Sect. 3.2.1 – The Sequence Retrieval System). With both search tools, it is also possible to query other databases located at SIB.

NCBI Protein Database

Another well-known protein sequence database is maintained at the NCBI. This database, however, is not a single database but a compilation of entries found in other protein sequence databases. For example, the NCBI database contains entries from Swissprot, the PIR database [pir], the PDB database [pdb], protein translations of the GenBank database, as well as from a number of other sequence databases. Its format corresponds to that of GenBank and queries are carried out analogously to those of GenBank via the Entrez system of NCBI.

UniProt

The information available for proteins continues to grow rapidly. Besides sequence information, expression profiles can be examined, secondary structures predicted, and biological/biochemical function(s) analyzed. All these data are stored in databases, some of which are quite specialized. Therefore, it can be time consuming to collect all the relevant information regarding any given protein. EBI, SIB, and Georgetown University, therefore, have built a consortium with the aim of developing a central catalog for protein information. The

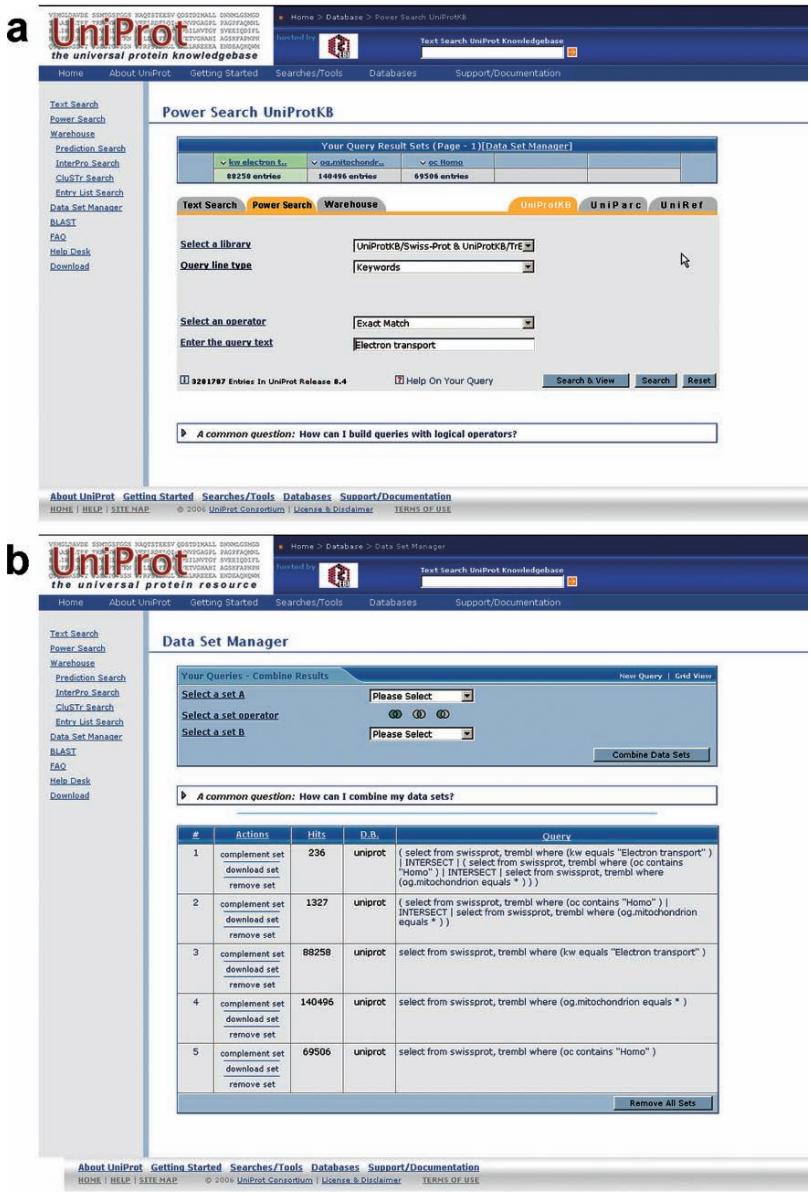


Fig. 3.6. Graphical power search (**a**) and data set manager (**b**) of the Uniprot database. (Printed with permission of the European Bioinformatics Institute)

result is the *Universal Protein Resource* (UniProt) [uniprot] (The UnitProt Consortium 2007), which unites the information in the three protein databases, Swissprot, TrEMBL, and PIR. UniProt consists of three parts, the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters Database (UniRef), and the UniProt Archive (UniPArc), a collection of protein sequences and their history.

UniProtKB is a comprehensive directory of protein annotations and is based on the Swissprot and TrEMBL databases. Querying the database can be done by either a full text search or complex queries (Fig. 3.6) using logical combinations of several search terms. Unlike the systems discussed so far, however, it is not necessary to connect the search terms by means of logical operators to build long search strings. Instead, single queries can be carried out with a single search term and then combined with the data set manager. Furthermore, help pages can be accessed at any time via context-sensitive hyperlinks.

UniRef is a nonredundant sequence database that allows for fast similarity searches. The database exists in three versions: UniRef100, UniRef90, and UniRef50. Each database allows the searching of sequences that are 100%, ≥ 90%, or ≥ 50% identical. The size of the database changes accordingly, thus making similarity searches much faster, for example, with BLAST.

3.3 Secondary Databases

3.3.1 PROSITE

An important secondary biological database is Prosite [prosite] (Falquet et al. 2002) resident at the SIB [expasy]. Classification of proteins in Prosite is determined using single conserved motifs, i.e., short sequence regions (10–20 amino acids) that are conserved in related proteins and usually have a key role in the protein's function. The search for such sequence motifs in

unknown proteins can provide a first hint of an affiliation to a protein family or function.

A motif is derived from multiple alignments (see Chap. 4) and saved in the database as a regular expression (Fig. 3.7). This is a formalized pattern for the description of a sequence

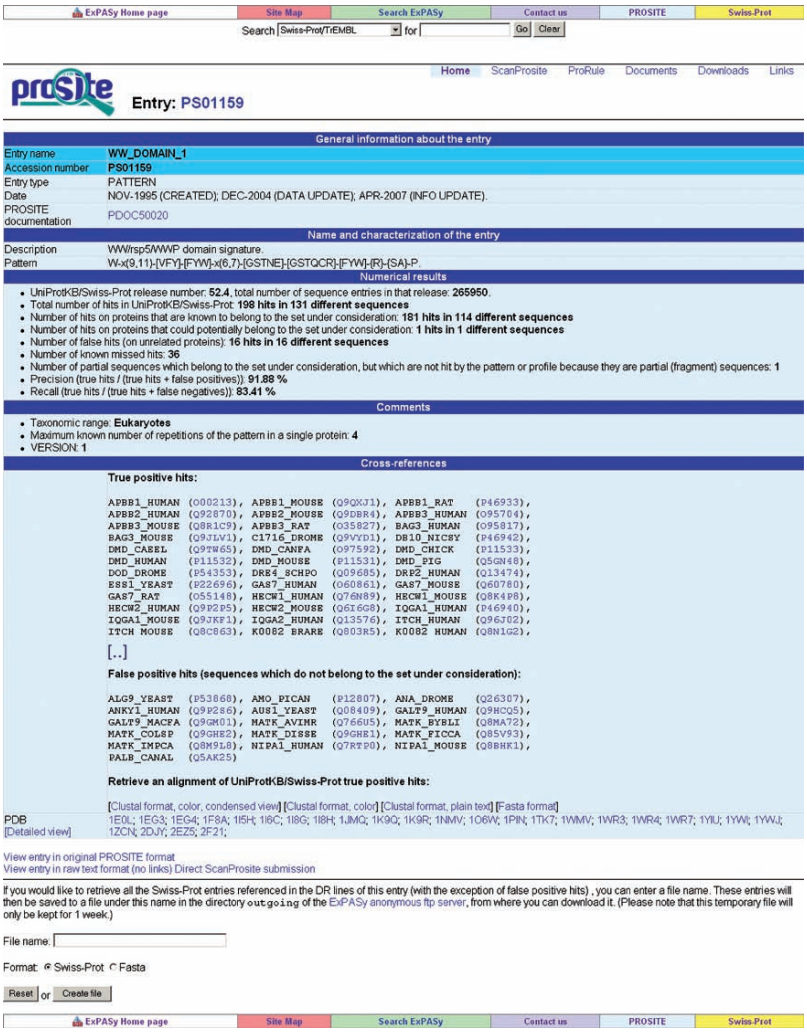


Fig. 3.7. *NiceSite view* of the Prosite database record PS01159. (Printed with permission of the Swiss Institute for Bioinformatics)

of characters. In a regular expression in Prosite, the individual amino acids are represented in the one-letter-code and separated by hyphens. If a position can contain more than one residue, then these are written between square brackets. Positions that can be filled by any amino acid are marked with the lower case letter *x*. Repetitions of the same amino acid are indicated within full brackets, followed by the number of repetitions. A typical regular expression in Prosite would have the following form: `[GSTNE]-[GSTQCR]-[FYW]-{ANW}-x(2)-P`. This regular expression has seven amino acid positions. The first amino acid can be glycine, serine, threonine, asparagine, or glutamate. The second position glycine, serine, threonine, glutamine, cysteine, or arginine, and the third position phenylalanine, tyrosine, or tryptophan. Position four can be any amino acid except alanine, asparagine, and tryptophan. In position five and six, any amino acid can follow, and position seven is occupied by proline. The Prosite user manual [prosite-manual] contains a complete description of the Prosite database as well as the syntax of the regular Prosite expressions. The Expasy Prosite WWW server [prosite] offers different possibilities to query the Prosite database. Besides searching for keywords, one can examine a sequence for the presence of Prosite motifs. Furthermore, using the algorithm ScanProsite, Prosite offers the possibility to search Swissprot, TrEMBL, and PDB for protein sequences that contain a user-defined pattern.

3.3.2 PRINTS

The Prints database [prints] (Attwood et al. 2003) uses fingerprints to classify sequences. Fingerprints consist of several sequence motifs, represented in the Prints database by short local ungapped alignments (see Chap. 4). The Prints database takes advantage of the fact that proteins usually contain functional regions that result in several sequence motifs per protein. By using fingerprints the sensitivity of the analysis increases, i.e., it is possible to evaluate the affiliation of a protein to a protein

family even in the absence of one of the surveyed motifs. Besides information on how to derive a fingerprint and judge its quality, Prints database also offers cross-references to entries in related databases, thus permitting access to more information regarding the protein family. Like Prosite, Prints database contains information about each protein family and, if available, the biological function of each motif in the fingerprint. Querying the database on the Prints-WWW server [prints] can be carried out via a keyword search. However, it can be more interesting to search for fingerprints in protein sequences. Like the Prosite WWW server, the Prints server offers tools for sequence analysis.

3.3.3 Pfam

The Pfam database [pfam] (Bateman et al. 2002) classifies protein families according to profiles. A profile is a pattern that evaluates the probability of the appearance of a given amino acid, an insertion or a deletion at every position in a protein sequence. Conserved positions are weighted more strongly than less conserved positions, i.e., a weighted scoring scheme. Pfam is based on sequence alignments. High-quality, manually checked alignments serve as starting points for the automatic construction of Hidden Markov Models (HMM). Further sequences are then automatically added to the individual alignments of the Swissprot database. The resulting alignments should represent functionally interesting structures and contain evolutionarily related sequences. Because of the partly automatic construction of the alignments, however, it is also possible that sequence alignments arise that have no evolutionary relationship to one other. Therefore, results of a search against the Pfam database should be carefully reviewed.

3.3.4 Interpro

The Integrated Resource of Protein Families, Domains, and Sites (Interpro) [interpro] (Mulder et al. 2007) integrates important

secondary databases into a comprehensive signature database. Interpro merges the databases Swissprot, TrEMBL, Prosite, Pfam, Prints, ProDom, Smart, and TIGRFAMs [tigr] and thereby allows a simple and simultaneous query of these databases. The result page combines the output of the individual queries. This makes for a fast comparison of the results while taking into account the strengths and weaknesses of the individual databases. The Interpro WWW server offers a number of intuitive query facilities for text and sequence searches.

3.4 Genotype–Phenotype Databases

For diseases to emerge and progress, several genes or their products are frequently required. The identification of genes relevant to disease is, therefore, of vital importance in a target-based approach for rational drug development. A number of genotype–phenotype databases have been established that record relationships between genes and the biological properties of organisms. The Online Mendelian Inheritance in Man (OMIM) database of the NCBI [omim] is perhaps the best-known genotype–phenotype database. A new database of this type, dbGaP [dbgap], has also been recently established at the NCBI. The data in this database are provided with analyses of the statistical significance of the respective genotype–phenotype relationship. The Online Mendelian Inheritance in Animals (OMIA) database [omia] at the NCBI also contains genotype–phenotype relationships of various animals, except the mouse and the humans. For mice, the relevant database is in the Mouse Genome Database (MGD) [mgd]. Genotype–phenotype relationships of the two important model organisms, *D. melanogaster* and *C. elegans*, are recorded in FlyBase [flybase] and WormBase [wormbase], respectively. Both databases also contain much more information than just genotype–phenotype data. A detailed description of all the above databases [nar] would be beyond the scope of this book. Below, therefore, only a genotype–phenotype database is discussed that semantically integrates the contents of the above mentioned databases.

3.4.1 PhenomicDB

The PhenomicDB database is a multiorganism genotype–phenotype database containing data from man and other important organisms such as the mouse, zebra fish (*Danio rerio*), fruit fly (*D. melanogaster*), nematode (*C. elegans*), baker's yeast (*S. cerevisiae*), and cress plant (*Arabidopsis thaliana*). PhenomicDB integrates data from the above and other primary genotype–phenotype databases. A complete listing of all underlying data sources can be found on the homepage [phenomicdb] and in Kahraman et al. (2005).

A characteristic of PhenomicDB is that cross-organism comparisons of genotype–phenotype relationships are possible. This is accomplished by incorporating orthology data and gene indices from the database, HomoloGene [homologene] at the NCBI. For example, the cause of porphyria, an inherited or acquired enzyme defect of humans, is a nonfunctional δ-aminolevulinate dehydratase. The respective gene has the symbol ALAD. As PhenomicDB indicates, a defect in the orthologous gene of baker's yeast (gene symbol: HEM2) leads to a very similar phenotype, characterized by the keywords auxotrophies, carbon and nitrogen utilization defects, carbon utilization, and respiratory deficiency. Of course, one cannot expect that distantly related organisms such as baker's yeast and humans show identical genotype–phenotype relationships in every case. Nevertheless, similar relationships can occur that might generate new hypotheses regarding disease pathogenesis or that allow the advancement of a disease model, thereby supporting the development of new drugs.

PhenomicDB is queried via a simple search interface. Search terms can be complemented automatically or manually with wildcards and restricted to certain database fields. Furthermore, it is possible to restrict the search to selected organisms. Provided that orthologs of a given gene are found, the result page offers a hyperlink to the corresponding database record, allowing for a fast comparison of the genotype–phenotype relationships across organisms (Fig. 3.8). Because of the semantic integration of the primary databases, some detail information can be lost, however, but this is compensated for by the interconnections of

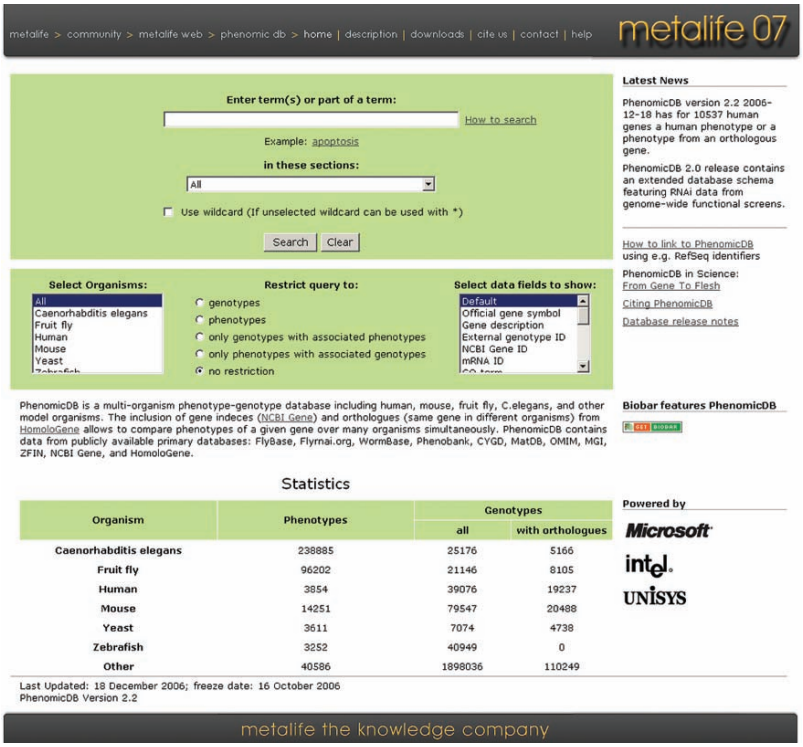


Fig. 3.8. Start page of PhenomicDB. (Printed with permission of Metalife AG)

the primary data and the breadth of information included. PhenomicDB can therefore be regarded as a metasearch engine for phenotypic information.

3.5 Molecular Structure Databases

3.5.1 Protein Data Bank (PDB)

The Protein Data Bank (PDB) is a database of experimentally determined crystal structures of biological macromolecules and is coordinated by the consortium located in the USA, Europe,

and Japan [wwpdb] (Berman et al. 2000). Probably, the best known web page of the PDB is that of the Research Collaboratory for Structural Bioinformatics [pdb]. The PDB was founded at the Brookhaven National Laboratory in 1971, reflected in the frequent use of the name Brookhaven Protein Data Bank.

About 46,000 macromolecule structures are stored in the PDB database (as of September 2007). These are predominantly proteins, but also include DNA and RNA structures and protein–nucleic acid complexes. Structures of other macromolecules, e.g., glycopeptides or polysaccharides, constitute only a very small proportion. As of 2002, only those crystal structures that have been solved experimentally are stored in the PDB database, whereas data of theoretical protein models are kept in their own section [pdb-models].

The PDB database offers a number of query options. A text-based search for a PDB-ID or a keyword can be initiated on the main page. Furthermore, a number of search options exist on the search database page, including detailed keyword and BLAST queries. A database record summarizes all of the information in the file which is then detailed on the following pages. In addition, the molecular structure can be visualized by means of different applets (Fig. 3.9).

3.5.2 SCOP

Proteins that perform a similar biological function and are evolutionary related must have a similar structural organization, at least in the region of their active centers. It should, therefore, be possible to predict the function of an unknown protein by comparison of its structural organization with that of known proteins. Two databases, SCOP and CATH, provide such predictions. SCOP (Structural Classification Of Proteins) [scop] (Murzin et al. 1995) classifies proteins of a known structure in a hierarchical manner. The three main classifications are families, super families, and folds. Families describe proteins with a clear evolutionary relationship to each other and are limited by a

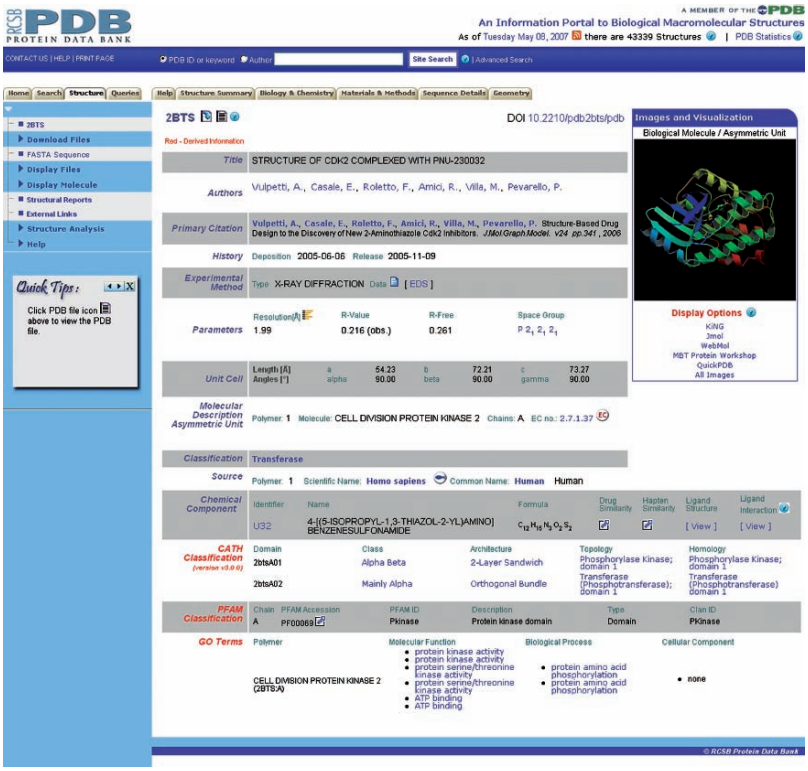


Fig. 3.9. Overview representation of the PDB entry 2BTS. (Printed with permission of the RCSB)

sequence identity that must be at least 30% over the total length of the proteins. Nevertheless, proteins that fall below this limit can be included in a family if relatedness can be shown due to proven similar structures and functions. Proteins with a very low sequence identity to one other, even with suggested relations due to structural and functional properties, are assigned into super families, however. Proteins that have the same arrangement of secondary structure elements in the same topology are classified into folds. It is unimportant if the proteins have a functional relationship or if the similarity of the fold is based on physico-chemical principles.

3.5.3 CATH

The CATH database [cath] (Greene et al. 2007) classifies protein structures hierarchically into four categories: Class (C), Architecture (A), Topology (T), and Homologous Superfamily (H). The classification of proteins into the Class category is mainly automatic, but can be complemented manually when required. In the Class category, the proportion of secondary structure elements is taken into account without consideration of their arrangement or connections. Four classes of proteins are distinguished: proteins that are composed mainly of helices (*mainly-alpha*); sheets (*mainly-beta*), or both helices and sheets (*alpha–beta*); and finally, proteins with very few secondary structure elements. The Architecture category describes the arrangement of secondary structure elements to one other and is curated manually. Its categorization is performed via simple descriptors such as *barrel, sandwich, beta-propeller,* etc. In the Topology category, protein form and the interconnections of secondary structure elements are described. Its categorization is based on an algorithm that uses empirically derived parameters for domain classification. The Homologous Superfamily category encompasses homologous protein domains, i.e., domains with a common origin. The similarity of the sequences is determined by sequence comparison followed by a structure comparison according to the classification in the Topology category. In addition to these four categories (whose first letters form the database name), a fifth category has been defined, the Sequence Families. Here, domains are classified based on high sequence identity (at least 35% identity over 60% of the length of the larger domain) and, thus, will likely possess similar functions.

3.5.4 PubChem

The PubChem database at NCBI [pubchem] stores small chemical molecules and information about their biological activities.

It consists of three components: PubChem Compound, PubChem Substance, and PubChem BioAssay. PubChem Compound contains approximately 11 million molecules (September 2007) together with their two-dimensional (2D) molecular structures. A query is performed graphically via a molecular structure editor that allows the drawing of the desired (partial) structure (Fig. 3.10). Furthermore, PubChem Compounds make possible a search for molecules that fulfill certain physico-chemical parameters, e.g., a particular molecular weight range, a given number of acceptors or donors for hydrogen bonds, a certain log *P* range, etc.

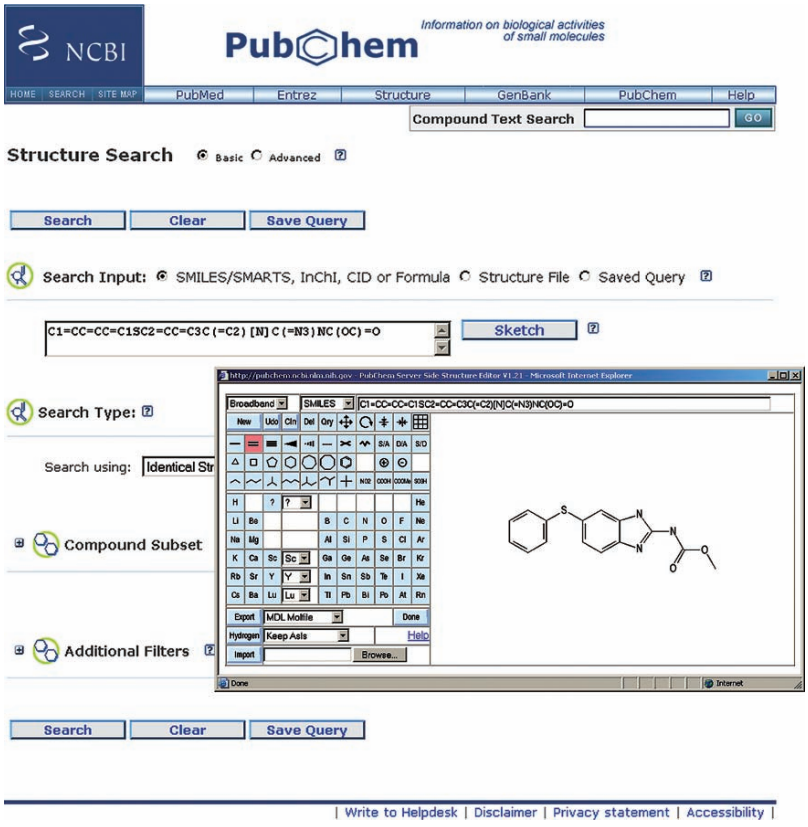


Fig. 3.10. 2D-molecular structure editor of the PubChem database. (Printed with permission of the NCBI)

PubChem Substance permits the search for substances produced by various manufacturers, compounds of unknown composition, and natural substances of unknown 2D molecular structure. The records of both databases are linked and include a link to the third database, PubChem BioAssay, provided that respective data are present. Information on biological assays and molecules that have been tested in these systems is recorded in PubChem BioAssay and this database can be queried via a text search in the Entrez system.

The PubChem databases have multiple applications due to internal and external database linking, including PubMed. For example, with a known enzyme inhibitor it is possible to find other similar potential inhibitors. Furthermore, small chemical molecules can be identified that have different structures yet shown to have similar effects in a biological test system.

3.6 Exercises

1. Search for a protein (enzyme) from the organism *Bacillus subtilis* that hydrolyzes terminal nonreducing arabinofuranoside residues. To do this, use the keyword search under Entrez (http://www.ncbi.nlm.nih.gov/entrez/).
   Note: hydrolysis, arabinofuranoside, hydrolases, glycosyl, terminal, nonreducing. At the Entrez page you can combine previous search results with new search terms using the hyperlink History. Possible combinations are AND, OR, NOT.
2. Locate the gene for the enzyme ABF2_BACSU from exercise 3.1 in the nucleotide database. If you are unable to find it, try to develop new search strategies from the results and hints provided.
3. Search for the protein with the following accession number in Entrez: P94552.
4. Search for the same accession number on the EBI home page (http://www.ebi.ac.uk/).

5. In the EBI-SRS system, look for the identifier ABF2_BACSU. To do this, select the databases UniprotKB and UniprotKB/Swissprot at the database selection page. Which entries can be found and how do they differ?
6. Have a closer look at the entry from exercise 3.5 and change to the TextEntry view. Which information can you obtain from such an entry? Describe briefly the information found. It is not necessary to characterize ABF2_BACSU any further.
7. In the SwissEntry view, under References, you will find a hyperlink to a publication in the journal Microbiology. Click on this hyperlink. What happens? Note: The hyperlink to the publication is also available in the TextEntry view.
8. In the literature, two genes for *arfI* and *arfII* are described that are homologous to α–L-arabinofuranosidase 1 and α–L-arabinofuranosidase 2. From which species are these two genes? Which other species are reported in literature to have homologous genes that are highly identical? To answer these questions go again to the NCBI page (http://www.ncbi.nlm.nih.gov/) and search in the PubMed database. The History function that was mentioned earlier (3.1) can also be used in PubMed and in all other database searches at NCBI.
9. In the PubMed database look for a publication of an author with your own last name. How many publications can you find? Are there several authors with your name? If you find nothing with your name, try it with the name *Blobel*. How can you restrict the results further using the name Günther Blobel, for example? How do you explain the differences with different search strategies?
10. Carry out a Prosite Scan (http://www.expasy.org/prosite) with the sequence of the database entry ABF2_BACSU. You can enter the sequence by cut & paste, or by imputing the Swissprot accession

number or ID. How many patterns are found and which ones? Which information about the motifs do you get on the result page? How can you obtain information about the biological role of the individual motifs in a simple way?

11. Go to the start page of the Prints database (http://bioinf.man.ac.uk/dbbrowser/PRINTS/) and perform a Fingerprint search against the Print database with the sequence ABF2_BACSU. Please note that the sequence must be entered in Raw format. When done, perform the same search with the sequence of the database record A1AB_HUMAN.
12. Go to the Blocks WWW server (http://blocks.fhcrc.org/) and initiate a database search with the Blocks Searcher using the sequence P35368. Because the search can take several minutes, possibly leading to a browser timeout, you should enter your e-mail address on the form. The results of the analysis will then be sent to you by e-mail. How many hits are found?
13. With the protein from exercise 3.12, query the Pfam database. The proteins of the Swissprot and TrEMBL databases are already present on the Pfam WWW server. You can, therefore, either retrieve the previously calculated result with the accession number or protein ID, or run a new analysis by providing the sequence in the FASTA format.
14. Repeat the search from above (3.13) with the Interpro database.
15. Retrieve the 3D structure of bovine rhodopsin (a GPCR) from the PDB database. How many entries can you find? Have a closer look at the entry with the best crystallographic resolution for the complete protein (detailed in the overview). At what temperature was the crystallization carried out and how many cysteine bonds does the protein have?

16. Is there an assay to check for the HERG channel activity of a molecule? How many compounds were tested in this assay and how many of them were active? Use the PubChem database to answer this question.
17. In how many assays was the molecule Fenbendazole tested and in how many of these was it active? What is Fenbendazole used for and how does the molecule differ from Albendazole?
18. Is there a genotype–phenotype relationship in *D. melanogaster* that resembles the humane genotype–phenotype relationship responsible for coproporphyria? Use PhenomicDB to answer this question.

WWW Links

bankit: http://www.ncbi.nlm.nih.gov/BankIt/
cath: http://www.cathdb.info/
cib: http://www.cib.nig.ac.jp/
dbgap: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gap
ddbj: http://www.ddbj.nig.ac.jp/
ddbj-srs: http://srs.ddbj.nig.ac.jp/index-e.html
ebi: http://www.ebi.ac.uk/
ebi-manual: http://www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html
ebi-srs: http://srs.ebi.ac.uk/
ebi-srs-help: http://srs.ebi.ac.uk/srs/doc/index.html
embl: http://www.ebi.ac.uk/embl/
entrez: http://www.ncbi.nlm.nih.gov/entrez/
entrez-help: http://www.ncbi.nlm.nih.gov:80/entrez/query/static/help/helpdoc.html
expasy: http://www.expasy.org/
flybase: http://www.flybase.org/
gb-sample: http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html
genbank: http://www.ncbi.nlm.nih.gov/Genbank/
homologene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene
interpro: http://www.ebi.ac.uk/interpro/
mgd: http://www.informatics.jax.org/
nar: http://nar.oxfordjournals.org/

ncbi: http://www.ncbi.nlm.nih.gov/
omia: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=omia
omim: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi? db= OMIM
pdb: http://www.pdb.org
pdb-models: http://www.pdb.org/pdb/search/searchModels.do
pfam: http://www.sanger.ac.uk/Software/Pfam/
phenomicdb: http://www.phenomicdb.de/
pir: http://pir.georgetown.edu/pirwww/dbinfo/pir_psd.shtml
prints: http://bioinf.man.ac.uk/dbbrowser/PRINTS/
prosite: http://www.expasy.org/prosite/
prosite-manual: http://www.expasy.org/prosite/prosuser.html
pubchem: http://pubchem.ncbi.nlm.nih.gov/
scop: http://scop.mrc-lmb.cam.ac.uk/scop/
sequin: http://www.ncbi.nlm.nih.gov/Sequin/
swissprot: http://www.expasy.org/sprot/
tigr: http://www.tigr.org/
uniprot: http://www.uniprot.org/
wormbase: http://www.wormbase.org/
wwpdb: http://www.wwpdb.org/

Literature

Attwood TK, Bradley P, Flower DR, Gaulton A et al (2003) PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 31:400–402

Bateman A, Birney E, Cerruti L, Durbin R et al (2002) The Pfam Protein Families Database. Nucleic Acids Res 30:276–280

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28: 235–242

Etzold T, Ulyanov A, Argos P (1996) SRS: information retrieval system for molecular biology data banks. Methods Enzymol 266:114–128

Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJA, Hofmann K, Bairoch A (2002) The PROSITE database, its status in 2002. Nucleic Acids Res 30:235–238

Greene LH, Lewis TE, Addou S, Cuff A et al (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35: D291–D297

Kahraman A, Avramov A, Nashev L, Popov D et al (2005) PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics. Bioinformatics 21:418–420

Mulder NJ, Apweiler R, Attwood TK, Bairoch A et al (2007) New developments in the InterPro database. Nucleic Acids Res 35:D224–D228

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540

The UniProt Consortium (2007) The universal protein resource (UniProt). Nucleic Acid Res 35:D193–D197

Wheeler DL, Barrett T, Benson DA, Bryant SH et al (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35:D5–D12

4 Sequence Comparisons and Sequence-Based Database Searches

4.1 Pairwise and Multiple Sequence Comparisons

The comparison of protein and DNA sequences is an important analytical method of applied bioinformatics. The annotations of new nucleotide and protein sequences, construction of model structures for proteins, design and analysis of expression studies as well as a variety of other bioinformatic and biological experiments are all based on these analyses. Nature acts conservatively, i.e., it does not develop a new kind of biology for every life form but continuously changes and adapts a proven general concept. Novel functionalities do not appear because a new gene has suddenly arisen but are developed and modified during evolution. Given this situation, therefore, one may transfer functional information from one protein to another if both possess a certain degree of similarity. However, this process must be carried out critically, as similar proteins may yet perform different functions, despite, for example, having arisen from a common ancestor.

Before analyzing whether sequences are possibly related, it is first necessary to define some terms. Related sequences are designated as being homologous; however, the term homology often leads to confusion. Homology is not a measure of similarity, but rather that sequences have a shared evolutionary history and, therefore, possess a common ancestral sequence (Tatusov et al. 1997). The definition of the terms ortholog and paralog in combination with the function of a protein is, however, controversially discussed (Jensen 2001, Gerlt and Babbitt 2001).

In general, genome biologists define these terms as follows: Homologous proteins from different species that possess the same function (e.g., corresponding kinases in a signal transduction pathway in humans and mice) are called orthologs. In contrast, homologous proteins that have different functions in the same species (e.g., two kinases in different signal transduction pathways of humans) are termed paralogs.

Homology is not quantifiable – either two sequences are homologous or not. The identity or similarity of two sequences is, however, quantifiable. Identity is the ratio of the number of identical amino acids or nucleotides in a sequence relative to the total number of amino acids or nucleotides. Unlike identity, similarity is not as simple to calculate. Before similarity can be determined, it must first be defined how similar the building blocks of sequences are to each other. This is done with the help of similarity matrices that are also known as substitution or scoring matrices. Similarity matrices specify the probability at which a sequence transforms into another sequence over time. Of course, this is dependent on the time elapsed and the mutational rate of nucleotides.

Before deciding upon the identity or similarity of two nucleotide or amino acid sequences, an alignment must first be calculated. The underlying principle of such an alignment is relatively simple (see Fig. 4.1). Two sequences are arbitrarily placed next to each other and the alignment judged according to the scoring matrices. The two sequences are then moved relatively to one other and for each position a score is calculated. This process is repeated until the best alignment is found.

For nucleotide sequences the simplest solution is an identity matrix (see Fig. 4.2a). Here, one assumes that the four nucleotides do not show any similarity to one other, and therefore, only identical nucleotides are factored into the similarity scoring. For protein sequences, an identity matrix is not sufficient to describe biological and evolutionary processes. Amino acids are not exchanged with the same probability as might be conceived theoretically. For example, an exchange of aspartic acid for glutamic acid is frequently observed; however,

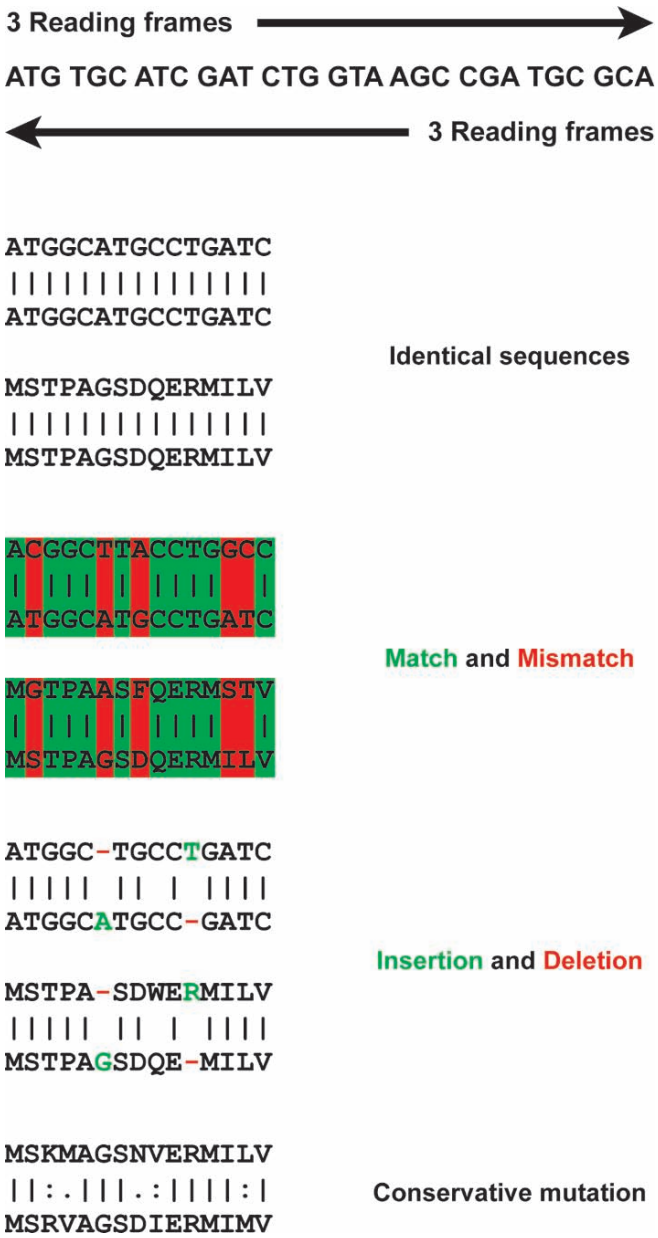
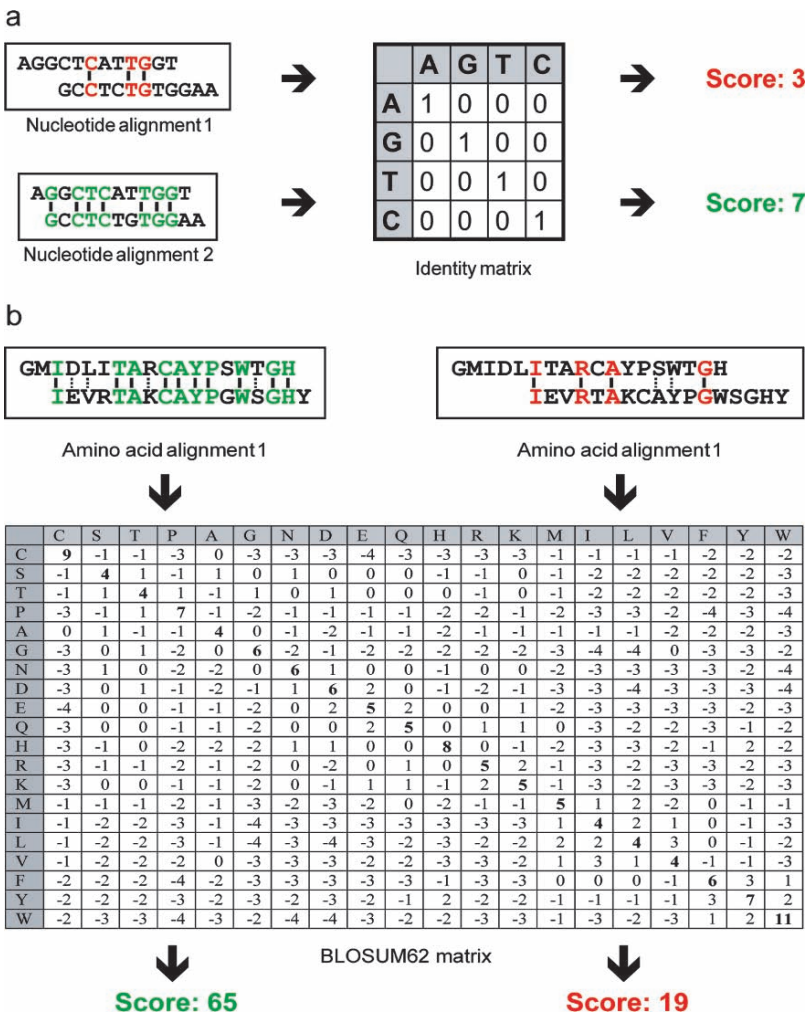


Fig. 4.1. Sequence alignments of nucleotide and amino acid sequences



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	**9**	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	**4**	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	**4**	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	**7**	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	**4**	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	**6**	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	**6**	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	**6**	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	**5**	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	**5**	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	**8**	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	**5**	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	**7**	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	**11**

Fig. 4.2. Scoring matrices allow the computation of optimal alignments. (**a**) Use of an identity matrix for the construction of an optimal nucleotide alignment. (**b**) Use of the BLOSUM62 matrix for the construction of an optimal amino acid alignment. Two potential alignments for each are represented whereby the optimal alignment is shown in *green*

a change from aspartic acid to tryptophan is seen rarely. One reason for this is the triplet-based genetic code (see Chap. 2). For an exchange of aspartic acid to glutamic acid to occur only a mutation of the last nucleotide in the triplet codon is required

(GAT/GAC to GA**A**/GA**G**). In contrast, a complete mutation of the whole triplet has to occur in order to exchange aspartic acid for tryptophan (**GAT/GAC** to **TGG**). Of course such a complete mutational substitution has a much lower probability of occurring. A second reason for the mutation of aspartic acid-to-glutamic acid to occur more often is that both have similar properties (see Chap. 2). In contrast aspartic acid and tryptophan are chemically different – the hydrophobic tryptophan is frequently found in the center of proteins, whereas the hydrophilic aspartic acid occurs more often at the surface. An exchange of aspartic acid for tryptophan, therefore, could greatly alter the tertiary structure of a protein and consequently its function. Such striking amino acid exchanges accompanied by a loss of function rarely happen. Amino acid substitution matrices, therefore, describe the probability at which amino acids are exchanged in the course of evolution. The most commonly used amino acid scoring matrices are the PAM (*Position Accepted Mutation*; Dayhoff et al. 1978) and BLOSUM groups (*Blocks Substitution Matrix*; Henikoff and Henikoff 1992) (see Fig. 4.2b). These matrices contain the logarithm for the relationship of two probabilities at which a couple of amino acids or nucleotides appears in an alignment, i.e., both the probability of a coincidental concurrence and the probability of an evolutionary event responsible for the occurrence are taken into account. Negative values in the matrix mean that the occurrence is rather coincidental whereas positive values suggest an evolutionary event. Because the matrix values are logarithms of relationships, addition of the numbers derives a conclusion for the complete alignment.

Alignments can be carried out both globally and locally (see Fig. 4.3). In global alignments, complete nucleotide or protein sequences are compared to one another over the entire length of the sequence. In Fig. 4.4, the calculation of a global alignment is shown. However, even very similar sequences can have single deletions or insertions and consequently a different number of amino acids or nucleotides. To represent these alignments appropriately, gaps must be inserted into the

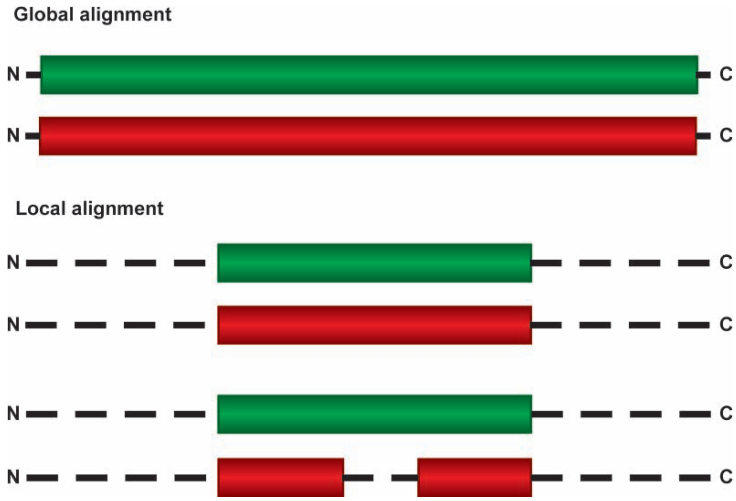


Fig. 4.3. Global and local sequence alignment. Gaps can appear also in a global alignment, as seen in the lower local alignment

sequences. Theoretically all possible sequences can be aligned by the introduction of gaps. To prevent this, scoring penalties are given for the introduction of gaps (gap opening) and their extension (gap extension). These are then subtracted from the alignment score to yield the total score. The alignment with the highest total score is considered the optimal sequence comparison. This method is based on the algorithm of Needleman and Wunsch (1970).

Sometimes, interest may focus solely on aligning the most similar stretches within two sequences – a local alignment. With this approach, protein domains and motifs (e.g., ATP binding sites, DNA binding domains, *N*-glycosylation sites) can be identified. In principle, a local alignment is calculated in the same way as a global alignment using a substitution matrix and the introduction and extension of gaps. However, the path through the matrix does not move from the lower right to the upper left, but starts and ends at arbitrary places (see Fig. 4.4). If a score

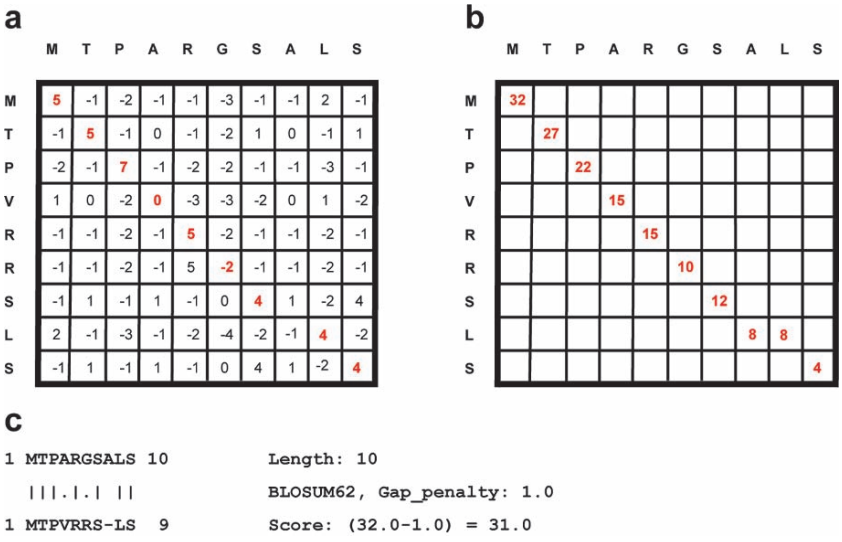


Fig. 4.4. Calculation of a global alignment of two similar protein sequences. (**a**) Both sequences are compared in a two-dimensional matrix and the similarity of the amino acids is determined using similarity matrices. Each alignment can be described as a path through the two-dimensional matrix, starting with highest-scoring amino acid pair at the N-terminus. (**b**) By adding the values corresponding scores for the different paths are obtained. The alignment with the highest score is considered optimal (shown in *red*). (**c**) The optimal alignment is obtained by the introduction of a gap and contains 10 amino acids, of which seven are identical. Using the BLOSUM62 similarity matrix and a gap penalty of 1.0 a score of 31.0 is achieved

can no longer be increased, the alignment stops and the value of 0 is given in the matrix. The local alignment identified in the matrix with the highest score is regarded as optimal. This method is based on the algorithm of Smith–Waterman (Smith and Waterman 1981).

To compare more than two nucleotide or protein sequences, one could compare all sequences pairwise and then further examine these alignments. However, it is quicker to perform a multiple alignment (Fig. 4.5). One well-known program is ClustalW (Thompson et al. 1994) that utilizes the fact that similar sequences are usually homologous. The program first generates a phylogeny that represents the relationships between the

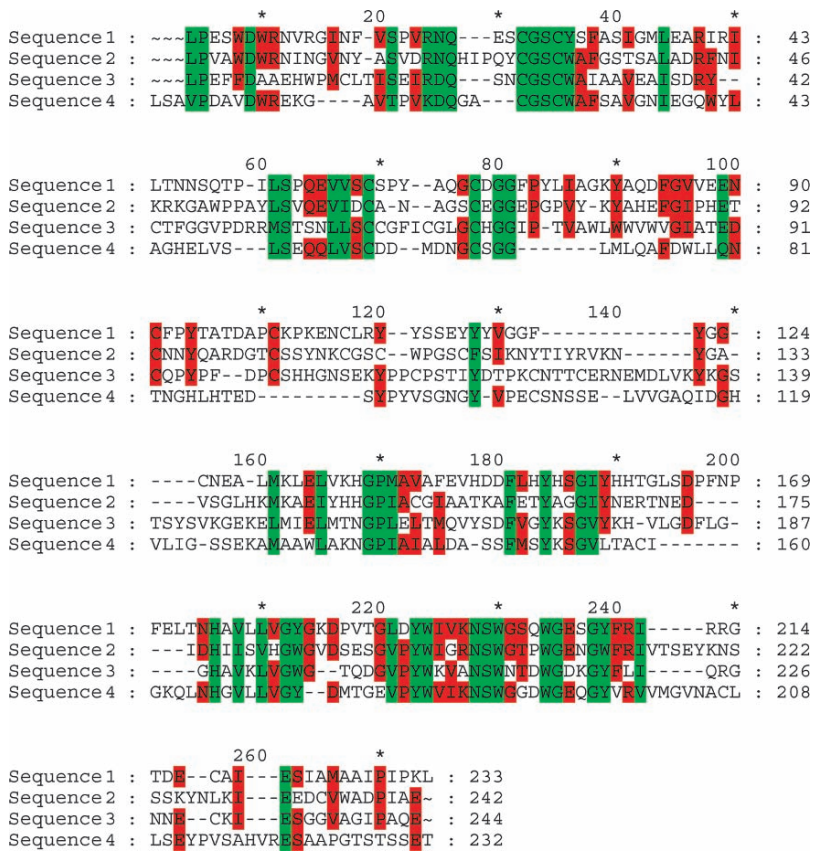


Fig. 4.5. Multiple sequence alignment of four related proteins. Amino acids conserved in all four sequences (or with conservative changes) are highlighted in *green*; those conserved only in three of four sequences are shaded *red*

sequences. Then pairwise alignments are carried out, beginning with the most similar sequences. Once all of the pairwise alignment scores relative to all the other sequences have been calculated, they are then used to classify the sequences into groups. Finally, the groups are presented as a multiple alignment. Because ClustalW calculates the evolutionary distance of the sequences, the program can also generate phylogenetic trees (see Fig. 4.6). These display sequence relationships as a graph in which the evolutionary distances correspond to the length of the horizontal branches.

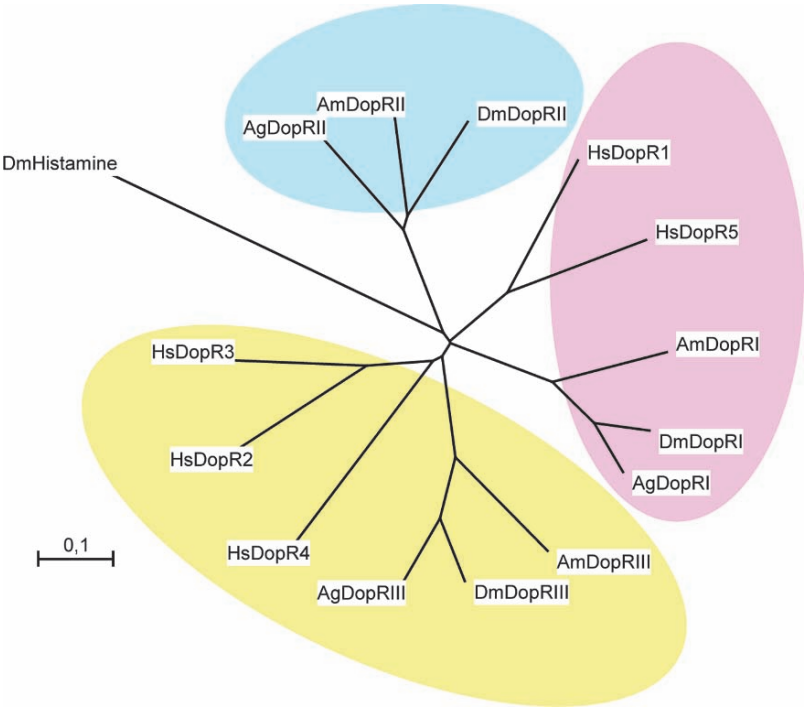


Fig. 4.6. Phylogenetic tree of dopamine receptor sequences. The evolutionary relationship between the sequences is reflected by the length of the branches. Dopamine receptor sequences of invertebrates (Dm, *Drosophila melanogaster*; Ag, *Anopheles gambiae*; Am, *Apis mellifera*) are compared with those of humans (Hs, *Homo sapiens*). Three clear clusters are formed. As a control, the phylogenetically distant sequence of the Dm histamine receptor was not found in any of the clusters

4.2 Database Searches with Nucleotide and Protein Sequences

A frequently used application of pairwise alignments is the search for similar protein or nucleotide sequences in sequence databases. With older dynamic alignment algorithms such as those designed by Smith and Watermann (1981) or Needleman and Wunsch (1970) this is too slow to perform on current computers. Instead, heuristic algorithms like BLAST (Basic Local

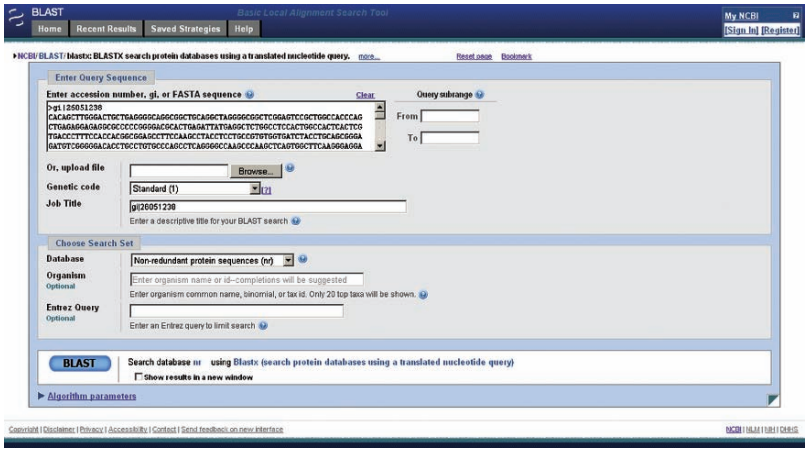


Fig. 4.7. Translated-BLAST start page at NCBI. The *blastx* algorithm was used to compare a nucleotide sequence with a protein sequence database (Printed with permission of the NCBI)

Alignment Search Tool; Altschul et al. 1990) are employed (Figs. 4.7 and 4.8). Heuristic methods make assessments to get almost exact results and utilize sequence and alignment statistics to make searches in large databases feasible. They do not guarantee an optimal alignment, however, but allow for sensitive and fast database searches.

BLAST comes in two versions: NCBI-BLAST and WU-BLAST. Both versions are based on developments from the NCBI [ncbi-blast]. WU-BLAST is an alternative modification of the NCBI-BLAST 1.4 by the Washington University [wublast] and performs slightly better when working with genomic sequences. NCBI-BLAST is more suitable for database searches and is used far more frequently.

To execute a meaningful search in a nucleotide or protein database, the corresponding algorithm must be chosen from the BLAST group and this depends on the aim of the search as well as the nature of the query sequence (nucleotide or protein; Table 4.1). For example, to query a nucleotide database with a protein sequence, every nucleotide sequence of the database must be translated into all six theoretically possible protein sequences (Fig. 4.1). Only then can the query sequence be compared with the

Fig. 4.8. Graphic representation of a BLAST result. The graph summarizes the number and length of hits with respect to the query sequence. The quality (alignment score) of the hits is represented by color-coding (Printed with permission of the NCBI)

database. This complex process is performed automatically by the algorithm *tblastn*. Depending on the nature of the query and the databases used, a total of five algorithms are possible (Table 4.1).

Within the BLAST family of algorithms, PSI-BLAST (Position Specific Iterated BLAST; Altschul et al. 1997), PHI-BLAST (Pattern Hit Initiated BLAST; Zhang et al. 1998), and bl2seq (blast two sequences; Tatusova and Madden 1999) are particularly interesting. The algorithm bl2seq carries out a local alignment of two sequences. PHI-BLAST allows to search for proteins in a protein database with sequence motifs similar to

Table 4.1. The most important algorithms of the BLAST group and their applications

Algorithm	Query sequence	Database	Remarks
blastp	Protein	Protein	–
blastn	Nucleotide	Nucleotide	–
blastx	Nucleotide	Protein	Query sequence is translated into all six reading frames
tblastn	Protein	Nucleotide	Database is translated into all six reading frames
tblastx	Nucleotide	Nucleotide	Query sequence and database are translated into all six reading frames

BLAST is usually performed first against a nonredundant database, which is a compilation of entries from different databases. In a nonredundant database, multiple entries are removed so that every record is available only once. Nonredundant databases exist for both nucleotide and protein sequences

those of the query. PSI-BLAST is a mixture of a pairwise and a multiple alignment. First, a normal BLAST search is executed. With the resulting multiple alignment of hits, a sequence profile is constructed, which is then used to continue the search for new sequences until no more are found. The interpretation of the results is frequently very difficult and occasionally misleading because sequences not directly related can also be taken into account. Therefore, PSI-BLAST results require careful examination. Hidden Markov Models (HMMs; Eddy 2004) operate in a similar fashion more slowly but with greater sensitivity. Again, results from HMMs must be checked critically. There are a number of species-specific BLAST applications for human, microbial, and other genomes as well as for the analysis of expression or immunological data and other special cases. These are available at the NCBI-BLAST web page [ncbi-blast].

4.2.1 Important Algorithms for Database Searching

- Needleman & Wunsch (1970): A global alignment that was first developed without gap functionality. The method is very time intensive due to its dynamic procedure. A dynamic procedure is a solution to a problem which is broken down into sub-problems and the best results then compared.
- Smith & Waterman (1981): A local alignment that was originally developed without gap functionality. The method is very much similar to that of Needleman & Wunsch and also time-consuming.
- FastA (Pearson and Lipman 1988): A local alignment that is very fast due to the use of a heuristic method (making assessments to get almost exact results). The method identifies short word regions and then uses a dynamic procedure to obtain a gapped alignment.
- BLAST (Altschul et al. 1990): A local alignment that can identify segment pairs of constant length quickly due to the use of a heuristic method. Segments are then prolonged until preset threshold parameters are reached. BLAST is up to 100-fold faster than the Smith & Waterman algorithm.
- Gapped BLAST (Altschul et al. 1997): A local alignment that looks only for a single segment pair. This segment pair is then prolonged by gaps in both directions. The gapped BLAST algorithm is three times faster than the ungapped BLAST algorithm.
- Patternhunter (Ma et al. 2002): A proprietary algorithm for homology searching in databases. Patternhunter is faster and more sensitive than gapped BLAST. The program was successfully used for the annotation of the mouse genome.

4.3 Software for Sequence Analysis

Besides gene and protein sequences, NCBI, EBI, and other publicly accessible servers also provide genomic sequences.

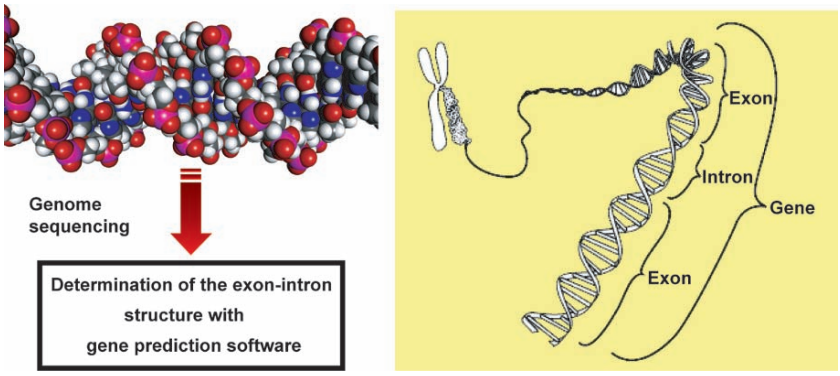


Fig. 4.9. The identification of new genes and proteins by genome sequencing

Such sequences are usually raw since they are published directly by sequencing units such as the Sanger Institute [sanger]. The advantage of raw sequence data is that predicted genes can be directly identified (Fig. 4.9). A number of software solutions for gene predictions are offered on the WWW. The Genscan server of the Massachusetts Institutes Technology [genscan] is particularly important and is based on neural networks that are trained to extract the exon–intron structure of eukaryotic genes from genomic sequences. A typical result of a Genscan analysis is shown in Fig. 4.10. Another software available for gene prediction in prokaryotic sequences is Glimmer [glimmer].

An interesting development in the area of sequence analysis is EMBOSS, the European Molecular Biology Open Software Suite (Rice et al. 2000) [emboss]. EMBOSS is an *Open Source Project* for different UNIX operating systems. The functional range of the software package grows steadily and is comparable with commercial packages such as that from GCG Wisconsin (Accelrys Inc.) or the DNA-Star (DNASTAR Inc.) and Vector NTI softwares (Informax Inc.). EMBOSS programs are also available as part of the commercial software package *Trinity* from the bioinformatics company Metalife AG [metalife]. This is a complete software solution for bioinformatics and contains algorithms for sequence analysis and the comparison of genomes. Moreover, it offers an extensive database collection (see Chaps. 1 and 3).

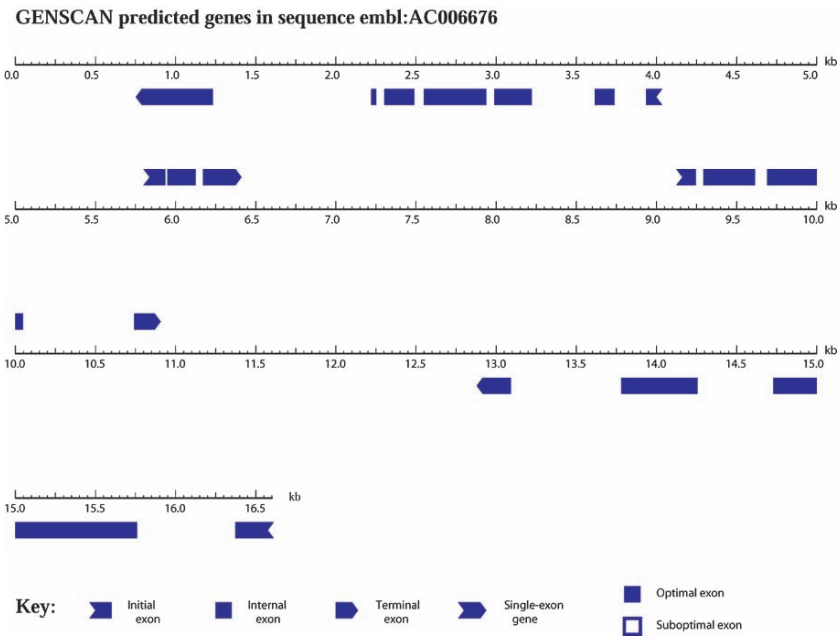


Fig. 4.10. Graphic version of the result of a Genscan analysis [genscan]

Expasy [expasy] and EMBnet [embnet] should also be mentioned. Besides databases, Expasy offers a number of hyperlinks to bioinformatic software. EMBnet is an association of 34 scientific groups in Europe and offers some free software for sequence analysis. In the exercise section below we will identify other software packages and their applications. A comprehensive compilation of bioinformatic applications available over the WWW is published once a year in the journal *Nucleic Acids Research* [nar] (*WEB server Issue*).

4.4 Exercises

1. Calculate the optimal alignment for the following sequences (see Fig. 4.4):
   Sequence 1: MTPARGSALS
   Sequence 2: MTPVRRSLS

Use the EMBOSS application *Needle* (http://bioweb.pasteur.fr/seqanal/interfaces/needle.html) to do this. Calculate the scores for the similarity matrices BLOSUM62, PAM250, and PAM30 using a gap penalty of 1. Do the suggested similarity matrices lead to similar alignments or are there differences?

2. Look for the Swissprot database record for the human 5-hydroxytryptamine 2A receptor in the NCBI protein database (http://www.ncbi.nlm.nih.gov) and save the protein sequence in FASTA format.
3. With the saved sequence from exercise 1, perform a BLAST search for similar sequences in the nonredundant protein database of NCBI. Do this by going to the NCBI-BLAST page (http://www.ncbi.nlm.nih.gov/BLAST/). How many similar sequences are found? What information can be extracted from the graph on the results page?
4. At the NCBI nucleotide database look for the entry with the AN AB037513 and save the nucleotide sequence in FASTA format. The sequence encodes a human 5HT2 receptor. Then perform BLAST searches using *blastn* and *tblastx* against the genome database *refseq_genomic*. Limit the search to the organism *Drosophila melanogaster*. How many similar sequences are found in each case? What can be stated regarding the quality of the hits? What are the differences of the two programs *blastn* and *tblastx* and how do the respective search results originate?
5. On the NCBI-BLAST page with *blast2Sequences* algorithm, perform a local alignment of the protein sequences P28223 and Q24511. The ANs can be entered directly so that no further database queries are required. The two sequences are the already-mentioned human 5HT2 receptor and its ortholog in *Drosophila melanogaster*. How can the result be interpreted?

6. Perform a multiple alignment with the sequences gi|543727, gi|7296517, and gi|10726392 with CLUSTALW (http://www.ebi.ac.uk/clustalw/). How can the result be interpreted? Note: sequences must be downloaded first from the NCBI database and then saved in FASTA format. Then go to the web page of Expasy (http://www.expasy.org) and under *Tools and Software Packages* click on the hyperlink *Alignment*. Under *Multiple,* follow the hyperlink to CLUSTALW at EBI and leave the default values unchanged. However, the output format should be changed to `gcg MSF`. Enter the three sequences into the corresponding text field and start the analysis. The results page is subdivided into several sections. In the middle section are found the multiple alignments of the sequences. Please save this alignment as a text file with the ending `.msf` using the browser. To do this, use the button *View Alignment File* and save the alignment with the ending `.msf`. Then open the file with the program *GeneDoc* or a similar program to visualize the alignment. Should there be problems in opening the file (e.g., file type not found), one reason can be that the file ending is not `.msf`. In this case, simply edit the file name in Windows Explorer by attaching the ending `.msf`. The alignment can be edited in *GeneDoc*, e.g., by changing the sequence names or the color representation of the alignments. *GeneDoc* can be downloaded and installed free of charge from the WWW (http://www.psc.edu/biomed/genedoc/). An alternative to *GeneDoc* for the visualization and processing of multiple alignments is *BioEdit* (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). When using *BioEdit*, however, the ending `.aln` is required in the output format. If program installation is neither possible nor desired, the CLUSTALW output may be viewed directly on the results page of

the EBI server. At Expasy a great variety of useful programs are freely available and a thorough study of this page, particularly the hyperlinks to Swissprot, Prosite, and InterPro, is highly recommended.

7. Perform a multiple alignment with the following sequences analogously to exercise 5 and calculate a phylogenetic tree for the proteins gi|19424144, gi|21245114, gi|2499874, gi|4503155, gi|1705638, gi|15214962. How can the result be interpreted? To what kind of proteins do the sequences belong to? Note: save both the alignment and phylogenetic tree as files using the browser. Have a look at the alignment with the program *GeneDoc*. Load the file with the ending `.dnd` into the program *Treeview* and study the calculated phylogenetic tree in the graphical outputs *radial*, *cladogram*, and *phylogram*. The program *Treeview* can be obtained free of charge from the WWW (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).
8. Find an entry for a eukaryotic cosmid in the NCBI nucleotide database, e.g. AC012088, and display the sequence in FASTA format. In a second browser window go to the *Genscan* server (http://genes.mit.edu/GENSCAN.html) and copy the sequence by cut & paste into the corresponding window. Then run *Genscan*. Try to interpret the result. Search for further cosmid sequences of different species and repeat the exercise.

WWW Links

bioedit: http://www.mbio.ncsu.edu/BioEdit/bioedit.html
clustalw: http://www.ebi.ac.uk/clustalw/
embnet: http://www.embnet.org/
emboss: http://bioweb.pasteur.fr/seqanal/EMBOSS/
expasy: http://www.expasy.org/
genedoc: http://www.nrbsc.org/gfx/genedoc/index.html
genscan: http://genes.mit.edu/GENSCAN.html

glimmer: http://cbcb.umd.edu/software/glimmer/
metalife: http://www.metalife.de/
nar: http://nar.oxfordjournals.org/
ncbi-blast: http://www.ncbi.nlm.nih.gov/blast/
sanger: http://www.sanger.ac.uk/
treeview: http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
wublast: http://blast.wustl.edu/

Literature

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Dayhoff MO, Schwartz RM, Orcutt BC (1978) In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol. 5, suppl. 3, p 345, NBRF, Washington/DC

Eddy SR (2004) What is a hidden Markov model? Nat Biotechnol 10: 1315–1316

Gerlt J, Babbitt P (2001) Respond: Orthologs and paralogs – we need to get it right. Genome Biol 2(8):1002.1–1002.3

Henikoff SB, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919

Jensen RA (2001) Correspondence: Orthologs and paralogs – we need to get it right. Genome Biol 2(8):1002.1–1002.3

Ma B, Tromp J, Li M (2002) PatternHunter: faster and more sensitive homology search. Bioinformatics 18:440–445

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Pearson WR, Lipman DJ (1998) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 4:2444–2448

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. Trends Genet 16:276–277

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 287:631–637

Tatusova TA, Madden TL (1999) Blast 2 sequences – a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 174:247–250

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Zhang Z, Schaffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV, Altschul SF (1998) Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res 26:3986–3990

5 The Decoding of Eukaryotic Genomes

5.1 The Sequencing of Complete Genomes

A new era in genome research started in 1995 with the publication of the first completely sequenced bacterial genome from the human pathogen, *Haemophilus influenzae*. For the first time one could analyze a complete genome, including both genes and their regulatory regions. Three years later, the sequencing of the first multicellular eukaryotic genome, from the nematode *Caenorhabditis elegans,* was completed. Eukaryotic genomes are larger and far more complex than those from bacteria (see Chap. 8). A comparison of the eukaryotic and prokaryotic genomes demonstrated that genes encoding proteins constitute a much smaller portion of the eukaryotic genome. Thus, in humans and mice just 1.4% of the genome actually encodes genes and only 5% of both genomes are highly conserved even though both share approximately 80% gene orthology. In addition to protein-encoding genes, the conserved regions contain important regulatory elements, non-protein encoding genes and regions important for chromosome structure. For the greater proportion of the genome, however, there are few data regarding function (Mouse Genome Sequencing Consortium 2002).

The relatively low number of genes identified in the human genome was, at first, surprising. At the beginning of the human genome sequencing project, it had been estimated that the number of genes would be in the order of 100,000 to 150,000. By completion, however, only 20,000 to 25,000 genes could be

demonstrated. A similar number of genes were also estimated for the mouse genome. Interestingly, humans possess only about 3,000 genes more than the nematode *C. elegans*. In view of the fact that a human body contains several billion cells while *C. elegans* has just 959 somatic cells, this small difference in the number of genes is remarkable.

5.2 The Characterization of Genomes using STS and EST Sequences

5.2.1 Sequence Tagged Sites are Landmarks in the Human Genome

It was a huge achievement to sequence the entire human genome. More than three billion nucleotides had to be sequenced and assembled in the right order. In a sense, the project could be compared to assembling a large jigsaw puzzle. It was first necessary to establish landmarks in the genome to allow for the correct placement of sequence regions. The most important landmarks in the genome are Sequence Tagged Sites (STSs), short DNA sequences 200–500 nucleotides in length that are present only once in the genome of an organism. STSs are generated by the polymerase chain reaction (PCR), a method for the amplification of specific nucleotide sequences. Because STSs are unique they can always be specifically amplified by PCR from genomic DNA.

DNA clones are examined by database searches for the existence of matching STS regions and are then positioned on chromosomes or in genomes. Using this approach a precise physical map of the humane genome could be generated.

As of 1994, a dedicated database for STSs has existed, the dbSTS [dbsts]. Here, one can find all the information available for individual STSs, including the STS name, sequences of the oligonucleotides necessary for PCR amplification, size of the PCR product, conditions for the PCR and nucleotide sequence of

the STS. Besides dbSTS, the database UniSTS [unists] is located at NCBI. This database contains a nonredundant collection of genetic markers and is based on the entries of dbSTS and other marker databases, such as the Genome Database [gdb].

Shortly after publication of the concept of STS-based mapping in 1989, it was recognized that STSs could also be generated from cDNA clones. Such cDNA clones originate from the cellular mRNA and, thus, correspond to the expressed genes of a cell. In addition to genome mapping, STSs derived from cDNA can also be used to localize genes within a genome. Indeed by 1996, a genetic map of the human genome had been assembled.

5.2.2 Expressed Sequence Tags

It was soon realized that partial sequences of cDNA clones are also well suited to the discovery of new genes (Adams et al. 1991). Because cDNA clones are derived from expressed genes, the sequences were called Expressed Sequence Tags (ESTs). ESTs are generated by end-sequencing of cDNAs (Fig. 5.1). ESTs are easy to produce at a reasonable price and many EST projects were initiated resulting in the identification of new genes. However, the concept of EST sequencing also met with opposition. Critics noted that sequencing just cDNA would miss important and non-expressed gene regulatory regions. Secondly, some ESTs are just too short to assign a gene function and, finally, ESTs, being automatically generated, can be of poor sequence quality. Frequently, not just nucleotide changes are, but also base insertions and deletions that lead to frameshift errors. Together, it was feared that many public EST databases would be of poor quality.

Despite these criticisms, EST projects became widely accepted. In particular, the speed with which ESTs could be generated on a high-throughput scale (due to the automation of DNA sequencing technology and plasmid DNA production) resulted in a real boom for EST projects. Important EST projects were initiated at Washington University (WU) [washington], for example. In collaboration with the American pharmaceutical company, Merck,

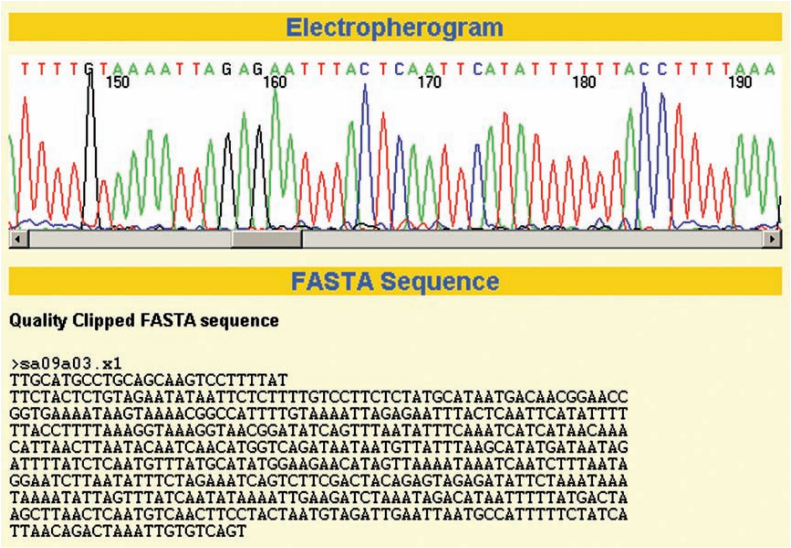


Fig. 5.1. Section of an electropherogram from a dideoxy DNA sequencing reaction with the corresponding nucleotide sequence of the expressed sequence tag. (Clipping from the database Ensembl, printed with permission of the EBI, Hinxton)

WU sequenced 580,000 human ESTs between 1995 and 1997. These ESTs were generated from cDNA libraries that had been made available by the IMAGE consortium. IMAGE stands for *Integrated Molecular Analysis of Genomes and their Expression* and is a merger of several academic research groups that produce high-quality cDNA libraries and make them available for other research, such as EST projects. The IMAGE consortium has the biggest collection of publicly available cDNA libraries worldwide [image].

As a reaction to the huge increase in EST data, dbEST [dbest] was established at NCBI to collect all publicly accessible ESTs. In 1993, less than 50,000 sequences were present in dbEST; today, however, more than 46 million ESTs from over 1,400 organisms are stored in this database (dbEST release 101207). One drawback of dbEST is that it contains redundant ESTs, especially for strongly expressed genes, for example, actin. For this reason, the UniGene database [unigene] was established in which

every cDNA and EST that originates from an identical gene is combined into a group or cluster. The result is a reduction in the number of entries down to the actual number of proteins produced in an organism. Because of its nonredundancy, therefore, UniGene is a useful basis for other databanks such as ProtEST and HomoloGene [homologene]. ProtEST is integrated into UniGene and provides information on whether cDNAs and ESTs that are assigned to an UniGene cluster are similar to known protein sequences upon translation. In contrast, the independent database HomoloGene provides information on whether human UniGene clusters have homologs in other species. A similar strategy to that of UniGene is pursued by the Gene Index Project of the Dana Farber Institute [tgi]. Again, in these databases, all available ESTs and cDNAs that can be assigned to an identical gene are combined into a cluster. This allows for a nonredundant representation of the sequences. The Gene Indices are ordered by species and can be searched by text queries and other sequences. The Human Gene Index contains more than 7 million sequences that are divided into approximately 330,000 clusters (Release 17.0).

Another NCBI database, dbGSS [dbgss], stores Genome Survey Sequences (GSSs). Like ESTs, GSSs are partial nucleotide sequences with a length of up to 1,000 bases and generated by end-sequencing individual clones. The difference between GSSs and ESTs is the nucleic acid source material: GSSs are prepared from genomic libraries, whereas for ESTs, cDNA libraries are used. Thus, GSSs differ from ESTs by potentially containing DNA fragments that lie outside of areas encoding genes. More than 21 million sequences from 700 organisms are stored in dbGSS (dbGSS release 101207).

5.3 The Implementation of an EST Project

At the beginning of an EST project, the starting material for the construction of a cDNA library is selected. This can be cells,

specific tissues or even whole organisms (Fig. 5.2). From this material total RNA is isolated, which predominantly comprises rRNA (ribosomal RNA), tRNA (transfer RNA), and mRNA (messenger RNA). Messenger RNA is most interesting for the construction of a cDNA library as it represents all active genes of a given cell or tissue. It is present only in very small amounts (approx. 3% of the total RNA). The very unstable mRNA is transcribed into the considerably more stable cDNA (complementary DNA) by the viral enzyme reverse transcriptase. The cDNA is then cloned into plasmids that serve as vectors. Usually cDNAs are cloned directionally, i.e., it is known at which end of the vector the 5′ and 3′ ends of the cDNA are located. Plasmids are amplified by transforming the bacterium, *Escherichia coli,* resulting in the desired cDNA library, which can then provide the basis for generating EST sequences. The transformed bacteria are plated and grown on nutrient media, and plasmid DNA is isolated from randomly selected individual clones. The cloned cDNA can then be sequenced either from the 5′ or 3′ end or from both ends simultaneously. The identified nucleotide sequence is then exported to a computer and the raw data are bioinformatically processed.

The quality of the data is first checked in a process called quality trimming. For example, quality trimming defines the minimum length that an EST must have and what number of ambiguous nucleotides (variable N) is allowed relative to the nonambiguous nucleotides (A/T/G/C). Modern sequencers permit the computation of quality scores that are a measure of the quality of the sequencing of each individual nucleotide. Using these values, sequence regions of poor quality, e.g., the ends of sequences are removed. Finally, any contamination with sequences from vector and bacteria are also removed.

Curated ESTs are a collection of random cDNA sequences of different lengths and many are derived from identical transcripts. Particularly, for highly expressed genes many ESTs will be found. To eliminate redundancy, therefore, alignments of these ESTs are generated to form overlapping sequences that are as long as possible (Fig. 5.3). These consensus sequences are compared again to other ESTs so that further identical ESTs are incorporated into

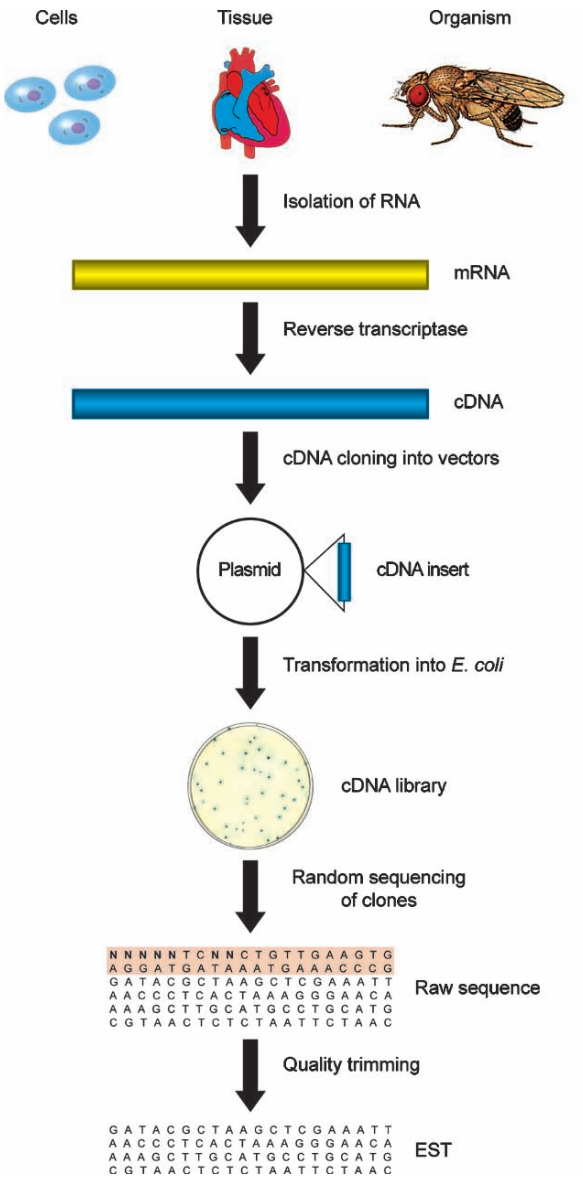


Fig. 5.2. Diagram for the establishment of a cDNA library and the generation of EST sequences. (*Drosophila melanogaster* from Patterson JT, Univ. Texas Publs 4313, 1943, printed with permission of the University of Texas; Heart from Schmidt, Thews Lang, Physiologie des Menschen, 28th edition 2000, printed with permission from Springer Verlag, Heidelberg)

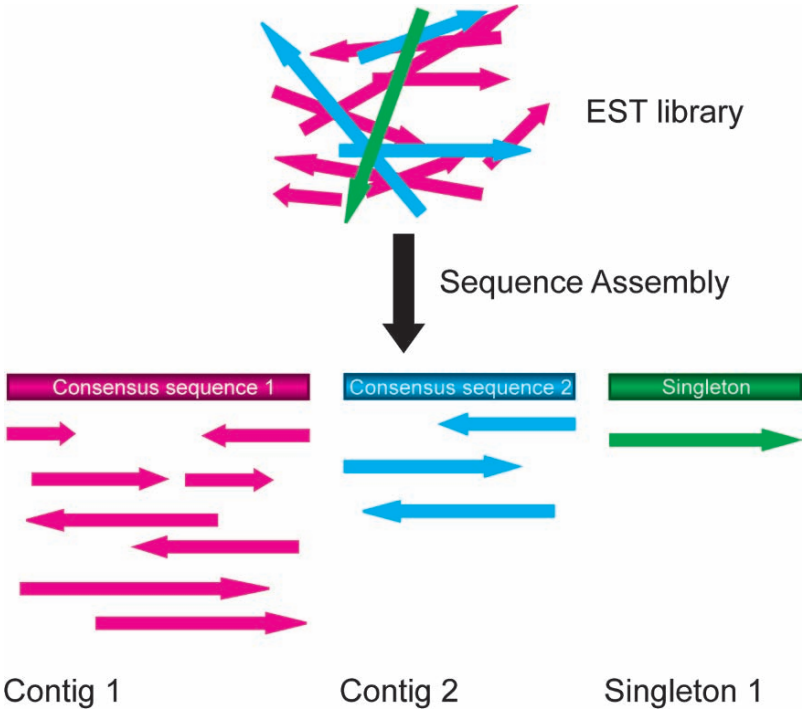


Fig. 5.3. Classification of expressed sequence tags into contigs and the formation of consensus sequences

the alignment. This iterative process is described as sequence assembly. Often sequence assembly programs such as CAP3 [cap] and Phrap [phrap] are used. Sequence assemblies are either contigs whose sequences correspond with the consensus sequences of the alignments or singletons that are not similar to other ESTs and, therefore, cannot be grouped in contigs.

For large EST datasets, it can be useful to subdivide the ESTs into groups or clusters first. Those clusters displaying identical nucleotides for a given region are summarized into groups. Finally, within these groups, a more stringent sequence assembly is performed to generate consensus sequences. Thus, ESTs that descend from alternatively spliced forms are arranged into the same clusters, but different contigs, better depicting the EST relationships. One useful program for sequence clustering is stackPACK [stackpack].

5.4 The Identification of Unknown Genes

Once ESTs are arranged into contigs the corresponding consensus sequences can be used to identify unknown genes. For this purpose annotation and sequence searches are carried out against various databases.

ESTs are usually first annotated, i.e., a potential function is assigned on both the level of the single ESTs and the assembled contigs by comparison with existing proteins of known function. Usually the BLASTx algorithm is applied whereby the EST nucleotide sequences are first translated into all six reading frames. This process is shown in Fig. 5.4 using an EST sequence obtained from bovine intestine. The EST was annotated by BLASTx against a nonredundant protein database and shows high similarity with part of murine caspase 6. Caspases are

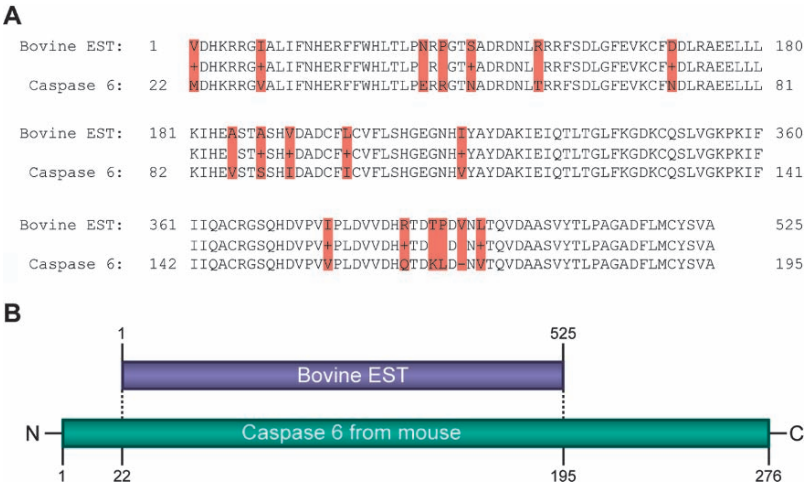


Fig. 5.4. Annotation of an EST sequence from bovine intestine. (**a**) The translated EST has an identity of 89% over a length of 175 amino acids (525 nucleotides) with murine caspase 6. Sequence differences are highlighted in *red*. The numbering of the EST sequence from 1 to 525 refers to nucleotides. In contrast, the numbering of the caspase 6 protein from 22 to 195 refers to amino acids. (**b**) Schematic of the alignment of the EST sequence with the sequence of murine caspase 6

proteases that function during programmed cell death (apoptosis). Because of the similarity to caspase, it can be inferred that the gene transcript from which the EST is derived encodes either a true caspase or a protein containing a caspase domain. It is important to state that ESTs are usually partial gene sequences and, therefore, alignments may not contain the entire length of a deduced protein. Indeed, ESTs often encode only the untranslated region (UTR) of mRNA and these are known as nonencoding ESTs (Fig. 5.5). These difficulties can be avoided, however, when ESTs are extended by sequence assembly, sometimes to the point where the entire protein can be identified.

By direct comparison of EST sequences between different organisms, similar or even new genes or proteins may also be identified. Generally, however, it is not advisable to attempt this at the nucleotide level (e.g. with BLASTn) as little similarity exists between species due to species-dependent codon usage (see Chaps. 2 and 8). However, sequences normally show greater conservation at the amino acid level. Therefore, sequences should be compared after translation of the nucleotide sequence into all six reading frames. For this, tBLASTx, which automatically carries out both the translation and the database comparison, is

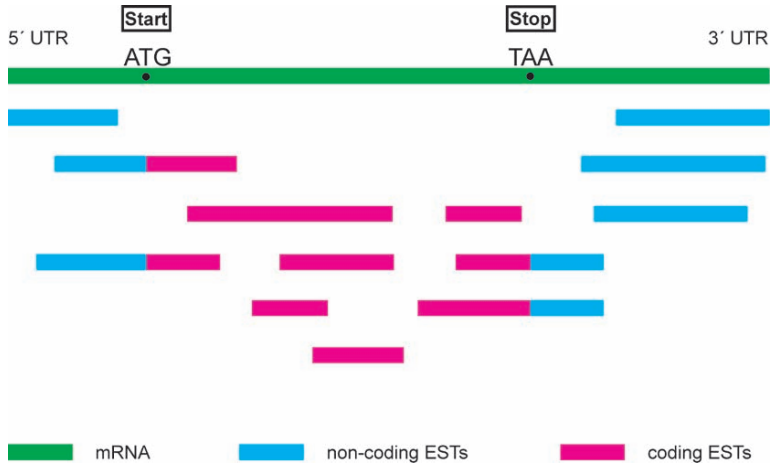


Fig. 5.5. EST sequences are derived from coding and noncoding segments of an mRNA

a good choice (see Chap. 4). However, when large databases are being queried, some time may be needed. An interesting example of a large comparison is the evaluation of EST sequences from different parasitic worms. At Washington University in St. Louis, the parasitic nematode sequencing project is underway in which more than 300,000 ESTs of different parasitic threadworms are being sequenced [nematode]. By comparison of these datasets it will be possible to find genes that are ubiquitous in all nematodes. Such an approach has been used to clarify evolutionary relationships within the phylum Nematoda (Blaxter et al. 1998).

Using EST data, new members of a protein family can also be identified. The procedure to identify new protein kinases in the nematode EST dataset is shown in Fig. 5.6. To start, one compares the peptide sequence of a known protein kinase (e.g. from mouse) with an EST database (e.g. dbEST). If a nematode EST sequence(s) of high identity to the mouse kinase is found then it is likely that the EST encodes a protein kinase. To determine whether the identified protein kinase(s) is novel, the ESTs must be compared with a nonredundant protein or nucleotide database. If no identical sequences are identified, then a new member of the protein kinase family has indeed been found.

5.5 The Discovery of Splice Variants

ESTs can not just help identify new genes but also alternative gene splice variants. Alternative splice variants can arise upon gene transcription and during the processing of the RNA primary transcript. During splicing, noncoding introns are removed from the primary transcript and the remaining exons joined to form the mature mRNA (see Chap. 2). During alternative splicing, one exon can be replaced by another thereby creating a new mRNA. In this way, different mRNAs, encoding different proteins, can arise from a single primary RNA transcript (Fig. 5.7). Alternative splicing, therefore, is an efficient strategy for producing

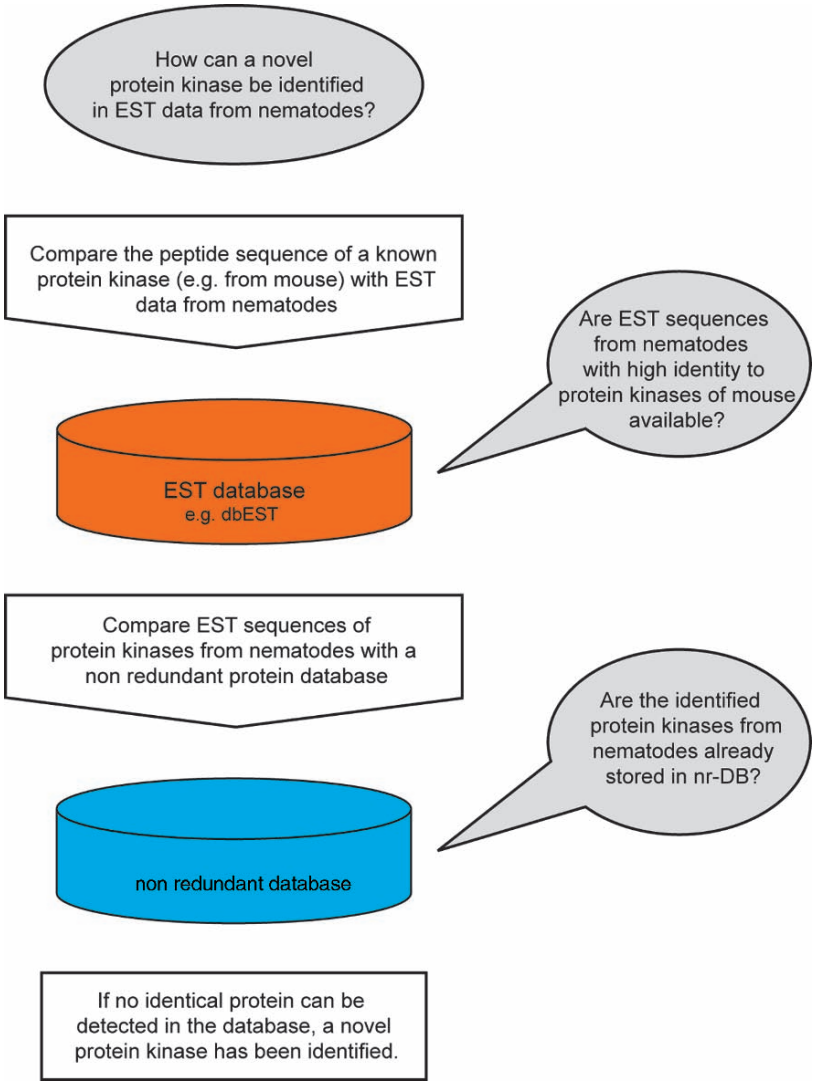


Fig. 5.6. Strategy for the identification of new members of protein families

several proteins from one gene. It is believed that alternative splice forms exist for one-third to two-thirds of all human genes (Yeo et al. 2004). For example, two mRNA transcripts are known for the F_c receptor that is important in immunology. During alternative splicing the cytoplasmic domain of the receptor is exchanged

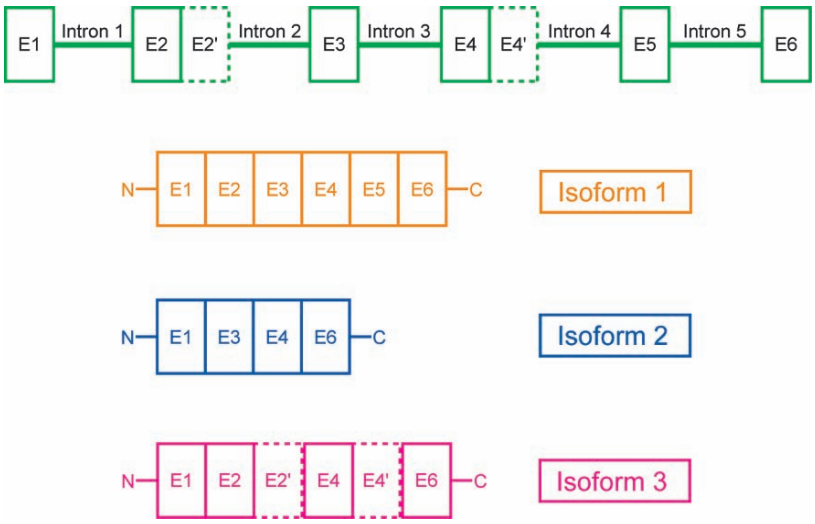


Fig. 5.7. Alternative splicing. The generation of several mRNA transcripts from a single gene by the combination of different exons (E) is called alternative splicing

for a second form. Because the individual cytoplasmic domains are crucial for signal transduction, alternative splicing generates domains with very different cellular functions.

ESTs derived from fully processed mRNAs can give valuable hints as to the identification of unknown splice variants. ESTs are compared with nucleotide databases that contain information for mRNA transcripts (e.g. Genbank) or protein databases (e.g. UniProt). In the case where otherwise identical sequences are found to differ in a few regions, e.g. by insertions or deletions, this can be evidence for alternatively spliced variants. Through such EST comparisons with known sequences in the public databases, numerous alternatively spliced gene variants have already been discovered. At the University of California Los Angeles, a database called Alternative Splicing Annotation Project has been established in which alternatively spliced genes, identified via EST sequences, are stored [asap]. Also, many gene prediction programs such as GrailEXP use EST sequences to correctly predict genes from sequenced genomes and derive information regarding splice sites [grailexp].

5.6 Genetic Causes for Individual Differences

A characteristic of eukaryotic genomes is the existence of mutations or genetic variations. These variations are responsible for the individual differences in a population. The most frequent variations are Single Nucleotide Polymorphisms (SNPs) caused by the exchange of a single nucleotide. Other polymorphisms are short deletions and insertions (deletion insertion polymorphisms) as well as variations due to repetitive sequences (short tandem repeats).

A consortium of commercial and noncommercial institutions has identified almost 1.8 million SNPs in the humane genome [snp syndicate]. Many of these SNPs lie outside genes and, therefore, do not alter cellular function. However, other SNPs lie within genes and are responsible for the occurrence of phenotypes. Example phenotypes are the color of eyes or hair, but also disease conditions. Functionally important SNPs are discovered by comparing the appearance of a phenotype with the frequency of a specific SNP. If a correlation is found, it is likely that this SNP is responsible for the phenotype. As individuals are randomly selected for these correlation analyses, the strategy is simpler and faster than classical pedigree analyses in which the appearance of phenotypes must be traced back in a family over several generations.

An example of a SNP-based disease is phenylketonuria in which the degradation of phenylalanine is disrupted. Point mutations in the enzyme phenylalanine hydroxylase lead to inactivation of the enzyme. Many different SNPs have been discovered in the human phenylalanine hydroxylase enzyme and these are collated in the database Phenylalanine Hydroxylase Locus Knowledgebase [pahdb]. Because of the missing enzyme activity, phenylalanine accumulates in the brain of newborns and infants, and ultimately leads to a mental defect. Newborns are therefore examined in many countries for high blood levels of phenylalanine. Disease symptoms are preventable by a phenylalanine-poor diet allowing those affected to live a normal life.

Genetic polymorphisms can also be an advantage. One example is the differential susceptibility of individuals to infection with the Human Immunodeficiency Virus 1 (HIV 1). In addition to the surface protein CD4, the virus requires additional co-receptors such as the chemokine receptor CCR5 to enter the cell. A mutant of this receptor with a deletion of 32 nucleotides was discovered in 1996. This mutation leads to a shift in the reading frame and subsequently to the translation of a nonfunctional protein that is no longer present at the cell surface. Humans who are homozygous for this mutation (both alleles disrupted) are more resistant to HIV 1 infection. Those who are heterozygous for the mutation (one functional allele) will develop AIDS later and have a longer life expectancy than those who lack the frame shift mutation. In the Caucasian population of the USA, this polymorphism is homozygous at a frequency of 1%; another 20% have a heterozygous allele. Unfortunately, among African and East Asian populations, this polymorphism is found only rarely (Berger et al. 1999).

SNPs are also excellent genomic markers because they are distributed over the entire genome and found at high density (on average every 300–500 nucleotides in the human genome). Moreover, SNPs have a low mutation frequency between generations and are detectable by high-throughput methods. SNPs, therefore, allow the generation of precise genetic maps of high resolution. This resolution facilitates the discovery of disease genes, particularly if several genes are responsible for the emergence of complex illnesses like cancer or diabetes.

A number of methods exist for the detection of SNPs or genotyping. Microarray genotyping is based on the principle that the denaturation temperature of hybridized DNA strands will decrease if nonidentical nucleotides are present. The advantage of this high-throughput method is the simultaneous and parallel analysis of many sequences. Other techniques for identifying SNPs are based on enzymatic reactions that show a very high specificity for their substrate and are thus, more accurate than hybridization-based methods. A commonly used enzyme-based genotyping technique is pyrosequencing [pyrosequencing]. Short

DNA segments are sequenced in real time without the necessity for time-consuming gel purification steps. The advantage of this method is that the entire vicinity of the SNP is sequenced and serves as an internal control for the sequencing reaction. An alternative enzyme-based technology is Single base primer extension, which provides precise quantitative results at a moderate cost. Short oligonucleotide sequences hybridize just next to the SNP. These oligonucleotides then serve as primers for polymerases that incorporate a labeled nucleotide at the position of the SNP. The incorporated nucleotide is then detected using colorimetric methods or by mass spectroscopy. Furthermore, SNPs can also be determined in silico, i.e., by graphically comparing similar EST sequences from different individuals. Using such multiple alignments, nucleotide exchanges are very easy to recognize. However, caution is advised when describing new SNPs using ESTs as these can contain sequencing errors interpretable as SNPs.

The database, dbSNP, is the NCBI repository for polymorphisms [dbsnp]. Each entry contains details of the genetic variation, adjacent nucleotides, and frequency of the polymorphisms. It also includes data about the experimental method and conditions used to identify the SNP. The dbSNP contains almost 88 million polymorphisms from 43 organisms, of which, 31 million are human (dbSNP Build 127). Moreover, a curated collection of human SNPs can be found in the Human Genome Variation Database [hgvbase]. These SNP entries have been subject to an additional quality check and are completely annotated. At present, approximately 40% of all human dbSNP entries are stored in this database.

5.6.1 Pharmacogenetics and Individual Medicine

Pharmacogenetics (or pharmacogenomics) deals with genetic variations that are responsible for how patients differ in their reactions to drugs. A study in the US in 1994 reported that 2.2 million patients suffered from serious medication side effects

and that over 100,000 patients died. Thus, there is a greater chance of dying from drug side effects than from most viral infections. Accordingly, the ability to predict how a patient might react to a drug prior to starting therapy would be a tremendous advance.

How a patient responds to drugs is a complex process involving many different proteins, including the receptors and enzymes that bind to and metabolize the drug, respectively. Genetic variations in such proteins can result in decreased or absent drug-binding or metabolism. Of particular importance are polymorphisms in proteins of the cytochrome P450 family. For example, the enzyme CYP2D6 is responsible for the metabolism of 20–25% of all prescription drugs. Mutations in CYP2D6 can influence the rate at which drugs are metabolized. Depending on the type of mutation one can distinguish patients with ultra-fast, extensive, medium, or slow drug metabolism. Clearly, therefore, genetic polymorphisms greatly influence the individual reactions of patients to drugs. As SNPs represent by far the most frequent genetic variations, the search for SNPs that influence a drug's effect or metabolism is of central importance to pharmacogenetics.

As stated, a major aim of pharmacogenetics is to predict unwanted side effects of a drug in advance of therapy. An important prerequisite for this is the development of diagnostic tests to understand the genetic disposition of a patient and how they might react to a specific drug. In such diagnostic tests the genotype of every patient is established, i.e., whether relevant proteins such as drug-metabolizing enzymes show distinct polymorphisms. The patient can then be classified into a corresponding group and a suitable therapy selected based on his/her genotype (see Fig. 5.8). This is also referred to as individual medicine because therapy is optimized for every patient. An example already practiced in many countries is the chemotherapeutic treatment of patients with acute lymphatic leukemia (ALL). Mercaptopurine and thioguanine are frequently used as drugs that are incorporated into the DNA of proliferating cells (especially cancer cells), leading to their eventual death. One

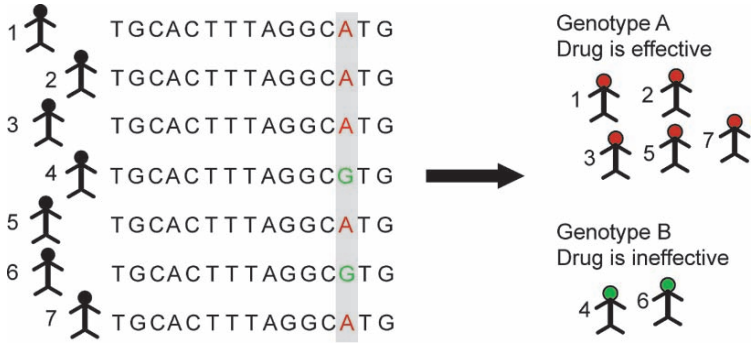


Fig. 5.8. Genotyping of patients by detecting single nucleotide polymorphisms

enzyme responsible for the metabolism of these compounds is thiopurine-S-methyltransferase. Clinical studies have shown that genetic polymorphisms greatly influence the activity of thiopurine-S-methyltransferase and, therefore, the toxicity and efficacy of mercaptopurine and thioguanine. Patients deficient in thiopurine-S-methyltransferase accumulate these drugs in blood cells at high concentrations eventually causing death. By contrast, in patients with high thiopurine-S-methyltransferase activity, mercaptopurine and thioguanine must be used at higher doses. Therefore, each patient is examined for polymorphisms in the gene encoding thiopurine-S-methyltransferase and the most effective dose determined before treatment with mercaptopurine and thioguanine. In addition to patients in the clinic, pharmacological research has also benefited from pharmacogenetics. Prior to approval for use in patients, every new drug candidate must be tested in extensive clinical studies using the strictest safety and efficacy criteria. Pharmacogenetics offers the possibility of excluding those patients unlikely to react to the therapy or who might experience undesired side effects before the start of each study. This selection process increases the chance of a drug reaching the market by appropriately selecting those patients who will benefit from the drug without unpleasant or even dangerous side effects. Moreover, pharmacogenetics will allow the development of new drugs

for patient groups that do not respond to existing therapies. Overall, it is expected that pharmacogenetics will accelerate the approval of drugs of increased quality.

It should also be considered that individual reactions to drugs can only be partly explained by genetic variations and that other factors may also influence drug efficacy and safety. These include the patient's age and nutritional status, whether additional disease conditions co-exist or other drugs are being taken. Thus, the field of pharmaco-metabonomics takes into account both genetic differences and environmental factors that can influence drug efficacy and safety (Clayton et al. 2006).

5.7 Exercises

1. How many ESTs does the database dbEST at NCBI contain (http://www.ncbi.nlm.nih.gov/dbEST/index.html)? Which two organisms have the most entries and what is their percentage of the total number of entries?
2. Determine by querying dbEST how many *Mangifera indica* ESTs there are. Note: Enter the name Mangifera indica on the home page of dbEST. Then repeat the search and this time enter Mangifera indica [ORGANISM]. Explain the differences of the two results.
3. Save the result of the second search in FASTA format.
4. Using the saved sequences perform a sequence assembly. Use the CAP3 sequence assembly program of the PBIL institute (http://pbil.univ-lyon1.fr/cap3.php). Note: The number of sequences should not exceed 100. How many contigs are built? How many ESTs does the contig with the most sequences contain? Also, are there ESTs that are not grouped into contigs (singletons)?
5. Annotate the ESTs by comparing the contigs with a non-redundant protein database using the BLASTx algorithm. To do this, go to the NCBI BLAST home page. Can reliable hits for all contigs in the protein database be found?

6. Search for an EST with the accession number AI590371 using the database query system Entrez at NCBI. Save the sequence in FASTA format.
7. Compare the saved EST sequence with the nonredundant nucleotide database of NCBI. To do this, use the NCBI BLAST home page. How many reliable nucleotide sequence hits can be found in this database?
8. Some nucleotide sequences have hyperlinks to the database UniGene at NCBI. Click on this hyperlink and examine the stored information. What is the name of this UniGene cluster and how many ESTs are found? Which protein does this cluster encode? With which disease is the protein associated and what human population is predominantly affected by it?
9. What can be learned about the expression of the protein from the EST?
10. Use the hyperlink to the database ProtEST in which the results of a BLASTx comparison between the nucleotide sequences of the UniGene cluster and the sequences of a protein database are stored. How many nucleotide sequences align over the full length of the protein? Why does the UniGene cluster have more EST sequences than ProtEST?
11. Using the database query system Entrez at NCBI, locate the protein sequence of the mouse proto-oncogene c-myc with the accession number P01108. Save the sequence in FASTA format.
12. Compare the saved sequence of the protein c-myc with an EST database from the mouse. Use the NCBI BLAST home page to do this. Are mouse ESTs found in the database? What is noticeable about the distribution of the ESTs and how can this be explained?
13. In addition to very good hits (alignment score > 200, red colored bars), many hits with an alignment score of 80–200 (magenta colored bars) are found. Do these

ESTs also encode the protein c-myc? Give reasons for the result. Note: Compare the nucleotide sequences of this EST with the protein database Swissprot.

14. At the NCBI book collection find *Genes and Disease*. Inside is found information about phenylketonuria. On which human chromosome is the gene for phenylalanine hydroxylase found? Click on the *hyperlink* to the database Entrez Gene. Which information does this database provide?
15. In the dbSNP database at NCBI (http://www.ncbi.nlm.nih.gov/SNP/) search for the reference cluster with the ID rs334. In which organism is this single nucleotide polymorphism found? Examine the category GeneView. Compared to the reference sequence (contig reference), which nucleotide exchange is found? Does it result in an amino acid exchange and, if so, which one? Which gene is affected by this SNP? Follow the link of the gene name to the database Entrez Gene. Which disease is caused by the mutation?

WWW Links

asap: http://www.bioinformatics.ucla.edu/ASAP/
cap: http://seq.cs.iastate.edu/
dbest: http://www.ncbi.nlm.nih.gov/dbEST/
dbgss: http://www.ncbi.nlm.nih.gov/dbGSS/
dbsnp: http://www.ncbi.nlm.nih.gov/SNP/
dbsts: http://www.ncbi.nlm.nih.gov/dbSTS/
gdb: http://www.gdb.org/
grailexp: http://compbio.ornl.gov/grailexp/
hgvbase: http://hgvbase.cgb.ki.se/
homologene: http://www.ncbi.nlm.nih.gov/sites/entrez?db = homologene
image: http://image.llnl.gov/
nematode: http://www.nematode.net/
pahdb: http://www.pahdb.mcgill.ca/
phrap: http://www.phrap.org/
pyrosequencing: http://www.pyrosequencing.com/

snp syndicate: http://snp.cshl.org/
stackpack: http://fling.sanbi.ac.za/CODES/
tgi: http://compbio.dfci.harvard.edu/tgi/
unists: http://www.ncbi.nlm.nih.gov/sites/entrez?db = unists
unigene: http://www.ncbi.nlm.nih.gov/UniGene/
washington: http://genome.wustl.edu/data/est.cgi

Literature

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H et al (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656

Berger EA, Murphy PM, Farber JM (1999) Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. Annu Rev Immunol 17:657–700

Blaxter M (1998) Caenorhabditis elegans is a nematode. Science 282:2041–2046

Clayton TA, Lindon JC, Cloarec O, Antti H, Charuel C, Hanton G et al (2006) Pharmaco-metabonomic phenotyping and personalized drug treatment. Nature 440:1073–1077

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Yeo G, Holste D, Kreiman G, Burge CB (2004)Variation in alternative splicing across human tissues. Genome Biol 5(10):R74

6 Protein Structures and Structure-Based Rational Drug Design

6.1 Protein Structure

Proteins are macromolecules whose monomeric subunits are the naturally occurring 20 amino acids. The amino acids are linked via peptide bonds (generated upon water release) to form a polypeptide (see Chap. 2). Polypeptides can range in length from three to several hundred amino acids. The amino acid sequence of a given protein, also known as the primary structure, is genetically determined. It becomes fixed during translation based on the information encoded in the mRNA.

The properties of the polypeptide chain correspond to a cross-section of those of the corresponding amino acids, i.e., the function of the respective protein cannot be determined solely from the primary structure, but also depends on the spatial arrangement of the amino acids to one another. Stretched polypeptide chains fold spontaneously into secondary structures and then into three-dimensional structures. The secondary structure can comprise two main structural features, the α-helix and the β-sheet. These structural elements are connected via nonrepetitive elements called loops. It is the combination of the positioning of the amino acid side chains and the peptide backbone of the secondary structure that forms the protein tertiary structure. If a protein consists of several subunits then the association of these subunits to form the functional protein is called the quaternary structure.

The function of a protein is mediated by its three-dimensional structure, which if known can allow the inference of

function. Ab initio prediction of protein tertiary structure based solely on the primary structure is not yet possible. Also, the experimental determination of structure is still difficult and the number of known protein structures comparatively small. Therefore, the prediction of the protein function based on tertiary or quaternary structure is limited. However, proteins show a variety of structural and topological features, which can be used to predict their properties and functions. Many of these features can be inferred or predicted from the primary structure by computational methods. Some of these properties and their predictions are discussed below.

6.2 Signal Peptides

For many proteins the site of synthesis is not the site of action. This applies to transmembrane proteins, proteins within the endoplasmic reticulum, and proteins that are secreted or imported into lysosomes. Prior to their activation, these proteins must first be transported to the site of action and this is facilitated by a peptide recognition signal for the cellular transport system. The recognition signal is an N-terminal leader sequence (signal peptide) that consists of approx. 15–30 amino acids placed on the N-terminus of the mature protein (Fig. 6.1). According to the signal hypothesis of Günter Blobel and David Sabatini (Blobel and Sabatini 1971), the signal peptide is recognized by a signal recognition particle, guiding the nascent polypeptide chain through the membrane of the endoplasmic reticulum. As soon as the signal peptide has passed the membrane, it is specifically cleaved from the nascent polypeptide by a signal peptidase. Proteins with a signal peptide are called preproteins or, in those cases where they also contain propeptides, preproproteins. Unlike signal peptides, propeptides are proteolytically removed to allow for protein activation (Fig. 6.1).

The presence of a signal peptide gives an important hint as to the site of action of proteins. This knowledge, in turn, can

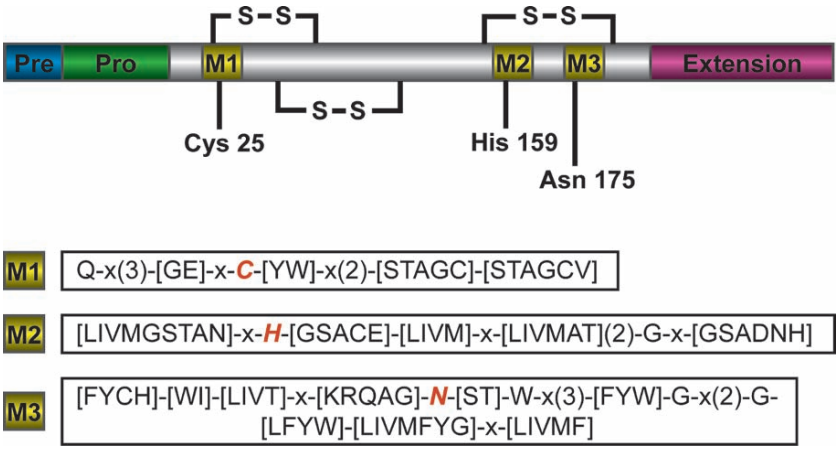


Fig. 6.1. Schematic illustration of a preproprotein exemplified by cysteine proteases of the papain family. The amino acids of the catalytic triad, Cys^{25}, His^{159}, and Asp^{175}, are each located within the characteristic sequence motifs of cysteine proteases (M1–M3). Only a few cysteine proteases have an additional C-terminal extension for which a function is still not known

help clarify function and, thus, assist in the choice of that protein being a target molecule for pharmacological research. For this reason, methods for the prediction of signal peptides in the primary structure have been developed. An example is the program SignalP from the Center for Biological Sequence Analysis (CBS) at the Technical University of Denmark [signalp] (Bendtsen et al. 2004). The recognition of signal peptides by the signal recognition particle is not due to a conserved amino acid sequence but depends on physicochemical properties. A signal peptide usually consists of three parts. The first region (the n-region) contains 1–5 usually positively charged amino acids. The second, h-region, is made up of 5–15 hydrophobic amino acids, and the third, c-region, has 3–7 polar but mostly uncharged amino acids. A classical sequence alignment method is therefore unsuitable for the prediction of signal peptides. The SignalP program is instead based on the use of neural networks as well as a *Hidden Markov Model* (HMM). Both methods were trained with a set of known sequences and are thus able to judge

the properties of amino acids in unknown sequences, thereby allowing the recognition of signal sequences.

Before the analysis is started it is important to choose the right organism group as the methods are trained on gram-negative, gram-positive bacteria or eukaryotes. In Fig. 6.2, the graphical output of the SignalP program for the *outer membrane protein C* (*precursor*) from *Salmonella typhimurium* (OMPC-SALTY, P0A263) is shown. Figure 6.2a shows the result of the neural networks for the predicted signal peptide as well as

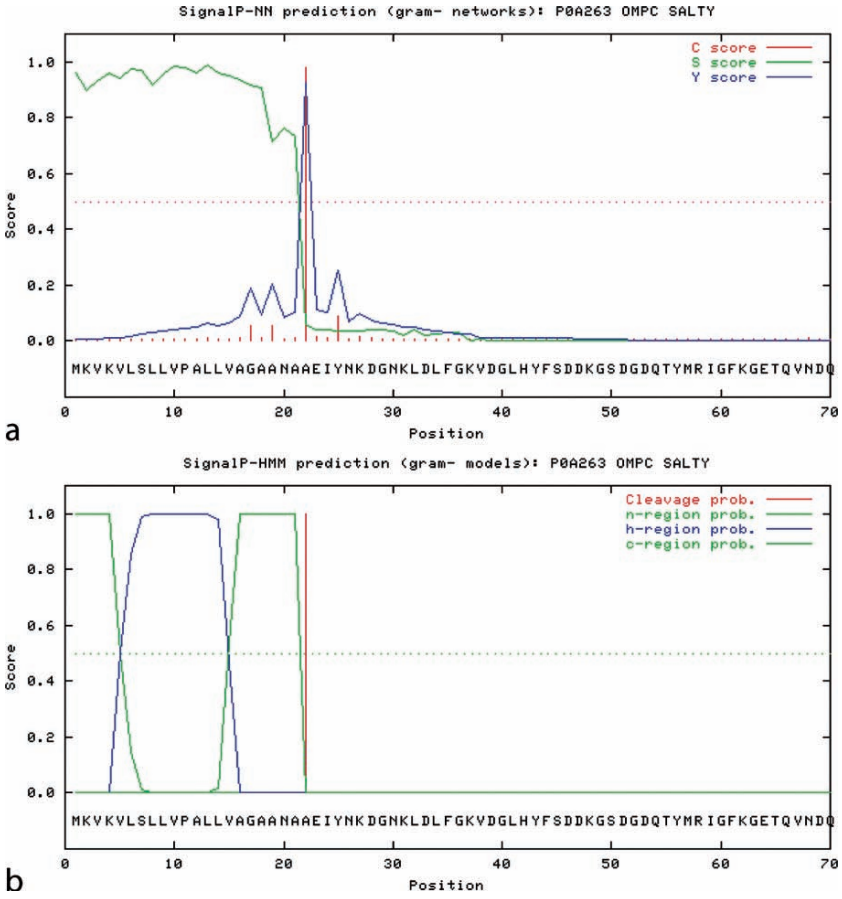


Fig. 6.2. Graphical output of the SignalP server [signalp] at CBS. (Printed with permission of the CBS)

the signal peptidase I cleavage site (SPase I). The graphical display contains three scores, *S*, *C*, and *Y*. The *S-score* is calculated for each amino acid and has a high value (between 0.5 and 1.0) if the corresponding amino acid is part of the signal peptide. Amino acids of the mature protein have a low *S-score* (between 0.0 and 0.5). So if a signal peptide is present the curve of the *S-score* will show an interval with a large negative slope. *C-score* stands for *cleavage site score* and predicts the cleavage site of SPase I. The *C-score* is high at the position of the first amino acid of the mature protein (position 22), i.e., the cleavage site is between positions 21 and 22. Because the *C-score* can be high at several positions, the correct prediction of the cleavage site from the *C-score* alone is difficult. In this case the simultaneous analysis of both the *C-score* and the *S-score* will help. At the correct cleavage site a high *C-score* coincides with the interval where the *S-score* shows its large negative slope. To make the analysis of the two scores easier, a third *Y-score* has been introduced. It combines the other two scores so that its value reaches a maximum at the predicted cleavage site. In addition, two more values are calculated, *S-mean* and *D-score*, which are given only as numbers, however. The *S-mean* is the average of the *S-scores* of all amino acids of the signal peptide. Consequently, if there is a signal peptide this value should be high. The *D-score* is the arithmetic mean of the *S-mean* value and the maximum value of the *Y-score*. It will also be high if a signal peptide has been predicted. It has been found that the *D-score* can better discriminate secreted from nonsecreted proteins than the *S-mean* value.

Figure 6.2b shows the graphical output of the HMM results. For each amino acid, the Hidden-Markov Model calculates the probability that it belongs to one of the three aforementioned regions of the signal peptide. From plotting these values it can be seen that the n-region goes from positions 1 to 4, the h-region starts at position 5 and extends to position 14, and the c-region goes from positions 15 to 21. The probability of the respective amino acids being part of one of the three regions is close to 1.0. Therefore, it is likely that a signal peptide is indeed

present. The cleavage probability curve indicates the probability that there is a SPase I cleavage site at the respective amino acid. For OMPC-SALTY the cleavage site is predicted with a probability of 1.0 between positions 21 and 22, in agreement with the prediction by the neural network.

6.3 Transmembrane Proteins

Biological membranes contain integral proteins that have various functions in the cell, such as acting as cell–surface receptors. The integration into the membrane lipid bilayer is accomplished by hydrophobic interactions between the protein and the nonpolar chain structures of the lipids. The polar head groups of the lipids build hydrogen bonds and ionic bonds with the protein. Integral membrane proteins are therefore always amphiphilic molecules that have both hydrophilic and lipophilic regions. These proteins are orientated asymmetrically in the membrane, i.e., some membrane proteins are only exposed on one side of the membrane, whereas others completely penetrate the membrane and are exposed on both the extracellular and intracellular sides, i.e., transmembrane proteins. The hydrophobic transmembrane domains are usually formed by α-helices.

The prediction of transmembrane proteins is of great importance for classification and defining function, as described above for signal peptides. The program TMHMM of the CBS server in Denmark can predict transmembrane domains. TMHMM is based on a HMM that has been trained to detect hydrophobic transmembrane helices. Furthermore, the program also predicts the orientation of the individual domains in the membrane (intracellular or extracellular) and, accordingly, of the whole protein.

The graphical output of such a prediction with TMHMM is shown in Fig. 6.3 for the transmembrane domains of the G protein-coupled receptor (GPCR) 5-hydroxytryptamine-1B receptor (5H1B-SPAEH) of the mole rat *Spalax leucodon ehrenbergi*. Such GPCRs

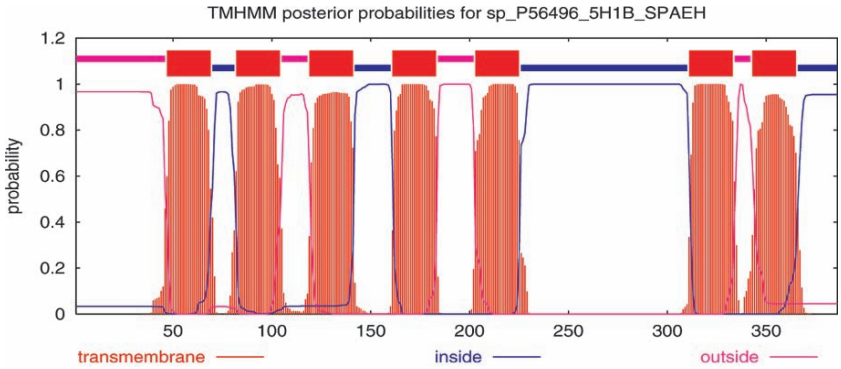


Fig. 6.3. Graphical output of the TMHMM server [tmhmm] at CBS. (Printed with permission of the CBS)

are integral membrane proteins with typical seven transmembrane domains. In the graph, the probability of a transmembrane helix and its intracellular or extracellular localization is plotted along the amino acid sequence. Additionally, in the upper part of the figure a schematic representation of the topology is inserted. The graphical representation of the probabilities also allows the recognition of transmembrane helices of relatively low likelihood.

6.4 Analyses of Protein Structures

As stated above, the prediction of protein three-dimensional structure from the amino acid sequence is currently not feasible and will not be for the foreseeable future. Therefore, experimental methods must be employed to determine protein structures. The two primary approaches are X-ray crystallography and high-resolution nuclear magnetic resonance (NMR) spectroscopy. A third approach using the electron microscope is useful only for very large proteins with molecular masses of several thousand kilodaltons. Overall, in spite of much technological progress, these methods are still very time-consuming and cost-intensive, and the successful resolution of a crystal structure is not guaranteed for every protein.

6.4.1 Protein Modeling

A useful and fast method for structure prediction is homology modeling of proteins based on sequence homology. The approach is based on the fact that related proteins within a protein family (e.g., cysteine proteases of the cathepsin family, see also Figs. 6.5 and 6.6) that have a high degree of amino acid sequence similarity also have similar protein folds. Proteins for which the three-dimensional structure is already known serve as reference proteins or templates. First, the amino acid sequence of the protein to be modeled is compared with the sequence of the reference protein(s) using pairwise or multiple sequence alignments (in case of several reference proteins). For sequences with identities of more than 70%, the modeled structures can be predicted very accurately. However, for sequences with identities of less than 30%, difficulties with the modeling can arise. The sequence identities of structurally conserved regions (SCRs) are frequently above those of less conserved loops and both influence the degree of identity of the complete sequence. Interestingly, areas of little conservation are often found at the protein surface and have a comparatively small influence on the SCRs, which are found inside the protein where most of the active centers are situated.

To identify SCRs in the reference proteins, a structural alignment of the amino acid sequences based on the secondary structure is performed. The sequence to be modeled is then arranged onto the oriented templates and the spatial coordinates of the SCRs are then transferred to the model sequence. The coordinates of the loops are usually taken from similar regions of other protein structures. The spatial orientation of the side chains of individual amino acids in the SCRs is maintained as in the templates. For all nonconserved side chains, the statistically most likely position is taken. The process of homology modeling is completed both by calculations that lead to energy minimization of the model and checking of the structural relevance of the resulting protein model.

6.4.2 The Determination of Protein Structures by High Throughput Methods

The number of experimentally determined protein structures stored in the world's only archive for structures of biological macromolecules, the *Protein Data Bank* (PDB), has grown enormously in the last 10–15 years [pdb] (Bermann et al. 2000, Westbrook et al. 2003). In 1971, PDB contained just seven structures, in 1992 the number was approximately 1,000, and by September 2007 it had grown to 46,051. This remarkable increase in information can be attributed mainly to the *Structural Genomics Initiative* (also called *Structural Proteomics*). This initiative is an international scientific consortium of 25 national initiatives from Japan, North America, and Europe. The aim is nothing less than to structurally solve all of the proteins encoded in the genomes of the most important organisms (archaebacteria, eubacteria, and eukaryotes) [rcsb-sg].

To solve the structures, X-ray structural analyses and NMR spectroscopy will be used in a high-throughput format. To decrease the number of protein structures to be experimentally solved, only representatives of the different protein families will be examined. The underlying idea is that proteins can be divided into protein families and that sequence similarity usually leads to structural similarity. The conclusion is that the number of different structural folds of proteins found in nature has to be limited. One estimate is that between 10,000 and 30,000 protein families exist and these contain approximately 1,000–5,000 different protein folds. Of these, approximately 700 folds are currently known. However, it has to be considered that similar protein structures do not inevitably have similar functions and that different protein structures may also perform similar functions. For example, the cysteine proteases are divided into three structurally different groups based on protein folding patterns: the papain-like proteases, the Picorna virus proteases, and the caspases.

To accomplish the ambitious goal of the *Structural Genomics Initiative*, the strategy is as follows:

1. All known protein sequences are grouped into protein families using bioinformatic methods.
2. Representatives of each protein family are produced in sufficient quantity by molecular biological methods.
3. The protein structures of these representatives are experimentally determined using protein crystallography or NMR spectroscopy.
4. All other protein structures of the respective protein family are generated by homology modeling.

Using this procedure it should be possible to decode almost all protein fold patterns in the foreseeable future, thereby making an important contribution to the functional elucidation of all the known proteomes. The benefits for modern pharmacological research are already substantial and will be invaluable in the future. Soon, for almost every drug-target, it will be possible to perform *structure-based rational drug design* and, thus, improve considerably the development of drugs (Burley and Bonanno 2002).

6.5 Structure-Based Rational Drug Design

From the sequencing of whole genomes and the generation of the corresponding biological information a modern approach to pharmacological research has been established. To initiate the development of a new drug, a drug target (*de facto*, a protein) that plays a key role in the disease must first be identified (see also Chap. 8). After the target's function has been experimentally confirmed (drug target validation), small-molecule chemicals are identified that influence the protein's function in such a way as to alleviate or cure the corresponding disease. The specific inhibition of an enzyme by a chemical inhibitor would be such an example.

The overlapping technologies of computer-assisted bioinformatics and cheminformatics have become essential components

of modern drug discovery efforts. Both strategies are indispensable for the identification and validation of drug targets as well as for the screening and the design of new small-molecules. Also, of special importance is the three-dimensional structure of the drug target to allow for structure-based rational drug design. Finally, the cheminformatic approach of *virtual screening*, which tests the protein target's interaction with chemical entities in large compound libraries, is incorporated into most discovery strategies. Unlike experiments conducted in the laboratory, virtual screening is automated being conducted at the computer, and many chemical substances can be tested for their activity spectrum relatively quickly. A variety of specialized software packages exist for structure-based rational drug design (e.g., docking, de novo design and pharmacophore analyses; Lyne 2002). The best known programs in use are DOCK, developed by Irvin Kuntz at the University of California, San Francisco [dock] (Ewing and Kuntz 1996) and GOLD from Peter Willett of the University of Sheffield [gold] (Jones et al. 1997) as well as the programs of the Flex group developed by Thomas Lengauer and colleagues at the GMD-SCAI (now Fraunhofer-SCAI) in Sankt Augustin, Germany [flexx] (Rarey et al. 1996). The word "docking" is the modern pictorial paraphrase of the lock and key concept postulated in 1894 by Emil Fischer (Fischer 1894). The specificity of the receptor–ligand complex is brought about by the geometric complementarity of both molecules.

Fischer's lock-and-key concept assumes rigid binding partners whose forms are fixed. In nature, however, this is frequently not the case as both the receptor and ligand are flexible and can change their form. The *induced fit hypothesis* takes this flexibility into account (Koshland et al. 1966) and it assumes that the ligand and binding sites adapt to each other's form. In the docking process, therefore, it would be necessary to account for the flexibility of both binding partners. In practice, however, only the flexibility of the ligand is incorporated whereas that of the receptor is neglected for computational reasons. False predictions resulting from this procedure are accepted during the primary virtual screen as its only purpose is to identify potential ligands from

a large library of substances. Single molecules of interest can be subsequently examined with methods that take the flexibility of the receptor into account.

6.5.1 A Docking Example

With DOCK all possible orientations of a ligand and its receptor can be generated. For example, the protein structure of an enzyme with a clearly defined active center can constitute a typical receptor. The structure of the ligand can originate from a database of chemical molecules such as the Available Chemicals Directory [mdl acd].

In the example shown, the cathepsin L-like cysteine protease of the infectious 3rd stage larva of the filarial worm *Brugia pahangi* serves as a receptor. This enzyme is important for the molting and development of this parasite. The protein structure was generated by homology modeling.

1. The first step is the characterization of the active center (site characterization, Fig. 6.4). To do this, the molecular surface

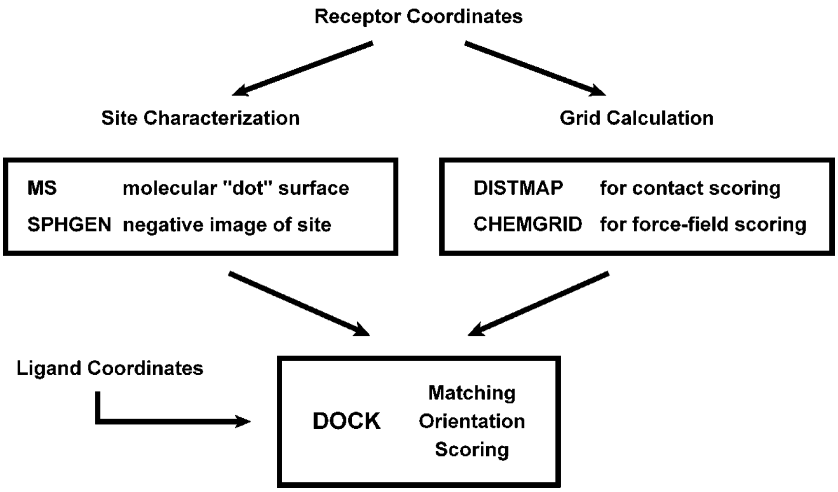


Fig. 6.4. Schematic representation of the mode of operation of the program DOCK [dock]

of the active center is generated first (sub-program *MS*) and converted to a negative image (sub-program *SPHGEN*). Overlapping spheres are then fitted into the active center (Fig. 6.5). The center of the spheres will be eventually replaced with the atoms of the ligand.

2. In the second step, a calculation of physical, chemical, and topological parameters is carried out at each nodal point of a space grid (grid calculation) in order to compute a score, which can be either a contact score or a force field score, respectively.
3. After these calculations, the actual docking can take place. This can be done in two modes; the single DOCK mode and

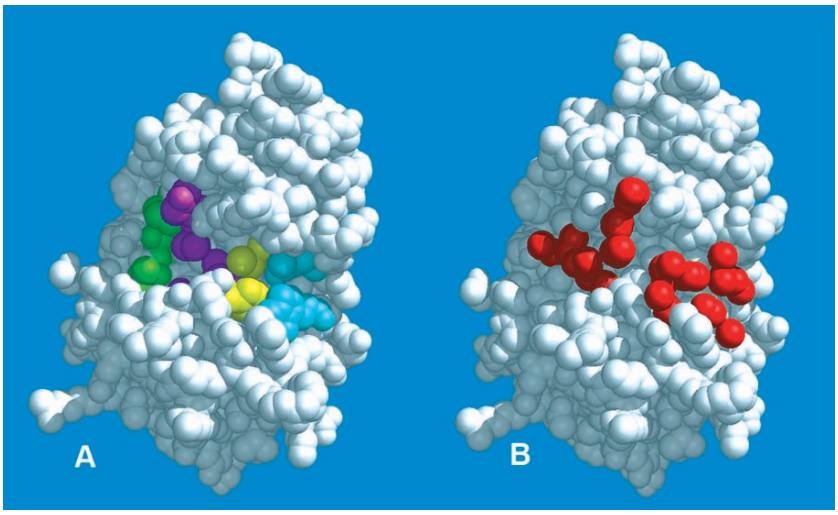


Fig. 6.5. (**a**,**b**) Spheres model of the cathepsin L-like cysteine protease of the filarial, *Brugia pahangi*. The underlying protein structure was generated by homology modeling. (**a**) The most important amino acids in the active site cleft, which is located between the two main domains of the protein, are indicated in color. The active site cysteine (*top*) and histidine (*bottom*) of the catalytic triad are highlighted in yellow. The active site asparagine is hidden in the structure. Important amino acids of the S' subunit are drawn in cyan blue and those of the S subunit in green and pink. (**b**) Graphical representation of the characterization of the catalytic cleft by the program DOCK (subprogram SPHGEN). The centers of the overlapping spheres, where later the atoms of the ligands will lie, are represented in red

the search DOCK mode. In the single mode, DOCK generates all possible orientations of a single ligand in the active center (see Fig. 6.6). In the search mode, large databases of chemical molecules are searched. To perform this, the best orientation of every ligand is first generated and then saved as a relative score. The connections with the highest ranking scores are examined for size, fit, and interaction with the active center. The best compounds can then be experimentally tested for activity in appropriate assays.

For the example of the cysteine protease of *Brugia pahangi*, a chemical database of known cysteine protease inhibitors was searched with DOCK and hydrazide compounds known to

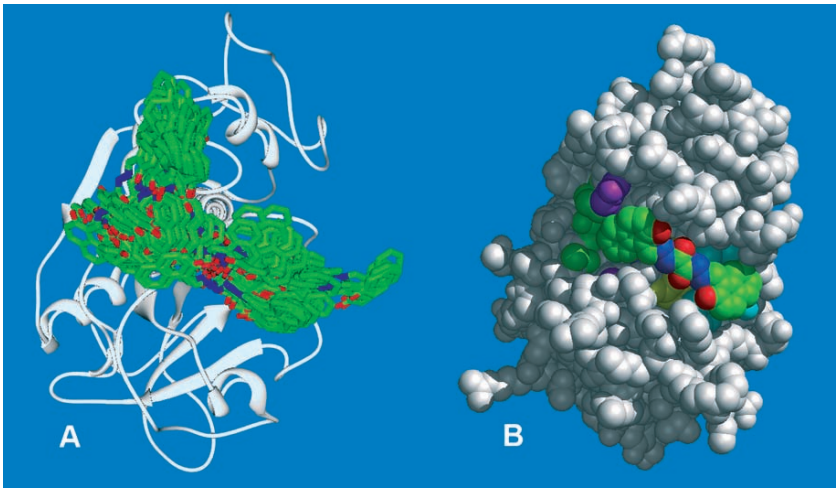


Fig. 6.6. (**a**,**b**) Model of the catalytic cleft of the cathepsin L-like cysteine protease of *Brugia pahangi*, into which a chemical compound was modeled using DOCK. (**a**) The protein is displayed in its secondary structure (*ribbon model*). In single DOCK mode, all possible orientations of a chemical compound (hydrazide) were generated. All overlapping orientations of this single compound are represented in the figure: carbon, *green*; oxygen, *red*; nitrogen, *blue*. (**b**) Based on the analysis of the docking experiment described in (**a**), the most likely orientation of the hydrazide in the catalytic cleft of the cysteine protease is shown. Protein and chemical compounds are represented as spheres. Coloration is similar to those in (**a**) and Fig. 6.5

inhibit the cysteine proteases of the parasites *Trypanosoma cruzi*, *Trypanosoma brucei*, *Leishmania major*, and *Plasmodium falciparum* had very high scores. The binding of the identified hydrazides was then more thoroughly examined in the single DOCK mode to identify the most promising inhibitors (Fig. 6.6). Subsequent experiments with the best predicted inhibitors prevented the development of the infectious 3rd stage larva to the 4th stage larva (Selzer 2003).

6.5.2 Pharmacophore Modeling

Knowledge of the 3D structure of a receptor is essential for docking in a virtual screen. In the absence of either 3D-structure or homology model, virtual screening can be, nonetheless, attempted. As long as some ligands of the receptor are known, the methods of pharmacophore modeling and pharmacophore screening can be used. A pharmacophore is either the spatial arrangement of the ligand's functional groups that are responsible for binding (Böhm et al. 2001) or the system of molecular properties responsible for the pharmacological effect (Nicklaus 2003). To determine a pharmacophore, the known inhibitors or ligands of a receptor are sterically superimposed so that as many similar functional groups as possible are congruent. The common functions are then identified and categorized. Possible categories are hydrogen bond acceptors and donors, aromatic systems, ring systems, hydrophobic areas, hydrophilic areas, etc. (Fig. 6.7). In the end, the spatial arrangement of the individual functions is analyzed and saved in a model called a pharmacophore hypothesis.

Because of its relatively low demand for computation time, pharmacophore screening is frequently used as a filter to reduce the size of the screening library and subsequently the computation time required for virtual screening via docking. In the screening process, every single molecule of the database is spatially fitted into the pharmacophore model and the overlaps of the functional groups of the molecule are compared with the

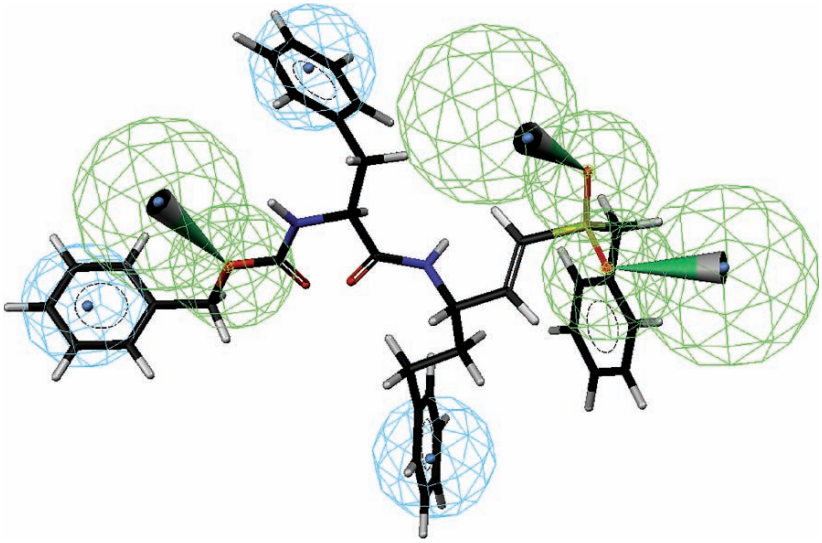


Fig. 6.7. Representation of a pharmacophore model. The functional groups of the pharmacophore are shown as colored spheres in a mesh representation. Blue spheres represent hydrophobic areas and green spheres represent hydrogen bond acceptors. Hydrogen bonds are directional interactions, which are represented by two spheres and a cone. A molecule is fitted into the pharmacophore model that possesses the required pharmacophore

properties required by the model. The quality of the agreement is assessed with a score and all molecules of a certain quality are listed as potential ligands. The program employed determines the exact classification and coding of the hypothesis as well as how the score is assessed. Programs for pharmacophore modeling and pharmacophore screening are part of nearly all cheminformatic software packages. Example programs are Catalyst [catalyst] (Greene et al. 1994), Galahad [galahad] (Clark et al. in preparation), MOE Pharmacophore Modeling [moe], and Phase [phase] (Dixon et al. 2006).

When the 3D structure of a receptor is known, receptor-based pharmacophore modeling can also be used. Here the binding pocket is evaluated for functional groups that are potentially involved in binding to a ligand. The spatial arrangement of these functional groups represents a negative

image of the sought-after pharmacophore, which must then be translated into a positive image for use in pharmacophore screening. In addition, there are also mixtures of ligand-based and receptor-based pharmacophore modeling that incorporate information derived from both the ligand–receptor complex structures and the structures of the ligands themselves (Wolber and Langer 2005).

6.5.3 Successes of Structure-Based Rational Drug Design

A frequently asked question is whether such virtual methods actually lead to drugs. The answer is clearly "yes." There are more examples than can be listed here where virtual technologies have contributed considerably to the development of drugs. One should keep in mind, however, that the development of a drug is a demanding process that involves many different steps. Rational drug design is only the first step on the long road to a marketable drug.

Dorzolamid (trade name Trusopt, (Merck) marketed since 1995), which is used for the treatment of glaucomas, is a carbonic anhydrase inhibitor that originated as the first drug from a program involving structure-based rational design. A second example, Captopril, is a drug that lowers blood pressure and whose lead structure was based on a natural substance that inhibits Angiotensin Converting Enzyme (ACE). Enalapril, another effective ACE inhibitor, is a further development of Captopril. Further examples are the HIV protease inhibitors, Saquinavir and Ritonavir (Norvir), from Roche and Abbott, respectively; the tyrosine kinase inhibitor, Gleevec from Novartis, that is used successfully in leukemia patients and the neuraminidase inhibitors, Tamiflu from Roche and Relenza from GlaxoSmithKline (Böhm et al. 1996).

There are a number of examples where the DOCK program has been used successfully. Particularly impressive have been studies with cysteine proteases. Using DOCK and homology models of the cysteine proteases of *Leishmania major*, small molecules could be identified that block these drug target

enzymes and stop the development of promastigote and amastigote *Leishmania* in cell culture without damage to host cells. In a mouse model of leishmaniasis, progression of the infection could be considerably delayed (Selzer et al. 1997, 1999). Similar results in animal models were achieved for cysteine proteases of *Plasmodium falciparum* (Shenai et al. 2002), *Trypanosoma cruzi* (Engel et al. 1998), and *Schistosoma mansoni* (Abdulla et al. 2007). For *Trypanosoma cruzi,* the success of cysteine proteases inhibitors has set the stage for clinical trials against Chagas' disease (Barr et al. 2005).

6.6 Exercises

1. Explore how many solved protein structures are currently present in the PDB database (http://www.pdb.org/).
2. How many consortia are part of the *Structural Genome Initiative*? How are they distributed worldwide?
3. Find the entry CHER_SALTY/P07801 in the Swissprot database (http://www.expasy.org/sprot/). Does this database record contain information about the tertiary structure of the receptor?
4. View the PDB database record of the receptor from exercise 6.3 (ID 1AF7) and display the molecular structure with one of the visualization programs. Which structures (primary, secondary, tertiary structure) are recognizable?
5. Choose the secondary structure view in the topmost options field on the left and then select two atoms from two arbitrary neighboring leaflets. To do this, click twice on the selected atoms (secondary structure window). What changes are displayed in the primary structure window and what do they mean? Which further representation options does the QuickPDB-Viewer offer?
6. Carry out some secondary structure predictions with the amino acid sequence of the database record CHER_SALTY.

The necessary programs can be found at http://www.expasy.org/tools/#secondary. Compare the predicted secondary structures with the experimentally determined secondary structure.

7. Does CHER_SALTY have a signal peptide? Give reasons for the assumption. Check the presence of a signal sequence (http://www.cbs.dtu.dk/services/SignalP/). Note: *Salmonella typhimurium* is a gram-negative bacterium.
8. Retrieve the database record P41780 from the Swissprot database and repeat exercise 6.7 with this sequence. How does the program SignalP work?
9. The prediction of transmembrane regions works in a very similar way to the determination of signal peptides. The appropriate program can be found at http://www.cbs.dtu.dk/services/. Determine the transmembrane regions of the G-protein coupled receptor (GPCR) with the Swissprot AN Q99527. How many transmembrane regions are detected? Compare this result with a secondary structure prediction for this receptor. Note: in general transmembrane regions are helices.
10. Perform homology modeling with the Swissprot sequence P29619. To do this, go to the SWISS model page of the Expasy server (http://swissmodel.expasy.org/) and follow the hyperlink *First Approach Mode* (on the top left). Fill out the input fields. Start the analysis. Note: all further communication by the server is via e-mail. Therefore, a valid e-mail address must be provided. Save the text file returned by the server with the ending .pdb and open this file with the DeepView SwissPDB viewer. The viewer is available free of charge at the WWW [spdbv]. Tutorials for operating spdbv can be found at the following URL: http://swissmodel.expasy.org/spdbv/text/main.htm. Another free visualization program is RASMOL [rasmol], which, however, offers far less functionality than spdbv.

WWW Links

catalyst: http://www.accelrys.com/products/catalyst/
dock: http://dock.compbio.ucsf.edu/
flexx: http://www.biosolveit.de/FlexX/
galahad: http://www.tripos.com/index.php?family=modules,SimplePage,,,&page=sybyl_galahad
gold: http://www.ccdc.cam.ac.uk/prods/gold/#ref1
mdl acd: http://www.mdli.com/products/experiment/available_chem_dir/index.jsp
moe: http://www.chemcomp.com/software-ph4.htm
pdb: http://www.pdb.org/
phase: http://www.schrodinger.com/ProductDescription.php?m ID=6&sID=16&cID=0
rasmol: http://www.umass.edu/microbio/rasmol/index2.htm
rcsb-sg: http://sg.pdb.org/
signalp: http://www.cbs.dtu.dk/services/SignalP/
spdbv: http://swissmodel.expasy.org/spdbv/
swissmod: http://swissmodel.expasy.org/
tmhmm: http://www.cbs.dtu.dk/services/TMHMM/

Literature

Abdulla MH, Lim KC, Sajid M, McKerrow JH, Caffrey CR (2007) *Schistosomiasis mansoni*: novel chemotherapy using a cysteine protease inhibitor. PLoS Med 4:e14

Barr SC, Warner KL, Kornreic BG, Piscitelli J, Wolfe A, Benet L, McKerrow JH (2005) A cysteine protease inhibitor protects dogs from cardiac damage during infection by *Trypanosoma cruzi*. Antimicrob Agents Chemother 49:5160–5161

Bendtsen J, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: Signalp 3. J Mol Biol 340:783–795

Berman HM, Westbrook J, Feng Z, Gilliland G, et al (2000) The protein data bank. Nucleic Acids Res 28:235–242

Blobel G, Sabatini DD (1971) In: Manson LA (ed) Biomembranes. Plenum, New York, USA, pp. 193–195

Böhm HJ, Klebe G, Kubinyi H (2001) Wirkstoffdesign. Spektrum, Heidelberg, Berlin, Oxford

Burley SK, Bonanno J (2002) Structuring the universe of proteins. Ann Rev Genomics Hum Genet 3:243–262

Clark RD, Abrahamian E, Abrams C, Brandt P, Gustavsson A-L, Homan E, Strizhev A, Wirstam M, Wolohan P Genetic algorithm with linear assignment for hypermolecular alignment of datasets. Manuscript in preparation

Dixon SL, Smondyrev AM, Rao SN (2006) PHASE: a novel approach to pharmacophore modeling and 3D database searching. Chem Biol Drug Des 67:370–372

Engel JC, Doyle PS, Hsieh I, McKerrow JH (1998) Cysteine protease inhibitors cure an experimental *Trypanosoma cruzi* infection. J Exp Med 188:725–734

Ewing TJA, Kuntz ID (1996) Critical evaluation of search algorithms for automated molecular docking and database screening. J Comp Chem 18:1175–1189

Fischer E (1894) Einfluss der Configuration auf die Wirkung der Enzyme. Ber Dtsch Chem Ges 27:3189–3232

Greene J, Kahn S, Savoj H, Sprague P, Teig S (1994) Chemical function queries for 3D database search. J Chem Inf Comput Sci 34:1297–1308

Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748

Koshland Jr. DE, Némethy G, Filmer D (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits. Biochemistry 5:365–385

Lyne PD (2002) Structure-based virtual screening: An overview. DDT 7:1047–1055

Nicklaus MC (2003) Pharmacophore and drug discovery. In: Gasteiger J (Ed) Handbook of Chemoinformatics, Vol 4, Wiley-VCH, Weinheim, Germany, pp. 1687–1711

Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261:470–489

Selzer PM (2003) Structure-based-rational-drug-design: Neue Wege der modernen Wirkstoffentwicklung. In: Lucius R, Hiepe T, Gottstein B (Hrsg) Grundzüge der allgemeinen Parasitologie. Parey, Berlin

Selzer PM, Chen X, Chan VJ, Cheng M et al (1997) *Leishmania major*: Molecular modeling of cysteine proteases and prediction of new nonpeptide inhibitors. Exp Parasitol 87:212–221

Selzer PM, Pingel S, Hsieh I, Ugele B et al (1999) Cysteine protease inhibitors as chemotherapy: Lessons from a parasite target. Proc Natl Acad Sci USA 96:11015–11022

Shenai BR, Semenov AV, Rosenthal PJ (2002) Stage-specific antimalarial activity of cysteine protease inhibitors. Biol Chem 383:843–847

Westbrook J, Feng Z, Chen L, Yang H, Berman HM (2003) The protein data bank and structural genomics. Nucleic Acid Res 31:489–491

Wolber G, Langer T (2005) LigandScout: 3D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J Chem Inf Model 45:160–169

7 Systems Biology: The Functional Analysis of Genomes

7.1 The Identification of the Cellular Functions of Gene Products

The draft sequence of the human genome project was published in 2001 and the number of genes is estimated to be between 20,000 and 25,000. Each human cell, except sperm and eggs, has a complete set of genes. Obviously, however, a blood cell differs in its morphology and physiology from a liver cell. How, therefore, can these differences be explained if all cells have the same genetic material? The answer is simple. Not every gene is transcribed and expressed in every cell. It follows that only those proteins that are required are present in a cell at a given time during the cell's life, i.e., the proteome of a cell or tissue is dependent on the cell type and its current state. It also follows that knowledge of the genome and its genes does not suffice to explain how a gene, a cell, or an organism works. To understand a complex biological system, one must study the regulation and expression of its genes, the function of expressed proteins, the quantitative occurrence of metabolites, and the effects of gene defects on an organism's phenotype. The study of this complexity is frequently termed Systems biology. Systems biology is a relatively new branch of the life sciences, which tries to understand biological organisms as a whole. Its aim is to obtain an integrated picture of all regulatory processes at all levels, from the genome to the proteome and organelles, and from behavior to the biomechanics of the complete organism. Modern methods for the functional analysis of genomes (functional genomics) are

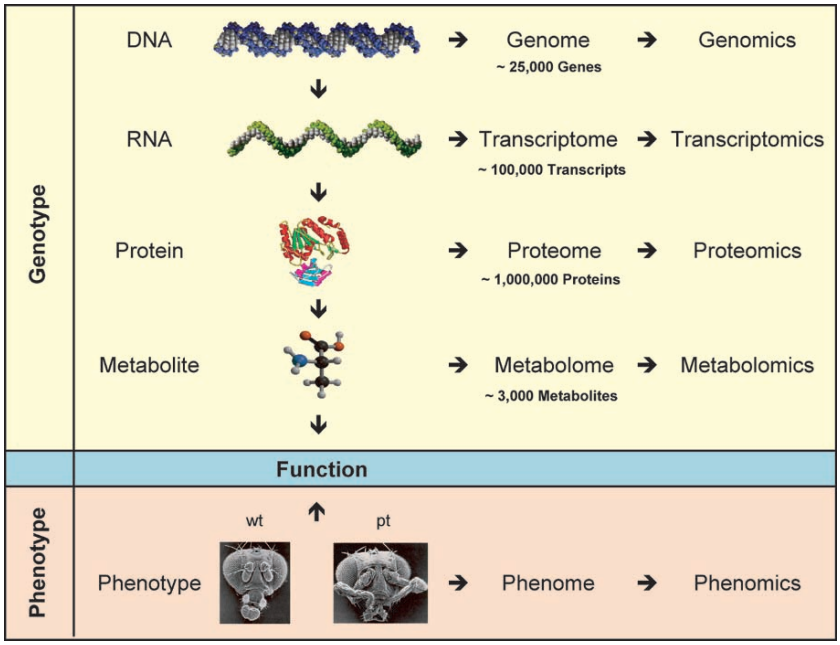


Fig. 7.1. Correlation between genotype and phenotype. From the genome via the transcriptome, proteome, and metabolome to the phenome. The example numbers in the genotype section are taken from *H. sapiens*. The example in the phenotype section is taken from *D. melanogaster* (graphics of DNA, RNA, and metabolite from Lehninger Biochemie, 3rd Edition 2001, printed with permission from Springer Verlag Heidelberg. *D. melanogaster* microscopy images printed with permission from F. Rudolf Turner of Indiana University)

called transcriptomics, proteomics, and metabolomics (Fig. 7.1). These are usually high-throughput procedures that place heavy demands on data-management and -analysis. These approaches are complemented by phenotypic analyses of model organisms and cells in vitro, also in a high-throughput format.

7.1.1 Transcriptomics

Unfortunately, the functions of many proteins based on nucleotide sequences alone are unknown. However, information regarding the gene-regulation and -expression can offer

insight into the functions of gene products in cells, tissues, and organisms. For example, because a gene is expressed exclusively in muscle cells it can be inferred that the gene product is important for the physiology of this cell type. Many techniques exist to analyze gene regulation and expression, e.g., the Northern Blot, which is a method for the detection of mRNA in agarose gels, utilizing nucleic acid hybridization, or reverse transcriptase Polymerase Chain Reaction (RT-PCR), a technique for the amplification of specific nucleotide sequences derived from mRNA. These methods, however, permit only the simultaneous analysis of just a few genes being unsuitable for the efficient analysis of large amounts of data. Therefore, it became necessary to develop high-throughput procedures that permitted a more time-efficient analysis of whole transcriptomes.

7.1.1.1 DNA Microarrays

An example of a high-throughput method is the DNA microarray that is well-suited for the determination of cellular gene expression. Because one can create a profile of every cell based on the genes expressed, this method is also referred to as expression profiling. The solid support material of a DNA microarray can comprise a glass slide on which many thousand nucleic acid spots are placed next to one another (Fig. 7.2a). Alternatively, other materials such as nylon membranes can be used as a support. Each DNA spot includes many copies of a unique single-stranded DNA, allowing its unambiguous assignment to a specific gene.

For the production of DNA microarrays, many techniques are used. In principle, one can distinguish between oligonucleotide arrays and cDNA arrays. For oligonucleotide arrays, short sequences 20–50 nucleotides in length are synthesized directly on the support material (Fig. 7.2b). The procedure involves photolithography, which was originally developed for semiconductor production and is still used in the computer industry. The glass slide is coated with linkers to allow for

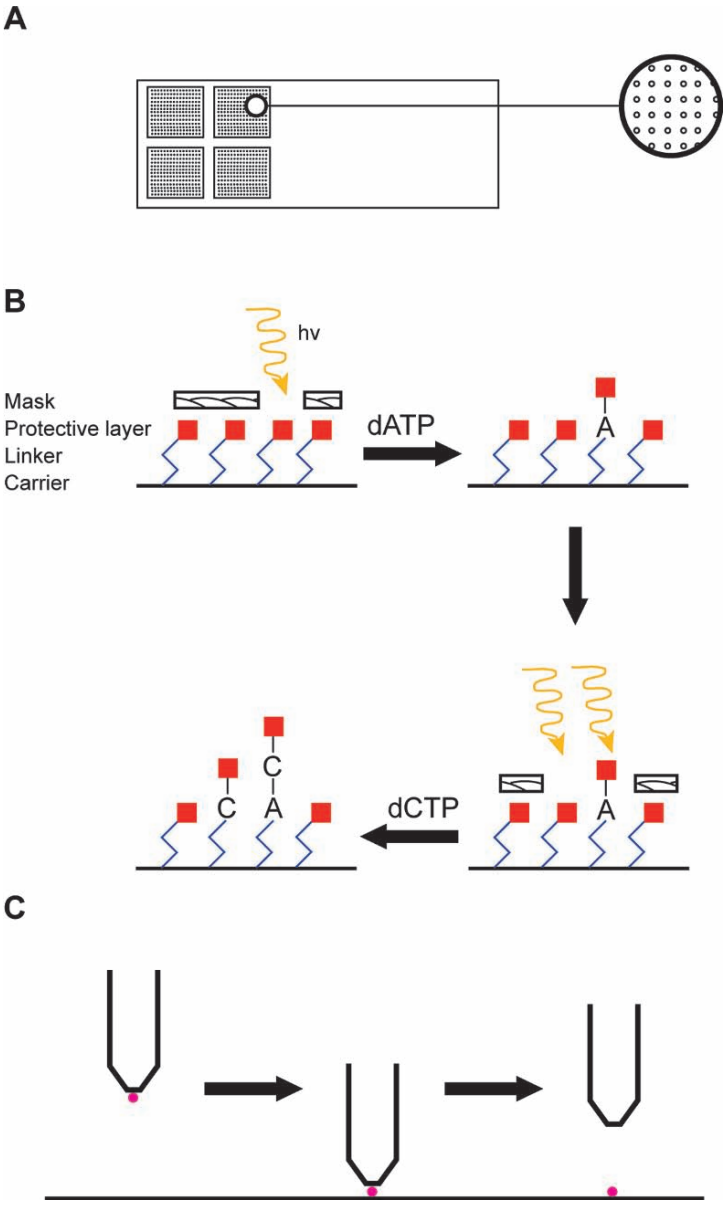


Fig. 7.2. DNA microarrays. (**a**) DNA microarray that consists of several thousand nucleic acid spots and arrayed in high density. (**b**) Schematic illustration of the production of an oligonucleotide array using photolithography. (**c**) cDNA solutions are spotted during array production onto the microarray plates by robots

covalent bond formation with the nucleotides. The linkers are blocked with a photolabile protecting group to prevent the nucleotides from binding nonspecifically. By selectively applying a photo mask the photolabile protecting group is removed, thereby specifically activating selected array sectors. Then the surface of the array is incubated with a nucleotide solution that contains only one specific nucleotide, e.g., dATP. At those positions that had been activated by the photo mask, the nucleotide can now bind covalently to the linker of the support material. The nucleotides themselves are also blocked at the 5′ end with a photolabile protecting group and these must be activated again before the following reaction can occur. Thus, by multiple repetitions and by the application of various masks, an oligonucleotide array of choice can be produced. This technique can produce densely packed microarrays with over 250,000 oligonucleotide spots per cm^2. The American company Affymetrix is the market leader in the field of oligonucleotide arrays, which are also called Genechips or Biochips [affymetrix]. Oligonucleotide arrays are distinguished by an extreme density of high quality spots. Because of the high density, for any given gene, several oligonucleotides can be placed on the array, permitting control of the results and increasing the precision of these arrays. The disadvantage of this technology is that they are usually not produced in-house and must be purchased, often for a considerable price. Moreover, one is dependent on the arrays offered by the manufacturer.

In contrast, for cDNA arrays, considerably longer cDNA probes are placed onto the array support (Fig. 7.2c). First, the cDNAs are amplified to a length of some hundred nucleotides by means of PCR in the laboratory. These are then applied in tiny volumes as DNA spots onto the array support by means of a robot, after which they are immobilized, e.g., by UV light. A number of suppliers of spotting robots are available that use slightly different procedures. One method is microspotting whereby the PCR products are applied with a capillary directly onto the array support. Another procedure is microspraying whereby the cDNA solution is sprayed like from an

ink-jet printer but without the nozzle ever touching the array support. A density of greater than 2,500 DNA spots per cm^2 can be obtained with cDNA arrays. The cDNA array technology is popular in many research laboratories because it is economical. Also, there is flexibility in the choice of the starting material (organism, tissues, or cells). Over the last few years the costs of spot microarrays have declined considerably, however, so has the self-manufacture of cDNA arrays.

The Performance of an Expression Profiling Experiment with cDNA Arrays

Many expression profiling studies compare the gene expression pattern of two different cell populations, e.g., that of healthy cells (cell type A) compared to tumor cells (cell type B; Fig. 7.3). The first step is to isolate total RNA from both cell populations. The mRNA is transcribed into cDNA by the enzyme reverse transcriptase and simultaneously labeled by the incorporation of nucleotides that have been coupled to different fluorescent dyes. Usually, control cDNA (in this case from healthy cells) is labeled with Cy3 dye and sample cDNA (from the cancer cells) with Cy5 dye. Cy3 and Cy5 emit light in the green and red spectrum, respectively. This method is referred to as direct labeling. In contrast, indirect labeling methods exist that are used only if very small quantities of starting material are available. In this case, modified nucleotides are incorporated during cDNA synthesis that bind special dyes with a high affinity.

The labeled cDNA pools are mixed, denatured, and the single-stranded cDNAs then incubated with the DNA microarray. DNA in the pools hybridizes to the complementary single-stranded DNA molecules making up the array. Laser activation of the microarray in the emission frequencies of the dyes followed by quantitative scanning of the emitted light measures the amount of bound cDNA. As a result one gets two pictures, one in the green and one in the red wavelength range. If both are

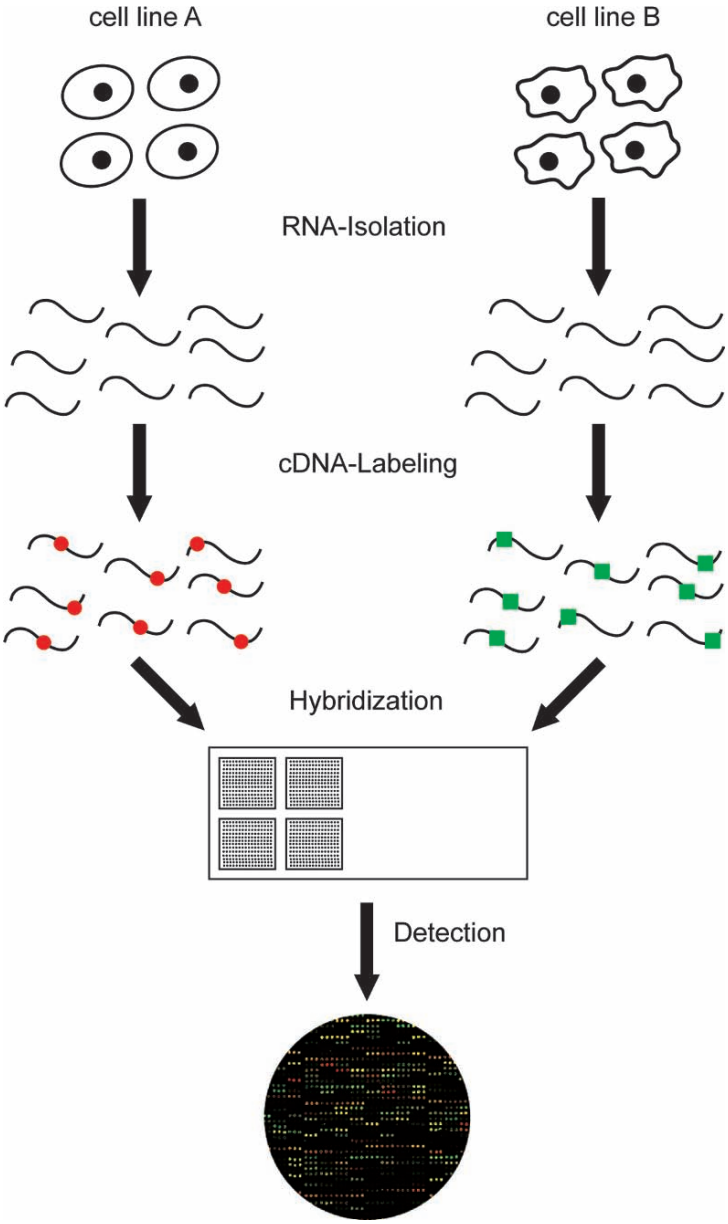


Fig. 7.3. Comparison of gene expression levels in two different cell lines as part of an expression profiling experiment using cDNA microarrays. (Iron chip, printed with permission from M. Muckenthaler, EMBL Heidelberg Germany)

superimposed a merged picture will result with colored spots (Fig. 7.3).

If genes are differentially expressed, i.e., in one cell population there are larger quantities of a specific mRNA, the spots will appear red or green. Spots will appear red if more Cy5-labeled cDNA is bound, i.e., an over-expression of those genes in the cancer cells compared to the controls. Conversely, spots fluoresce green if genes are less expressed in the cancer cells than in controls. Spots appear yellow if red and green fluorescent cDNAs have hybridized to the spotted DNA in equal amounts. This means that the corresponding genes are expressed in the control and cancer cells at equal levels. Spots for which no complementary cDNAs are present in the pools appear black. Thus, it is obvious that the expression of a gene is a relative value between two samples; absolute quantities are not possible with cDNA arrays. This differs from oligonucleotide arrays that allow for an absolute quantification.

The Interpretation of an Expression Profiling Experiment

Although the idea behind microarrays is simple, their execution and analysis of the results are more complex. This is due to the numerous sources of error, including statistical errors, based on stochastic fluctuations that cannot be influenced and systematic errors that lead to measurement deviations. Such systematic errors can be due to incorrect calibration of the instrument or changing environmental conditions (e.g., fluctuations in the temperature or atmospheric humidity) during its operation.

Errors can be minimized with proper experimental design and repetition of the experiments. Samples should be freshly prepared each time to ensure that each experiment is independent. Systematic errors can be minimized by a sophisticated experimental design and control experiments. One control experiment is dye swapping for which cDNAs are labeled with the opposite dye used in the original experiment (reciprocal labeling). Specifically, if in the original experiment the cDNA from the cancer and control cells was labeled with Cy5 and Cy3, respectively, then in the dye swapping control experiment, the cDNA of the cancer

cells should be labeled with Cy3 and that of the control cells with Cy5. Because the same cDNA preparation is used for both the original and dye swapping control experiments and only the label differs, similar results should be obtained. Using the dye swapping control experiment, one can investigate whether an error occurred during labeling of the samples and if so, the extent of the error can be taken into account in the analysis of the results (Churchill 2002).

Interpretation of the data starts with analysis of the displays made by the microarray scanner. The intensities of each spot must be measured in order to convert them to numeric values. This is a complex and comparatively difficult step. The many thousands of spots must be unambiguously identified. To do this, the peripheries of the spots and the fluorescence intensities in the two light channels have to be measured and both then compared with background. Atypical spots that have irregular shapes or contain clumps of red and green color can be marked and ignored for further analysis. All of these processes are usually carried out by the software of the microarray scanner.

Considering the different protocols and the complicated experimental setup (split into numerous individual steps), it is not surprising that microarray data contain systematic errors. Examples are the uneven distribution of the hybridization solution on the array, which leads to the nonhomogeneous staining of some areas of the array, or the different half-lives of the dyes, which can lead to inaccuracies when measuring spot intensities. To compensate for such systematic errors the expression profiling values must be normalized. Normalization is based on the hypothesis that most genes are not differentially expressed in the samples. Normalization not only adjusts the results but also ensures the comparability of experiments carried out on different days or in different laboratories. There are numerous algorithms for normalization and they all have advantages and disadvantages. The choice of algorithm depends on the experience and preference of the researcher (Quackenbush 2001).

It has long been the subject of discussion whether microarray platforms from different suppliers can be compared at all.

In spite of these concerns, several researchers have shown that comparability is actually possible with an adequate experimental setup. However, standardized protocols and adequate controls are essential (Ji and Davis 2006). To oversee quality control, two consortia have been formed with members from academic research groups, the microarray industry and US agencies. The MicroArray Quality Control Project [maqc] establishes standard controls that are aimed at facilitating the comparability of microarray experiments. The External RNA Controls Consortium [ercc] has similar aims. ERCC develops external RNA controls that are added to the experimentally isolated RNA before cDNA synthesis. In this way, the extent to which the results of a microarray experiment agree with defined minimal criteria can be verified.

The next step in data analysis is the identification of genes for which expression is significantly different between the two samples. For simplicity in early microarrays, it was assumed that all those genes for which expression in the samples varied by at least twofold were differentially expressed. Today more complex statistical procedures are used to identify those genes with significant differences in expression levels. These methods have the advantage of identifying genes with low yet significant differences in expression levels. After these statistical analyses, a final number of genes that is differentially expressed is obtained. Importantly, the results should be validated by independent methods, such as RT-PCR analysis.

The determination of the differential expression of individual genes is not the only interesting aspect of microarrays, however, but also the recognition of patterns in gene expression profiles. The idea is that genes that belong to a pathway or react in concert to a given environmental stimulus are coregulated and, therefore, display a similar expression profile. Using cluster analyses all genes with similar expression profiles can be combined into groups or clusters. Figure 7.4 shows such an analysis for 164 bacterial genes that are divisible into 13 clusters. Cluster analyses provide valuable insights into the function of proteins. If genes, for which no function for its products are currently known, are, however, clustered with well-characterized genes,

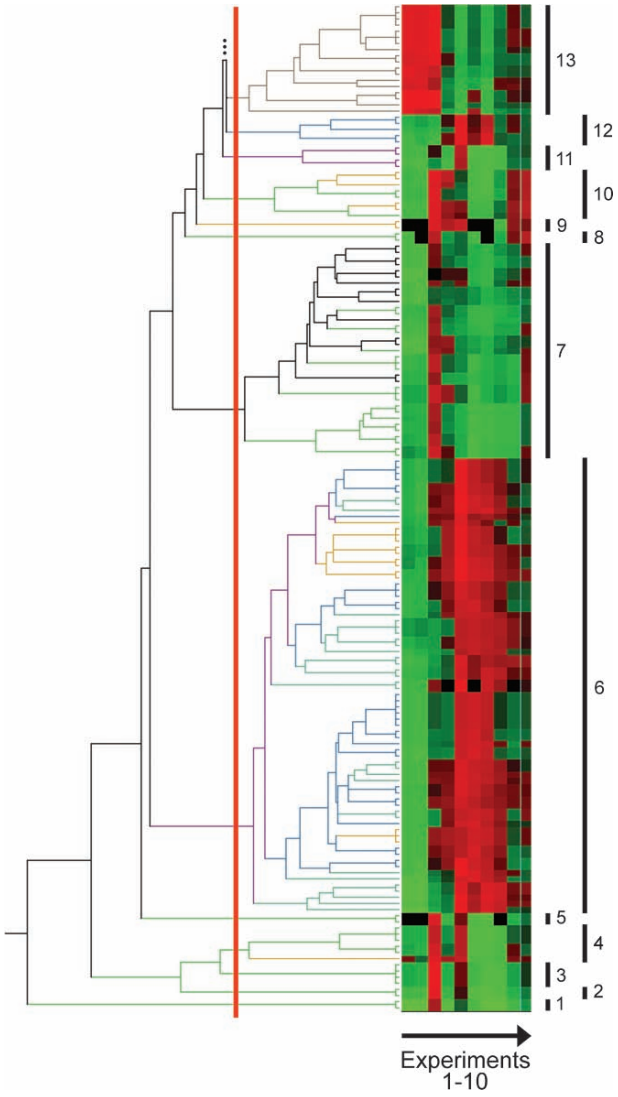


Fig. 7.4. Clustering of genes with similar expression profiles. Expression of 562 bacterial genes was measured in 10 different experiments. The expression profiles were then compared and genes with similar expression patterns grouped into clusters. In this figure, 13 clusters (*black bars*) with overall 164 genes are shown. For instance, cluster 13 contains 18 genes, which are highly expressed in the first three experiments (*red*), but then subsequently decreases (*green*). The red bar represents the threshold selected to define a cluster

then coregulated expression can indicate a similar function or a common pathway to those unknown gene products. The unknown proteins can then be specifically examined for these properties.

Each expression profiling experiment generates an enormous amount of data. One experiment can include dozens of microarrays, which in turn consist of many thousands of spots. Therefore, the resulting several hundred thousand or even millions of measurements must be managed and analyzed using special databases in which the data can be saved and retrieved at any time. Example databases are the Gene Expression Omnibus of NCBI [geo] and the ArrayExpress of EBI [arrayexpress]. In addition to results, one can also find the unprocessed raw data as well as the protocols and the conditions under which the experiments were performed. These data should comply with the *Minimum Information About a Microarray Experiment* [miame] protocol in which the minimum requirements for an explicit interpretation and reliable reproduction of the microarray experiments are defined.

In summary, performing microarray experiments, inclusive of the bioinformatic component, is complex and places high demands on the experimenter. Therefore, a variety of software solutions exist that simplify the analysis of the data. Known commercial programs for the analysis of microarray data are the Rosetta Resolver Gene Expression Analysis System from the company Rosetta Biosoftware [rosetta] and the GeneSpring Software Suite from Agilent Technologies [agilent]. Frequently used software packages that were developed in the academic environment are Bioconductor [bioconductor], the TM4 suite [tm4], and GenePattern [genepattern].

Besides expression profiling there are a variety of other applications for microarrays (Gershon 2005) that have gained increasing importance in tumor medicine, for example. The optimal treatment of a cancer patient is critically dependent on an as precise diagnosis as possible, which, at present, is based on a combination of clinical and histopathological data. In some cases, however, an exact

diagnosis is difficult as tumors frequently have atypical properties. In such cases microarrays can help classify tumors according to their gene expression profiles. An example is acute leukemia. This cancer of leukocytes can be sub-divided into acute lymphatic leukemia (ALL) and acute myeloid leukemia (AML) by using clinical and morphological data for diagnostics. The distinction of these sub-types is essential because each is treated with different chemotherapeutics. An initial study (Golub et al. 1999) examined whether reliable results could be obtained by molecular diagnostics with the help of DNA microarrays compared to classical methods. The gene expression profiles from patients with a known diagnosis were analyzed and then compared with those from patients with an unknown diagnosis. The result demonstrated that the microarray diagnostic tool was reliable. In addition, a patient with a diagnosed atypical acute leukemia was also examined. Here the microarray diagnostic tool showed that this patient's gene expression profile was completely different to those of the other patients. Its profile pointed more to a cancer of muscle tissue than to an acute leukemia. As cytogenetic examinations also disagreed with an acute leukemia diagnosis and favored a muscle tumor, the final diagnosis and chemotherapy were changed accordingly. Thus, the classification of tumors based on DNA microarrays provides validated support to the standard diagnostic techniques.

Another important field for the application of microarray technology is toxicology. Toxicological analyses are designed to identify the damaging consequences of chemical substances on cells. For example, a potential new antibiotic might not only kill the infectious bacterium, but also damage the cells or whole organs of the patient. Therefore, any new potential drug is studied for its toxicological properties by comparing them with existing toxins. These comparisons include gene expression profiling in DNA microarrays. If overlaps in the expression profiles between the known toxins and new compound occur, then the new substance will be classified as being potentially toxic. Analyses of toxicological characteristics using DNA microarrays are also known as toxicogenomics.

7.1.1.2 Serial Analysis of Gene Expression

Like DNA microarray technology, Serial Analysis of Gene Expression [sage] is a high-throughput technology for measuring gene expression. SAGE facilitates the comparison of gene expression in different cells or tissues and, therefore, the identification of differentially expressed genes. SAGE also requires the isolation of total RNA from cells or tissues and the conversion of that mRNA into cDNA using the virally sourced enzyme, reverse transcriptase. The cDNA is not cloned, however, but instead is treated with certain restriction enzymes that cut the DNA at specific sites. This results in the generation of short DNA fragments from each individual cDNA pool with a length between 10 and 14 nucleotides, a tag. Despite being so short, a tag is usually sufficient to unambiguously identify a specific mRNA. The tags are connected into long serial molecules and subsequently cloned into plasmids for sequencing. In a SAGE experiment, the frequency with which a tag appears in a sample is used as a measure of the magnitude of expression of the corresponding mRNA. For example, if the gene tag is found five times in a sample from healthy cells but twenty times in a sample from cancer cells then one assumes that this gene is approximately fourfold over-expressed in the cancer cells. SAGE results can be saved in the database SAGEmap at NCBI. There, the information about each tag can be found, including its DNA sequence, frequency in tissues or cells, and the specific transcript from which the tag was derived [sagemap].

The great advantage of SAGE over DNA microarrays is that all mRNA transcripts of a cell can be analyzed including unknown transcripts (e.g., new splice variants), i.e., no a priori knowledge of mRNA sequences is required, as is the case in DNA microarrays. Another advantage of SAGE is its steady reproducibility between experiments. One disadvantage of SAGE compared to DNA microarrays is the limited capacity of SAGE experiments (approx. 50,000 tags per SAGE experiment). This could change in the future as systems are being developed to analyze more than one million tags (Bonetta 2006).

7.1.2 Proteomics

The quantification of mRNA by DNA microarrays or SAGE provides important information about potential cellular functions of gene products. Measuring mRNA alone, however, is not sufficient to completely and precisely describe complex biological systems. Ultimately, cellular activities like metabolic processes are mediated by proteins of the proteome and not by genes of the genome or mRNA of the transcriptome. Analogous to the DNA microarray technology, therefore, high-throughput procedures have been developed for the parallel functional analysis of proteins, i.e., proteomics. Proteomics is classified into two categories: classical or quantitative proteomics and functional proteomics. Classical proteomics deals with the identification and quantification of proteins in cell lysates, whereas the aim of functional proteomics is the determination of protein function.

7.1.2.1 Classical proteomics

Classical proteomics is similar to expression profiling, which is why it is also termed protein profiling. Both technologies permit the molecular fingerprinting of a cell based on the genes expressed at the mRNA or protein level. By comparing two or several such fingerprints, differentially expressed genes and proteins can be identified. Both technologies have advantages and disadvantages. Protein profiling detects the proteins that ultimately perform cellular functions. Also, the quantitative modifications in a protein's composition based on either new synthesis or breakdown (protein turnover) can be measured. Other advantages of protein profiling are the ability to verify post-translational modifications (e.g., phosphorylation and glycosylation) and to determine the protein composition of cellular compartments (e.g., of mitochondrion or nucleus). One disadvantage, however, is that not all proteins are soluble, particularly transmembrane proteins. A second limitation is the limit of detection such that weakly expressed proteins

can be missed. In contrast, complete genomes can be analyzed in a few DNA microarray experiments, yet the assumption that the quantity of mRNA stochastically reflects with that of the protein is often not the case. Moreover, the quantity of mRNA cannot provide information about protein turnover. Therefore, where possible, both expression and protein profiling should be performed as complementary techniques.

A common procedure for protein profiling combines two-dimensional gel electrophoresis (2D gel electrophoresis or 2DE) with mass spectroscopy. In 2DE, cell proteins are first separated through a separating matrix (e.g., a polyacrylamide gel) according to their individual charges generated by an electrical field. Separation is, therefore, possible due to two inherent properties of proteins, charge and mass. Charge depends on amino acid composition, e.g., cytochrome c contains many basic amino acids and is, therefore, positively charged at neutral pH. The net charge a protein carries depends on the pH of its surroundings and the pH at which both the positive and negative charges of a protein are equal (i.e., a net charge of zero) is called the isoelectric point (pI). Accordingly, a protein will not migrate in an electrical field when its pI equals the pH of its surroundings. Because each protein has a characteristic pI value, one can separate a protein mixture in a pH gradient using an electrical field. This method, called isoelectric focusing, is used in 2DE as the first dimension for the separation of proteins. In the second dimension, proteins are separated only according to their molecular mass. Peptides with a low molecular mass move faster than larger proteins through the pores of the polyacrylamide gel. In this way, up to 10,000 different proteins can be separated in high-resolution 2D gels. After separation, proteins are made visible using differing staining procedures (e.g., silver staining or staining with fluorescent dyes; Fig. 7.5). The gels are then digitized and evaluated with bioinformatic methods. Programs such as Melanie [melanie] at the Expasy proteomics server allow the automatic detection and precise quantification of protein spots. Furthermore, Melanie allows the comparison of several 2D gels. Co-localized protein spots are identified and their quantitative differences measured based on spot intensity.

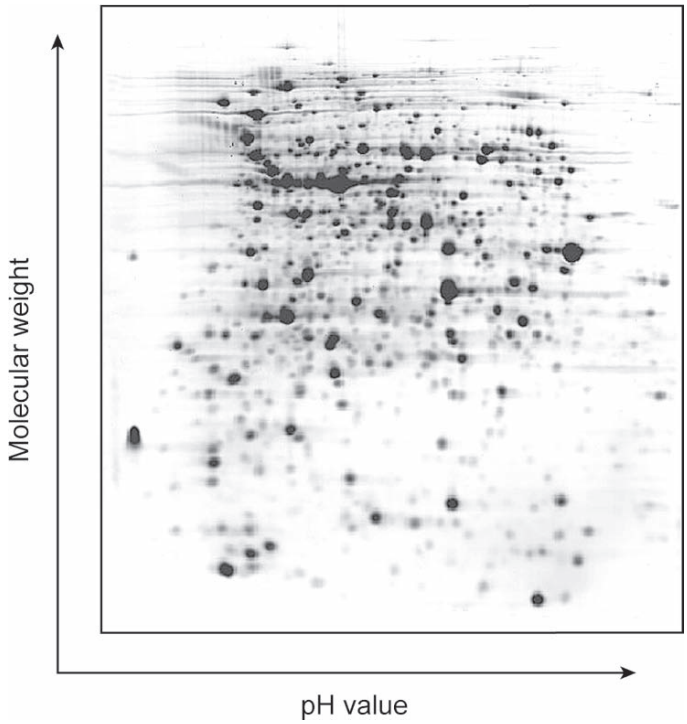


Fig. 7.5. Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE or 2DE). A protein lysate of a bacterium was separated along a pH-gradient (pH 3–10) in the first dimension and by molecular mass in the second dimension. The resolved proteins were then visualized by silver staining

Melanie also contains algorithms for normalization and statistical analyses with which the significance of the results can be judged to identify differentially expressed proteins.

The bioinformatic evaluation of 2D gels yields a list of expressed proteins for which only the isoelectric points and molecular masses are known. While the identity of some of these proteins can be determined using this information, for most proteins a partial determination of amino acid sequence is required. This sequence is compared with a protein database and, if the protein already exists, the identity can be confirmed.

Different techniques are used for the determination of amino acid sequence. A reliable method is sequencing by Edman degradation, for which, however, relatively large amounts of protein

are required. A second method is mass spectroscopy for which proteins are ionized and the charged particles analyzed in a mass spectrometer.

One technology for the ionization of proteins is Matrix Assisted Laser Desorption/Ionization (MALDI). MALDI is sensitive enough to require only picomolar amounts of protein. Stained protein spots are excised from the polyacrylamide gel

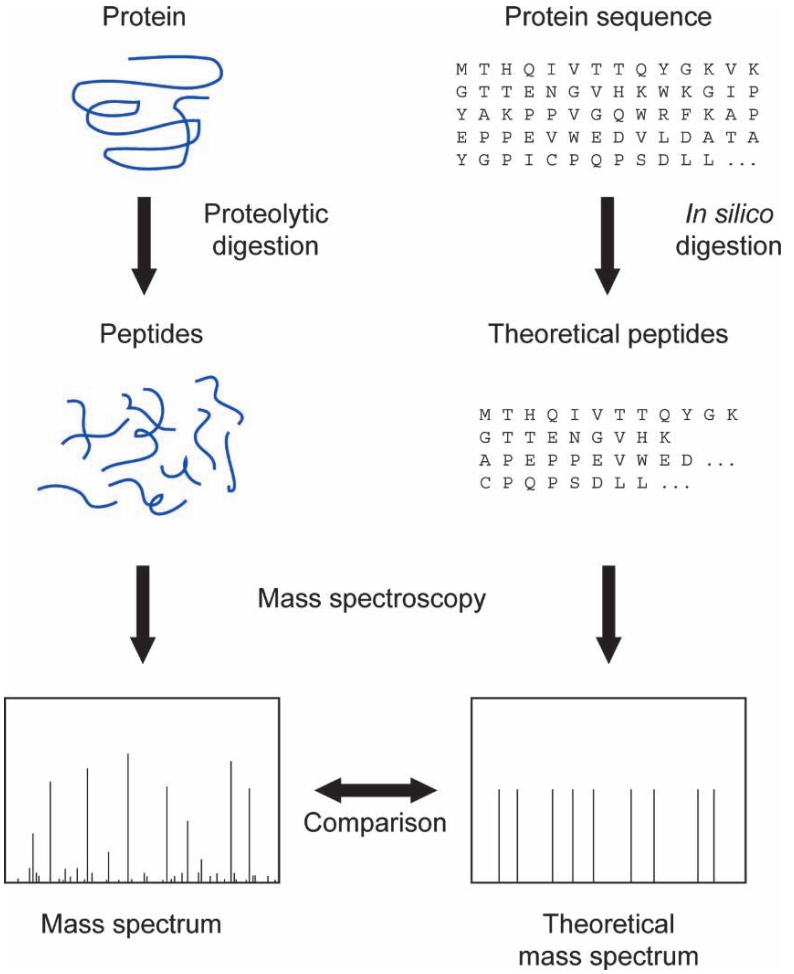


Fig. 7.6. Identification of proteins by cross-referencing data from mass spectroscopy experiments and mass spectra that were theoretically computed

and incubated with a protease (e.g., trypsin), which hydrolyzes each protein into a specific peptide pattern. The peptides are extracted from the gel and analyzed after MALDI in a Time-of-Flight (TOF) spectrometer. For each peptide, a specific peptide mass spectrum is generated (Fig. 7.6). At the same time, all proteins in a database are digested into peptides in silico based on the same cleavage specificity of trypsin and the theoretical mass spectra of these fragments are calculated. The experimentally determined MALDI mass spectra are then compared with the theoretical spectra and those mass spectra that are identical are selected. Because a MALDI mass spectrum can result from more than one protein, definitive identification of the protein requires the spectra of several peptides. Thus, if several of the mass spectra determined by MALDI and those determined theoretically agree, then the experimentally analyzed protein is the same protein as that identified in the database.

An alternative protein ionization technique is electrospray ionization (ESI). ESI is sensitive and particularly suited to the analysis of high molecular mass compounds like proteins. The advantage of ESI over MALDI is that one can couple ESI to a liquid chromatographic system (LC). The latter can fractionate protein solutions and, at least with samples of moderate complexity (i.e., with a limited number of different proteins), replace the laborious 2DE. By the direct coupling of an LC system to the mass spectrometer (LC/MS), protein identification is accelerated. Disadvantages of ESI are its strong sensitivity to alkali contamination and the somewhat more problematic assignment of distinct masses.

An innovative development in the field of mass spectroscopy is tandem mass spectroscopy (MS/MS). Here, two mass spectrum analyzers are run consecutively, which greatly improves the sensitivity and selectivity of the system. For example, protein samples are ionized by ESI. Then, in the first spectrometer, ions of a given mass are selected and excited for further fragmentation and detailed analysis is performed in the second spectrometer. Because of the combination of analyzers, therefore, an initial chromatographic separation may be unnecessary. In practice, however, these systems are frequently coupled as part

of an LC-MS/MS system or even of a 2D LC-MS/MS system, which further increases the sensitivity and selectivity.

7.1.2.2 Functional Proteomics

The aim of functional proteomics is to elucidate the function of proteins, e.g., identifying protein–protein interactions. Many cellular processes are governed by such interactions and their identification is an important topic for the understanding of protein function overall. Examples are the allosteric inhibition of enzymes, the regulation of signal transduction cascades by protein kinases, and the assembly of structural protein complexes to form the cytoskeleton. Numerous methods allow the analysis of such interactions, such as affinity chromatography and the yeast two-hybrid system. Their applications, however, are usually confined to studying the interactions of a limited number of proteins. It is only in the last few years that these methods have advanced to the point where they can be used to dissect protein–protein interactions in complete proteomes (Fig. 7.7). In this context, the term interactome of an organism applies and this type of research is also called interactomics.

To detect the interaction of two fusion proteins, the yeast two-hybrid system is commonly used (Fig. 7.8). Protein X, for which an interacting protein(s) is sought, is coupled to the DNA-binding domain of a transcription factor. Protein X is then mixed with the expression products translated from a cDNA library that have been fused to the cognate transcription factor's activating domain. Neither X nor Y alone are capable of forming a complete and functional transcription factor. Only when proteins X and Y interact are both domains brought together and a functional transcription factor results that can activate the transcription of reporter genes. Their expression can be measured by activity tests and is thus indicative of an interaction between proteins X and Y. Using yeast two-hybrid, a large proportion of the human proteome was tested for protein–protein interactions. Approximately 2,800

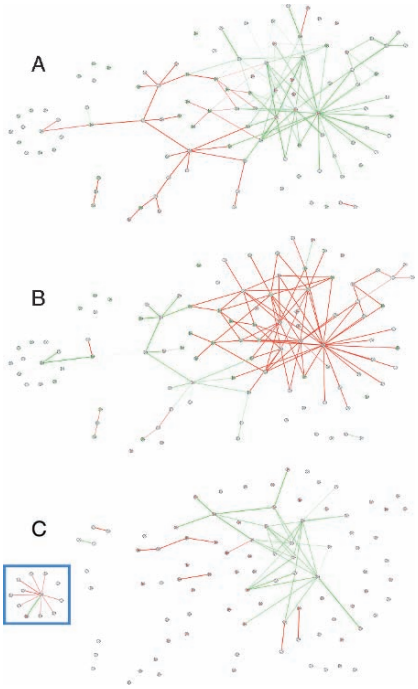


Fig. 7.7. Effects of pharmaceuticals on molecular networks. (**a**) Molecular network comprising proteins and lipids in healthy patients. Most connections are labeled in green, indicating a negative correlation between analytes. (**b**) Molecular network in disease groups. The majority of correlations are labeled in red representing a change from a healthy to a diseased condition. (**c**) Molecular network in drug-treated patients. Many of the green links seen in healthy patients have been restored. However, in a second pathway new network links appear (blue box). This is due to off-target effects of the drug treatment (printed with permission from BG Medicine Inc. USA)

protein–protein interactions for 1,549 proteins were identified (Rual et al. 2005).

Tandem affinity purification (TAP) is another technology that is suitable for the analysis of multi-protein complexes. This technique is based on the combination of affinity chromatography and mass spectroscopy. The target gene is modified so that the gene product is labeled with a short peptide sequence or tag that facilitates the isolation of the labeled protein from

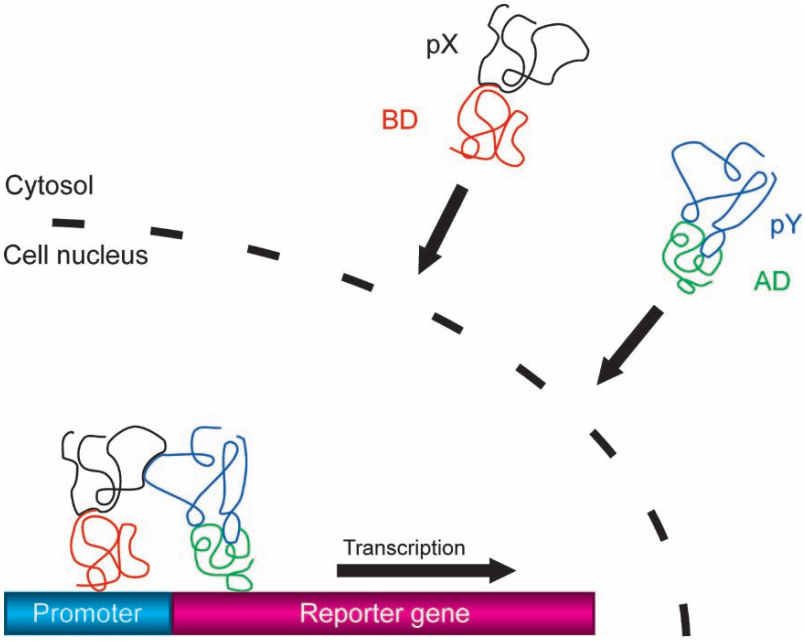


Fig. 7.8. Identification of protein–protein interactions using the yeast two-hybrid system. Transcription of a reporter gene can only be activated when a fusion protein comprising the DNA binding domain of a transcription factor (BD) and a random protein X (pX) interact with a second fusion protein containing the cognate transcription factor's activating domain (AD) and a random protein Y (pY)

the protein lysate. The procedure is gentle and simultaneously co-purifies those interacting cellular proteins that were bound to the labeled protein. The isolated multi-protein complex is then separated by gel electrophoresis and the individual components are analyzed by mass spectroscopy. In this way, 232 different multi-protein complexes could be identified in the yeast *Saccharomyces cerevisiae*. Some of the multi-protein complexes consist of over 40 individual components. Furthermore, a potential function could be assigned to some unknown proteins based on their interaction with proteins with well-characterized known cellular function(s) (Gavin et al. 2002).

As is the case for every high-throughput experiment, the large quantity of data generated by interactomics requires the development of special databases. Example interactome databases are the Biomolecular Interaction Network Database [bind] and IntAct [intact]. To ensure that all relevant data of an experiment are included in the databases, the *minimal information required for reporting a molecular interaction experiment* (MIMIx) protocol regulates the minimal requirements for the storage of protein–protein interaction data.

7.1.2.3 Protein Arrays

An alternative method for the analysis of proteomes is based on protein array technology (Eisenstein 2006). Protein arrays are built similar to DNA microarrays. Onto a coated glass plate or membrane, spots are applied at a high density. These spots consist of reagents that have a high protein-binding affinity (e.g., antibodies). Protein arrays are also suitable for the generation of a protein profile, whereby three different variants of protein arrays are distinguished:

- One variant is the sandwich assay (Fig. 7.9a), whereby antibodies are directly coupled to the protein arrays. The arrays are then incubated with a protein lysate. If a protein is present in the lysate for which an antibody has been spotted onto the array then the protein will bind to that antibody. The detection of this binding is carried out with a secondary antibody that is directed against the same protein but to a different epitope than the primary antibody. The secondary antibody is labeled (e.g., with an enzyme that catalyzes a visually detectable reaction) to allow for detection and quantitation of the binding.
- The second variant is the antigen capture assay (Fig. 7.9b). As above, the primary antibodies are directly bound to the matrix. It differs from the sandwich assay in that proteins in the lysate are already labeled (e.g., with fluorescent dyes). With this assay two cell lysates can be compared by labeling the proteins of the respective lysates with different dyes. Both lysates are

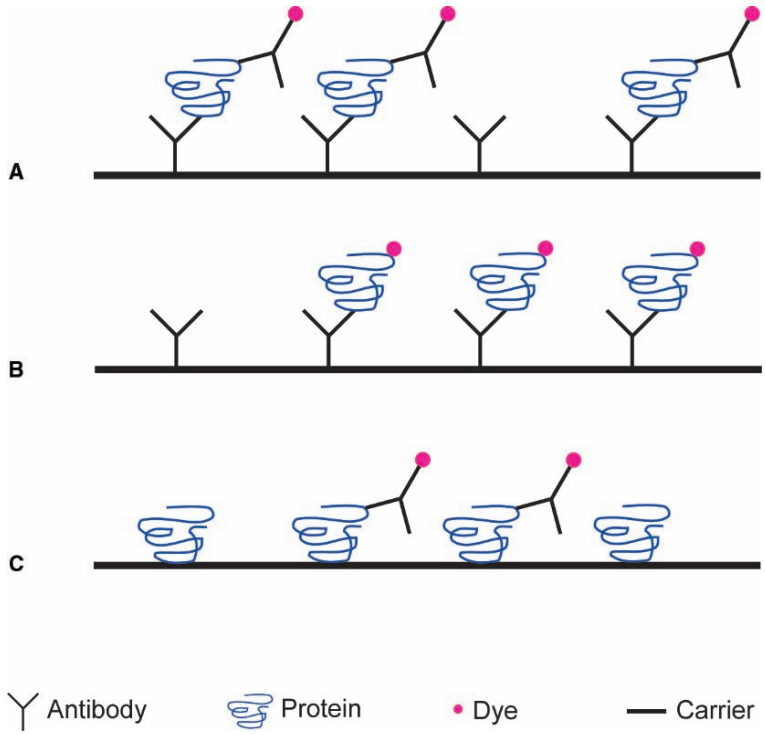


Fig. 7.9. Protein arrays. (**a**) In a sandwich assay, antibodies that are attached to array plates selectively bind to antigenic proteins after incubation with a protein lysate. Detection of the captured protein is performed with a secondary antibody that binds to a different antigenic site(s) on the protein. (**b**) In an antigen capture assay, the antigenic proteins are labeled directly prior to incubation with the protein array, thus dispensing with the need for a labeled secondary antibody for detection. (**c**) In direct or reverse-phase arrays, proteins are directly coupled to array plates and detected with labeled antibodies

mixed and incubated with the protein array. Depending on the amount of labeled bound protein, one or other lysate will contain more labeled protein. The basic concept of this procedure is analogous to that of an expression profiling experiment.

- In the third variant, the direct or reverse-phase assay, proteins and not antibodies are coupled to the protein arrays followed by the use of labeled antibodies. This way, proteins that interact with the antibodies are identified (Fig. 7.9c).

Protein arrays can also identify protein–protein interactions, as described in the previous section. Unlike the yeast two-hybrid system and Tandem Affinity Purification, however, it is an in vitro method. Protein interactions are analyzed outside the cell and under in vitro conditions – conditions that may lead to interactions that do not occur in vivo. On the other hand, protein arrays have the advantage that they can be produced in large quantities, allowing for multiple repetitions of experiments and modification of the conditions (pH, temperature, protein concentration, availability of ions and cofactors). Moreover, with such arrays, thousands of proteins and even whole proteomes can be analyzed at the same time. For example, proteins from *Saccharomyces cerevisiae* were sought for that could interact with the calcium-binding protein calmodulin. The array contained 5,800 of the possible 6,200 yeast proteins (Zhu et al. 2001). Thirty-nine proteins were identified as potential interacting partners, of which just six had already been described as calmodulin-binding proteins. The example highlights how protein arrays can define novel protein–protein interactions. Furthermore, protein arrays can also aid the detection of protein interactions with glycosides, lipids, nucleic acids, or other general ligands.

7.1.3 Metabolomics

If one compares tumor cells with normal cells then it is striking that in the former metabolic enzymes are frequently overexpressed. This should not be surprising, however, as cancer cells grow faster and thus have a greater need for metabolites. The notion is, therefore, that by quantifying cellular metabolites cells can be profiled in a similar way as with either microarray or proteomic techniques. The total metabolite pool of a cell is called the metabolome and the research field dealing with metabolic profiling is termed metabolomics (Fig. 7.1).

Metabolomics is a relatively new research area, although in 1970 Robinson and Pauling had already described experiments

to identify and quantify the metabolites in human urine. Today, over 2,000 metabolites of man are known and the identification of new metabolites is the aim of many research groups. The Human Metabolome Project [hmp] has already discovered approximately 1,000 new metabolites, the identities of which are deposited in the Human Metabolite Database.

Despite the metabolome being rather concise compared to the genome, transcriptome, or proteome, the technical demands required for metabolomics are particularly high. The reason for this lies in the extreme diversity of the various physical properties of the metabolites to be measured. Some metabolites are relatively small and hydrophilic (e.g., vitamin C), and others have a much higher mass and are nonpolar (e.g., cholesterol esters, Fig. 7.10). At present, no single technology exists to identify and quantify all the metabolites simultaneously. However, the technological progress over the last few years has resulted in methods that can measure a small number of metabolites in parallel. Usually, the relative quantities of the metabolites in two different samples are compared to each other, similar to the approach with DNA microarrays. In addition, more sensitive equipment and adequate standards also allow the absolute quantification of metabolites in a single assay.

Two methods are primarily employed to measure metabolites. Sensitive nuclear magnetic resonance (NMR) spectroscopy can generate physical, chemical, electronic and, particularly, structural data from molecules. The method more frequently employed, however, is mass spectroscopy (see Sect. 7.1.2.1). Usually a chromatographic step (e.g., gas chromatography (GC)) to separate the metabolites is done first. Then, with the help of highly specialized equipment, more than 4,000 raw data peaks can be measured, corresponding to approximately 1,800 metabolite peaks (Kell 2006).

Metabolomic experiments generate huge data amounts, which must be analyzed and converted to biologically useful. Many researchers have expressed the opinion that metabolomics describe the functions within a cell better than genomics, transcriptomics, or proteomics. They justify their opinion by saying that genes encode transcripts; transcripts in turn encode proteins and these are eventually responsible for the production of

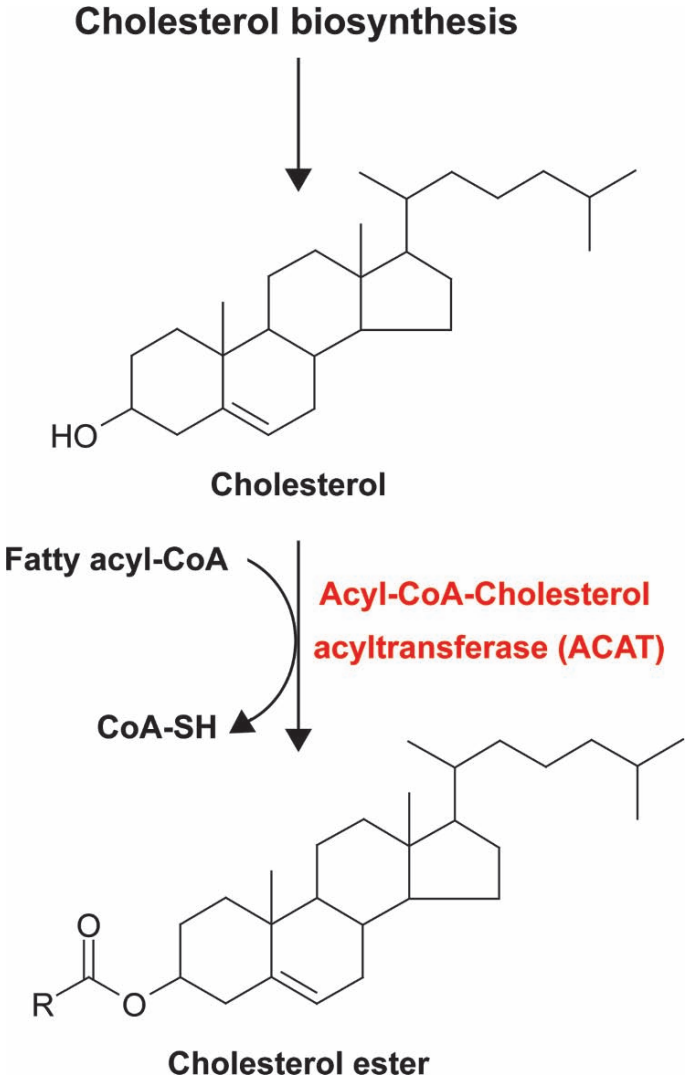


Fig. 7.10. Cholesterol ester biosynthesis catalyzed by Acyl-CoA-cholesterol acyltransferase

metabolites. Therefore, metabolites are at the end of the information chain and, thus, closely connected to their function. A further argument is the amplification of information. It has been experimentally determined that even small changes in the concentration of a few enzymes can lead to significant changes

in the concentration of many metabolites (Raamsdonk et al. 2001). Reasons are that synthesis and turnover of metabolites in general is catalyzed by several enzymes and one metabolite can be involved in many different reactions. In this respect one also refers to a metabolic network (see Figs. 8.3 and 8.4).

A strength of metabolomics is the possibility to construct models of quantitative changes in the metabolome due to its networked structure. Indeed, many models have already been devised, particularly for well-studied organisms such as the baker's yeast, *Saccharomyces cerevisiae*. For example, with the help of a metabolic model that represents 750 genes and 1,149 reactions in baker's yeast, 4,154 growth phenotypes were predicted. A comparison with experimental results showed that the model had, in fact, correctly predicted 83% of the phenotypes (Duarte et al. 2004). The generation of such metabolic models overlaps to some extent with another area, namely *Systems Biology,* which is described in more detail in Sect. 7.2.

7.1.4 Phenomics

The phenotype or physical appearance is the sum of all extrinsic visible features of an individual (see Fig. 7.11). It refers to both morphological and physiological properties. Consequently, the visible and measurable properties of an organism or cell that are based on interactions of the genotype with the environment constitute the phenotype (see Sect. 5.1.2). By this definition, therefore, metabolomics, being measurable, are also a manifestation of a phenotype that is based on interactions of the genotype with the environment.

Many methods exist in the context of the functional genomics that define protein function based on phenotypes. This research area is also called phenomics if it is carried out in a high-throughput format. Initially, forward genetic screens were used in which genomes were randomly mutated, the resulting phenotypes recorded and the genes responsible for the modi-

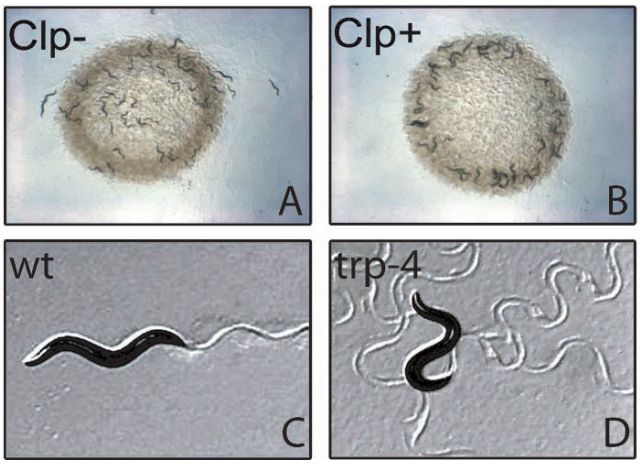


Fig. 7.11. Phenotypes in the roundworm *Caenorhabditis elegans*. (**a**) Most strains are solitary feeders and do not show a Clumping phenotype (Clp−). (**b**) Some strains aggregate on the border, recognized as the Clumping phenotype (Clp+). The phenotype is caused by a naturally occurring genetic polymorphism in a single gene. (**c**) Phenotype of a moving wild-type worm (wt). (**d**) Phenotype of a *trp-4* knock-out worm with an abnormal body posture. The ion channel mutants have a greater frequency of body movement with more pronounced flexing. ((**a**,**b**) Printed with permission from Marie-Anne Félix, Institut Jacques Monod, France. (**c**,**d**) Printed with permission from X. Z. Shawn Xu, University of Michigan Medical School, USA)

fied phenotype identified. Using this approach many thousands of genes could be identified and characterized.

The arrival of sequenced whole genomes offered alternative approaches for performing genetic screens for those genes without an ascribed function(s). The strategy that links a distinct gene with its function(s) is called reverse genetics. Often knockout experiments are carried out whereby genes are selectively mutated ("switched off") so that no functional protein is encoded. The consequence can be an altered phenotype whose properties can then be accurately documented. If a gene encodes an essential protein, the resulting phenotype may be lethal, i.e., the cell or organism dies. Such knockout experiments are usually performed mainly in cell lines or in model organisms such as the fruit fly *Drosophila melanogaster* [genedisruptionproject], the

nematode *Caenorhabditis elegans* [geneknockoutconsortium], or the house mouse *Mus musculus* [geneknocksmouse]. The disadvantage of this method is the complicated and time-consuming experimental approaches required, reflected by the fact that for only a few organisms (e.g., baker's yeast) are complete and comprehensive genome-wide knockout data available. Analogous to knockouts are "knock-in" experiments to elucidate the function of gene products. In this case, genes are transfected into cells or organisms and then observed to determine whether they cause phenotypic changes [geneknocksmouse]. The knock-in strategy is frequently used as additional proof of the protein's function. If the phenotypic change of a prior knockout can be reversed by a knock-in experiment, then there is little doubt as to the protein's function. For example, a bacterium, in which a specific flagellar protein has been knocked out, is rendered immotile. If the same bacterial clone has the gene restored in a knock-in experiment and subsequently recovers motility, then the evidence is solid that the protein is essential for proper flagellum function. Unfortunately, as for knock-out strategies, knock-in technology is laborious and not amenable to high throughput.

The discovery and experimental application of RNA interference (RNAi) has resulted in a revolution for reverse genetic screening. RNAi is an evolutionarily conserved mechanism that involves repression of gene expression by double-stranded RNA (dsRNA; Vanhecke and Janitz 2005). After gaining access to the cytoplasm of the cell, dsRNA molecules are first cut into lengths of 21–23 nucleotides, termed small interfering RNAs (siRNAs), by the enzyme Dicer (Fig. 7.12). The single-stranded siRNA is then loaded into the enzyme complex called the RNA-induced silencing complex (RISC). The activated enzyme complex, guided by the siRNA strand, binds specifically to the complementary mRNA, which is cut by the endonuclease activity of the RISC complex. In this way, the expression of the target gene is specifically blocked preventing translation of the cognate protein. Because transcription blockade by RNAi may not always be complete, the term gene knock-down applies.

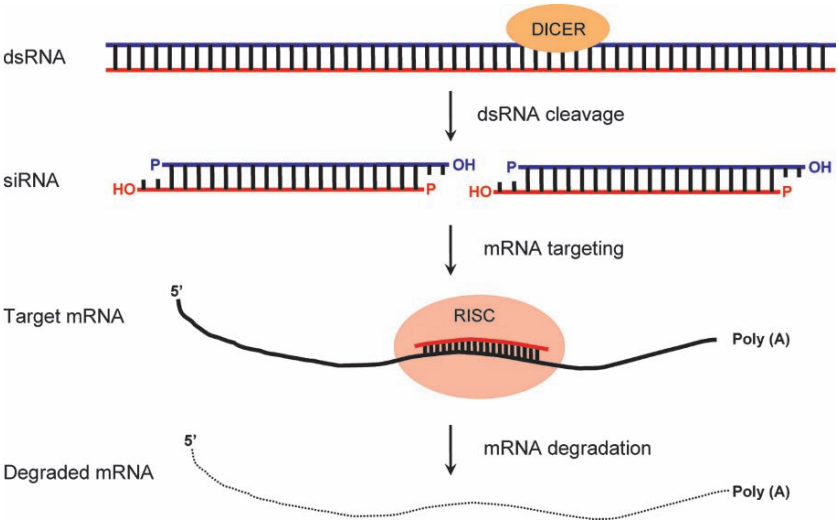


Fig. 7.12. Specific degradation of mRNA by RNA interference (RNAi). A type III ribonuclease (Dicer) binds to and cleaves double-stranded RNA into 21–25 base pair duplexes, termed small interfering RNA (siRNA). The siRNA is incorporated into the multi-protein complex called RNA-induced silencing complex (RISC) that also contains an RNAse. RISC unwinds the siRNA, releases the sense strand, and facilitates hybridization of the antisense strand of the siRNA to the complementary strand of the cognate messenger RNA (mRNA). The binding activates the nuclease activity in RISC, leading to cleavage of the target mRNA. The damaged mRNA is then degraded significantly reducing the expression of the target gene

An advantage of RNAi technology is its efficiency. Experiments are fast, simple, cost-efficient and, importantly, amenable to high-throughput formats. Numerous publications have analyzed complete genomes using RNAi. For example, 86% of all genes of the nematode *Caenorhabditis elegans* were examined by means of RNAi (Kamath et al. 2003). Approximately 10% of the targeted genes led to a change in phenotype, of which approximately one third were already known. In another study, new modulators of p53 that causes cell cycle arrest in human cells were searched for by RNAi. Of the 8,000 genes analyzed, five new modulators were discovered (Berns et al. 2004).

Unfortunately, not all RNAi results are absolutely reliable. For instance, it is known that the efficiency of RNAi is dependent on the incorporated nucleotide sequence. In some cases, the target mRNA is either only partly degraded or not at all. This can lead to false negative results. The experimenter will not see any change in phenotype and infer that the gene product has no important function. Importantly, such data should be checked by an independent method, such as RT-PCR, to determine whether the target RNA has in fact been degraded. Conversely, RNAi can also generate false positive results. In this case, the siRNAs produced by the nuclease Dicer can hybridize with more than one target mRNA, which in turn leads to the degradation of several mRNAs. Therefore, changes in phenotypes cannot be assigned unambiguously, and in the worst case, might lead to incorrect functional predictions of gene products.

To store phenotypic data dedicated genotype–phenotype databases have been established (see Sect. 3.4). PhenomicDB is a very interesting and integrated database in which phenotypes from different organisms that were generated by various methods (e.g., knockout, knock-in, knock-down) are integrated into one database and linked to genotypic data.

7.2 Systems Biology

The foregoing discussion on high-throughput procedures have established genomics, transcriptomics, proteomics, metabolomics, and phenomics as important technologies that facilitate the functional determination of gene products. Like all high-throughput experiments, however, these approaches produce false negative and false positive results. False negative results can lead to information being missed, whereas false positive results might lead the experimenter in the wrong direction. Therefore, to identify the valid results, the idea was to integrate all available data from the above-mentioned technologies and analyze them together (Fig. 7.13). This integration of experimental data

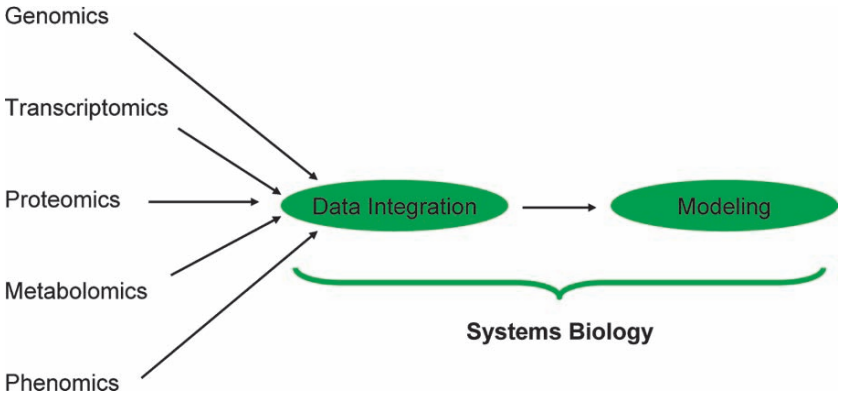


Fig. 7.13. Systems biology integrates data derived from different experimental technologies and generates computational models

improves reliability and the generation of more reliable hypotheses. The research field that focuses on the integration of various high-throughput data is known as Systems biology because it analyzes an entire biological system. Systems biology aims to produce as accurate a picture as possible of all the regulatory processes within a cell or organism by analyzing the interactions between component parts of the biological system e.g., metabolic pathways, organelles, cells, and tissues.

An example of a Systems biology approach is the analysis of phagosomes, which are special organelles found in phagocytosing cells (e.g., macrophages). After phagocytosis, particles such as bacteria are transported into phagosomes where they are destroyed. In a study by Stuart et al. (2007), the phagosome of a cell line derived from the fruit fly *Drosophila melanogaster* was analyzed. Proteins of the phagosome were identified by classical proteomics methods. Construction of a protein–protein interaction networks complemented the results, which were finally validated by RNAi experiments. With the help of this Systems biology approach a detailed model of the phagosome was built and new regulatory proteins and pathways associated with phagocytosis identified.

However, Systems biology frequently goes beyond the mere description and interpretation of experimental data. The ambitious aim is to develop computer models that simulate biological

systems and predict consequences upon changing parameters (e.g., changing the concentration of a specific metabolite). One of the first mathematical models in biology was published in 1952 by Alan Hodgkin and Andrew Huxley, which explained the transmission of action potentials. Since then, the increasing availability of high quality data (both quantitative and qualitative) and larger computer capacities have allowed for more realistic models to be developed. For example, a model has been generated to simulate glycolysis in baker's yeast (*Saccharomyces cerevisiae*). When compared with experimental data, most metabolite concentrations were correctly predicted within a maximal deviation of two (Teusink et al. 2000).

Even more demanding are computer models that perform complete cell simulations (Ishii et al. 2004). A well-known model is the E-Cell system that developed a "virtual bacterium" consisting of 127 essential genes from the genome of *Mycoplasma genitalium* (Fig. 7.14). This bacterium has less than 500 genes and is therefore excellently suited for the construction of a cell model. With the model, the transport of extracellular glucose through the cell membrane could be simulated, in addition to the metabolism of the sugar and the accompanying ATP production. The model also produced a surprise. When the concentration of extracellular glucose was set to zero, the model predicted a temporary increase in the intracellular ATP concentration before a final drop. This was contrary to the expectation that the ATP concentration would drop immediately upon depletion of the glucose. After much speculation, the conclusion was that the model's prediction was correct. During glycolysis, two molecules of ATP are generated from each molecule of glucose. Looking in more detail, it became apparent that in the first part of glycolysis, two molecules of ATP are spent before the production of four ATP molecules in the second part of the reaction. At the moment when the glucose concentration is lowered to zero, the consumption of ATP molecules stops before the generation of new ATP molecules, which are then consumed. Thus, the model recognized the short temporal shift and correctly predicted the temporary increase in ATP concentration.

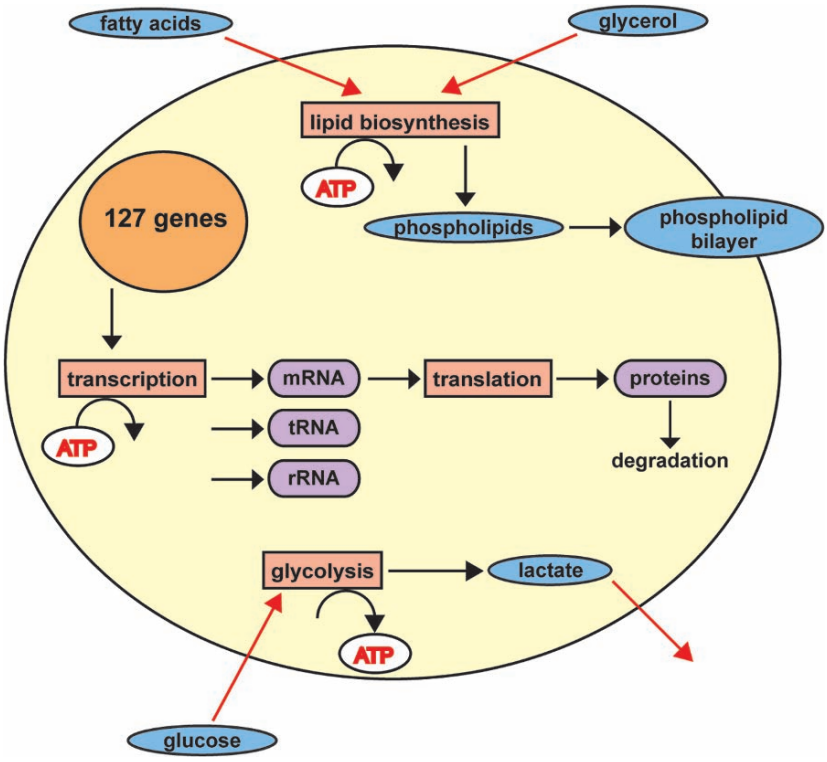


Fig. 7.14. Overview of metabolism in the E-cell model. The model cell has pathways for glycolysis and phospholipid biosynthesis, transcription, and translation

The emergence of Systems biology has been accompanied by the development of a dedicated programming language, the Systems Biology Markup Language (SBML). SBML is an XML-based computer language in which biological networks are represented. The central idea of SBML is the creation of a standardized format that permits the simple exchange of data between the many different software applications. Therefore, each calculated model can be tested in different software environments without additional effort. At the same time, specialized databases have been established in which the computer models can be stored and are accessible by every interested scientist. One example database is the BioModels Database at EBI [biomodels].

7.3 Exercises

1. Go to the NCBI home page and search for the gene `CG15848` in the category *GEO Profiles*. Have a look at the *Value/Rank Plot* of the data set *GDS191*. This study deals with the expression of genes in the fruit fly *Drosophila melanogaster* in different developmental stages from embryo to adult. Which protein does the gene encode and what is its putative function? What is striking about the course of expression of the gene and what hypothesis can be derived from it?
2. As for exercise 1, search for the gene `CG3557` in the study *GDS191*. What is noticeable about the expression of the gene by looking at the *Value/Rank Plot*? Search for genes that show a similar expression profile as `CG3557` by following the link *Profile Neighbors* just above *Value/Rank Plot*. How many genes with a similar expression profile are found and what do these genes have in common? Then go to the NCBI database *UniGene* and search for information about the expression of gene `CG3557`. In this database, the expression is determined using ESTs. Do the microarray results agree with those of the EST analysis?
3. In the *GEO Datasets* database find the entry `GDS1399`. *GDS1399* is a microarray experiment that examines the effect of distinct gene mutations on global gene expression in *Escherichia coli*. The *dam* mutant lacks the enzyme *DNA adenine methyltransferase*. This enzyme transfers methyl groups to sites with a characteristic short sequence in the *E. coli* genome and thereby exerts a significant influence on the regulation of gene expression. An *E. coli* strain without any genetic alteration is referred to as the wild-type. How many replicates of the wild-type and *dam* mutants were used in the experiment?
4. Determine the number of genes whose expression in the *dam* mutant is increased or decreased about two-

fold or more compared to the wild-type strain. To do this go to the drop down menu in the section *Subset and Sample Info* on the right side of the table. Select the entry *Query mean group A vs. B by values*. Then choose *2+ fold* and next to it, *higher* or *lower*. In the columns to the left and right, check B for the wild-type and A for the *dam* mutant. Start the search with *Query A vs. B*.

5. For how many genes is the expression in the *dam* mutant significantly different from that in the wild-type? To answer this question, go to the drop down menu on the right side of the table in the section *Subset and Sample Info*. Select the entry *Two-tailed t-test (A vs. B)* and further down, *0.050 significance level*. In the columns to the left and right, check B for the wild-type and A for the *dam* mutant. Start the search with *Query A vs. B*.
6. How many genes in the *dam* mutant are significantly up- or down-regulated compared to the wild-type? Look again for the drop down menu on the right side of the table in the section *Subset and Sample Info*. Select the entry *One-tailed t-test (A > B)* and below that, *0.025 significance level*. In the columns to the left and right, check B for the wild-type and A for the *dam* mutant. Start the search with *Query A vs. B* to identify those genes that are significantly up-regulated. To find the down-regulated genes select the entry *One-tailed t-test (A < B)* (please note: this time it is A < B). In the columns to the left and right, check B for the wild-type and A for the *dam* mutant. Start the search with *Query A vs. B*.
7. Go to the *Stanford Microarray Database* (SMD) and become familiar with the tutorials (http://genome-www5.stanford.edu/help/tutorials_subpage.shtml). They offer an excellent introduction to the topic of microarray analysis.

8. At the EBI web page is found the microarray software *Expression Profiler*. Its module EPCLUST allows the generation of gene clusters and their visualization. At the URL, *http://www.bioinf.ebc.ee/EP/Docs/dist_clust/*, the authors point out how cluster generation is affected by the choice of different algorithms. As an example, 10 genes, for which their expression was measured in four experiments (e.g., at four different time points), are compared to each other. Go to the category *Simpler distance measures* and for gene 04 compare the results of the algorithms *Euclidian distance, Euclidian distance squared, Manhattan distance, Average distance*, *Square Root of Average distance*, and *Number of attributes with opposite sign*. Use only the representation *distance average cluster tree* to perform this. Does gene 04 build clusters with other genes? If yes, which genes are these and which algorithms were used?
9. Go to the web page of the BROAD institute and download the software *GenePattern* (http://www.broad.mit.edu/cancer/software/genepattern/). On this page is found a tutorial as well as the corresponding data sets used in the tutorial. Test the program using these data sets.
10. Go to the homepage of Expasy. Search for the database *SWISS-2DPAGE* under *databases* and query it using *by description* under *Access to SWISS-2DPAGE*. Enter hsp60. *HSP60* stands for *Heat Shock Protein 60*. Select *CH60-HUMAN* and under *2D PAGE maps for identified proteins* click on the picture of the HepG2 gel. The spots corresponding to *HSP60* are marked in red. How many spots are found for *HSP60*? How can it be explained that several spots exist for one protein?
11. Next click on the picture of the 2D electrophoresis from liver. How many spots correspond to *HSP60* in this case? Why are fewer spots found now?

12. Next go to *2D PAGE maps for unidentified proteins* and click on the picture *HepG2 secreted proteins*. Can *HSP60* be found on this gel? Give reasons for the result.
13. Start the query *by clicking on a spot* at the entry page of SWISS-2DPAGE. Then under *HUMAN* select *2D-PAGE of nucleolar proteins from Human HeLa cells*. Click on the highlighted spot with the lowest molecular weight and a pI value of approx. 5.7. Which protein is it? What is the molecular weight of the protein?
14. Stay within the SWISS-2DPAGE database and under *Access to SWISS-2DPAGE* click on the hyperlink *retrieve all the protein entries identified on a given reference map*. Select *HUMAN:HepG2* and create a table by clicking on *Execute query*. What methods were used to identify the proteins?
15. In the table, search for the unknown protein that represents spot 106 and click on its *SWISS-2DPAGE Access Number P31929*. Look for the subject *Cross references* and follow the Swissprot hyperlink. What is the partial amino acid sequence of the protein that was identified by microsequencing?
16. Go to the home page of the program *PeptideMass* (http://www.expasy.org/tools/peptide-mass.html). Run an in silico digest of the human protein kinase src (*Accession number P12931*) with the enzyme trypsin. How many peptides with a mass of >1,000 Dalton (Da) are generated by this digest? What is the peptide mass of the largest peptide?
17. By 2D gel electrophoresis of a protein extract from bovine placenta, a protein of 38 kDa with a pI value of approximately 7.0 has been isolated. After incubation with the protease trypsin, peptides with masses of 1.433, 1.422, 1.088, and 1.030 Da are detected by mass spectroscopy (accuracy of measurement ± 0.5). Using the program *Aldente* (http://www.expasy.org/tools/

aldente/) in the database UniProt/SwissProt, identify those proteins that generate peptides of similar masses after an in silico tryptic digest.

18. Go to the *YEAST protein complex database* (http://yeast.cellzome.com/) and click on the button *enter as guest*. How many multi-protein complexes are stored in the database? Look at complex 116. How many proteins are found in this multi-protein complex? What is the function of this protein complex?
19. Click on the name of the protein NHP10. Is NHP10 also found in other multi-protein complexes? What functions do these multi-protein complexes have?

WWW Links

affymetrix: http://www.affymetrix.com/
agilent: http://www.chem.agilent.com/scripts/pds.asp?lpage = 27881
arrayexpress: http://www.ebi.ac.uk/arrayexpress/index.html
bind: http://www.bind.ca/
bioconductor: http://www.bioconductor.org/
biomodels: http://www.ebi.ac.uk/biomodels/
ercc: http://www.cstl.nist.gov/biotech/Cell& TissueMeasurements/Gene Expression/ERCC.htm
genedisruptionproject: http://www.fruitfly.org/p_disrupt/index.html
geneknockoutconsortium: http://www.celeganskoconsortium.omrf.org/
geneknocksmouse: http://www.knockoutmousestation.com/
genepattern: http://www.broad.mit.edu/cancer/software/genepattern/
geo: http://www.ncbi.nlm.nih.gov/geo/
hmp: http://www.metabolomics.ca/
intact: http://www.ebi.ac.uk/intact/site/index.jsf
maqc: http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/
melanie: http://www.expasy.org/melanie/
miame: http://www.mged.org/Workgroups/MIAME/miame.html
rosetta: http://www.rosettabio.com/products/resolver/default.htm
sage: http://www.sagenet.org/
sagemap: http://www.ncbi.nlm.nih.gov/SAGE/
tm4: http://www.tm4.org/
yeastproteincomplex: http://yeast.cellzome.com/

Literature

Berns K, Hijmans EM, Mullenders J et al (2004) A large-scale RNAi screen in human cells identifies new components of the p53 pathway. Nature 428(6981):431–437

Bonetta L (2006) Gene expression: an expression of interest. Nature 440(7088):1233–1237

Churchill GA (2002) Fundamentals of experimental design for cDNA microarrays. Nature Genetics Suppl 32:490–495

Duarte NC, Herrgard MJ, Palsson BO (2004) Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. Genome Res 14(7):1298–1309

Eisenstein M (2006) Protein arrays: growing pains. Nature 444(7121):959–962

Gavin AC, Bosche M, Krause R et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868):141–147

Gershon D (2005) DNA microarrays: more than gene expression. Nature 437(7062):1195–1198

Golub TR, Slonim DK, Tamayo P et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–537

Ishii N, Robert M, Nakayama Y et al (2004) Toward large-scale modeling of the microbial cell for computer simulation. J Biotechnol 113(1–3): 281–294

Ji H, Davis RW (2006) Data quality in genomics and microarrays. Nat Biotechnol 24(9):1112–1113

Kamath RS, Fraser AG, Dong Y et al (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature 421(6920): 231–237

Kell DB (2006) Systems biology, metabolic modelling and metabolomics in drug discovery and development. Drug Discov Today 11(23–24):1085–1092

Raamsdonk LM, Teusink B, Broadhurst D et al (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nat Biotechnol 19(1):45–50

Rual JF, Venkatesan K, Hao T et al (2005) Towards a proteome-scale map of the human protein–protein interaction network. Nature 437(7062):1173–1178

Quackenbush J (2001) Computational analysis of microarray data. Nature Rev Genetics 2:418–427

Stuart LM, Boulais J, Charriere GM (2007) A systems biology analysis of the *Drosophila phagosome*. Nature 445(7123):95–101

Teusink B, Passarge J, Reijenga CA et al (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. Eur J Biochem 267(17):5313–5329

Vanhecke D, Janitz M (2005) Functional genomics using high-throughput RNA interference. Drug Discov Today 10(3):205–212

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A et al (2001) Global analysis of protein activities using proteome chips. Science 293:2101–2105

8 Comparative Genome Analyses

8.1 The Era of Genome Sequencing

The extraordinary achievements of genome-based biology within the last 15 years can be explained for the most part by the technological progress in DNA sequencing as well as the developments in hardware and software to store and annotate the huge amounts of data. The total number of all freely accessible nucleotides in GenBank [genbank], the DNA sequence database at NCBI, is 80 billion bases within 76 million DNA sequences (Version 161.00, August 2007). The number of all protein sequences in the world's largest nonredundant protein database UniprotKB [uniprotkb] at EBI adds up to 4.9 million (Release 12, July 2007).

The first completely sequenced genomes, from the microbial organisms *Haemophilus influenzae* (Fleischmann et al. 1995) and *Mycoplasma genitalium* (Fraser et al. 1995), were published in 1995. Today 589 sequenced microbial genomes are available (542 from bacteria and 47 from archaebacteria) and another 1,143 microbial genomes are being sequenced [gold, cmr] (September 2007). Among these are the complete genomes of both virulent and nonvirulent strains of the same bacterium, which facilitates the identification of virulence factors. It is assumed that, within the next few years, all important pathogenic microorganisms of man, animals, and plants will have been sequenced. This flood of data will lead to new possibilities in the production of antimicrobial agents, vaccines, and diagnostic tests, all of which

should aid the ongoing fight against infectious diseases (Selzer et al. 2000).

Meanwhile, the complete genomes of 70 eukaryotic organisms are known. These include *Saccharomyces cerevisiae* (baker's yeast), *Caenorhabditis elegans* (nematode), *Drosophila melanogaster* (fruit fly), *Arabidopsis thaliana* (mouse-ear cress), *Takifugu rubripes* (tiger puffer), *Homo sapiens* (man), and *Mus musculus* (mouse). Furthermore, as of September 2007, 775 eukaryotic genome sequencing projects are underway. These data will eventually contribute to the decoding of the secrets of biology and thereby help combat serious diseases of humans and animals.

8.2 Drug Research at the Target Protein

Systematic research into active substances as novel drugs dates back to the second half of the 19th century. A prime example is acetylsalicylic acid, which was synthesized in 1897 by two chemists, Felix Hoffmann and Arthur Eichengrün of the company Bayer. It is now world-famous under the trade name aspirin. It is still a disputed question as to which of the two was the actual inventor of the synthesis of acetylsalicylic acid. Regardless, this substance has neither lost its economic nor scientific importance. Since then the identification of active compounds, including those with bioactivity against infectious diseases, has been dominated by direct testing (screening) in biological systems, mostly laboratory animals. Many antibiotics in use today were discovered in the first half of the twentieth century. However, since about the 1960s the number of new drugs has steadily declined. There are a number of reasons for this, including the constant decline in the success rate of nontargeted screening, the increased costs for research and development, and the higher standards of safety required. Furthermore, in the area of infectious diseases, the situation has been worsened by the emergence and increased spread of drug resistance. However, at about the same time, a new era of molecular research began in

1953 with the deciphering of the three-dimensional structure of the DNA double helix by James D. Watson and Francis H.C. Crick.

By sequencing whole genomes and the ensuing biological information, the approach to drug discovery has changed. Thus, in the target-based approach (Fig. 8.1), in which a target protein is used to search for new active compounds, the first step is to identify those proteins that are essential to the survival of the

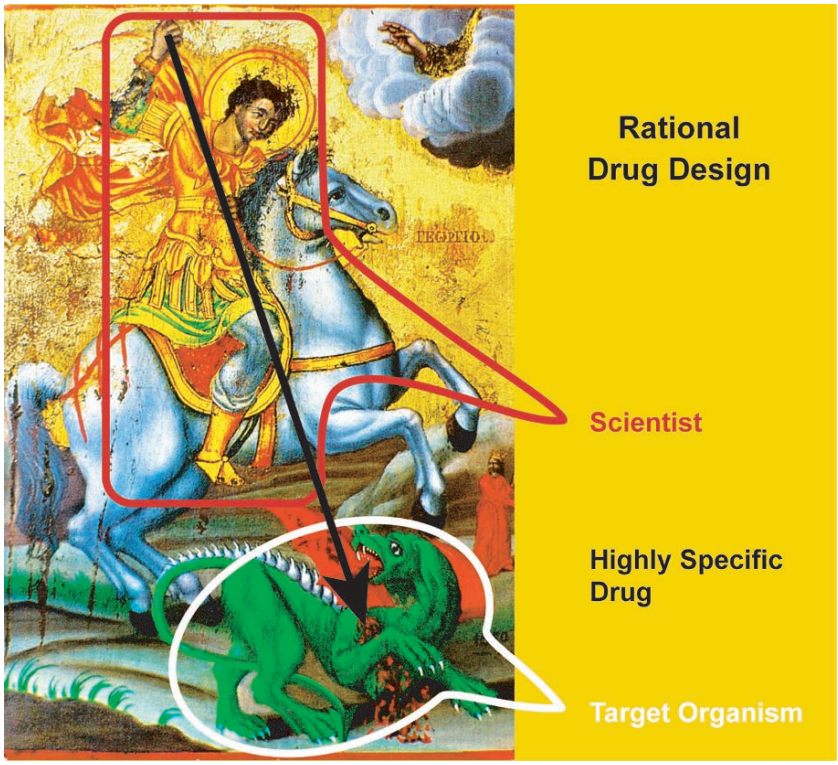


Fig. 8.1. The analogy of a Christian icon for *Target Based Approach to Drug Development*. The icon shows Saint George as a dragon-slayer. The dragon symbolizes the target organism that can only be killed by a precise stroke to the heart (target protein). All other targets are irrelevant. Based on this realization, Saint George (the scientist) uses his horse (scientific tools) to guide his lance (a highly selective drug) into the target. The original icon is in the Preveli monastery, Crete (Greece)

pathogenic organism. The second step is to find active chemical substances that influence the isolated target protein in the desired way. Only after such optimized chemical substances with the desired activity spectrum have been found using these in vitro methods will further testing be performed in a biological system (see also chapter 6). For example, to develop a new antibiotic, an ideal prerequisite would be that the target protein is essential to the survival of the pathogenic bacteria under study and that the host organism does not also possess the same or similar protein that may also be targeted potentially resulting in toxicity. In this scenario, comparative whole genomic analysis would be well-suited to identify pathogen-specific targets. Indeed, this approach was taken by Huynen et al. (1998) in their work on the genomes of three bacteria, *Escherichia coli, Haemophilus influenzae*, and *Helicobacter pylori*. Orthologous proteins were identified in either all three or in two of the three organisms, in addition to species-specific proteins. For *H. pylori*, the major causative agent of gastric and duodenal ulcers, the authors predicted that 123 proteins were involved in interacting between the pathogen and the host, i.e., represented potential targets for the development of an antibiotic. In pharmacological research, conserved targets usually lead to the development of broad-spectrum antibiotics, whereas with species-specific targets, narrow spectrum antibiotics are generated.

Because of the increasing number of completely sequenced bacterial genomes, it is clearer which genes are generally conserved among bacteria and which are specific for certain bacterial species. However, it is not always easy to settle on the threshold of sequence similarity that blocks the pursuit of a target based drug discovery approach due to potential toxicity arising from an unwanted interaction with the human protein counterpart. For example, bacterial dihydrofolate reductase has a sequence identity of 28% at the amino acid level to the corresponding human protein, yet the antibacterial drug, trimethoprim, is a very selective inhibitor of only the bacterial ortholog.

8.3 Comparative Genome Analyses Provide Information about the Biology of Organisms

Comparative genome analyses are frequently referred to as *Comparative Genomics* whereby two or more genomes are compared to one another (Beckstette et al. 2004). The goal is to find similarities and differences between these genomes that yield information about the biology of the respective organisms. Another important aim of Comparative Genomics is the description of genome structure and the identification of coding and noncoding regions (Wei et al. 2002).

8.3.1 Genome Structure

Analysis of genome structure of one or more genomes includes statistical measurements such as size and nucleotide composition, frequency of codon usage, and identification of conserved regions between two or more genomes. The percentage and frequency of guanine and cytosine (GC) content or adenine and thymidine (AT) content differs between groups of organisms and seems to have changed considerably in the course of evolution from microorganisms to multicellular organisms. Likewise, the codon usage for encoding identical amino acids is not the same in every organism (see Chaps. 2 and 4).

Many comparative studies of the genomes of man and mouse have shown that their organization, to a large extent, is similar. This indicates that, since the last common ancestor, the structural organization has been conserved. To describe such similarities between evolutionarily related chromosomal segments among species, various terms have been defined or broadened in their definition. If two or more genes lie on the same chromosome then one speaks of syntenic genes, or synteny. This definition only applies, however, within a species. Between species, the definition was expanded such that when syntenic genes of

orthologous proteins on a single chromosome are conserved across species the term conserved synteny applies. The order of the genes on the chromosomes is not considered (Fig. 8.2). If, in addition, the order of the genes on the chromosomes is also conserved, then the regions are called conserved segments or conserved linkages.

With the growing number of completed eukaryotic genome sequences it has become apparent that conserved segments are present in all mammals. Although syntenic regions are observed between species such as man and the puffer fish which separated approximately 450 million years ago, no larger blocks of conserved genome organization have been described thus far for such distantly related organisms (Frazer et al. 2003).

8.3.2 Coding Regions

The comparative analysis of coding regions between different genomes includes not just the identification of protein encoding regions but also the direct comparison of the types and numbers of orthologous and paralogous proteins. The identification of genes in prokaryotes is comparatively simple as there are relatively few noncoding regions. Normally 85% of a bacterial genome encodes proteins or RNAs with the smaller portion encoding regulatory units or noncoding regions. In contrast the prediction of genes in eukaryotes is far more difficult because noncoding regions have increased during evolution. Eukaryotic genomes possess a large number of intergenomic regions as well as a multitude of noncoding repeats. Furthermore, eukaryotic genes contain introns and exons and different proteins frequently arise as a consequence of alternative splicing (see Chap. 2 and 5). For instance, the genome of the prokaryote *Escherichia coli* has approximately 4,300 genes at a genome size of 4,600 kilobases (kb) with, on average, one gene for every 1 kb in length. In contrast, the genome of the eukaryotic unicellular yeast, *Saccharomyces cerevisiae,* has approximately 6,300 genes at a genome

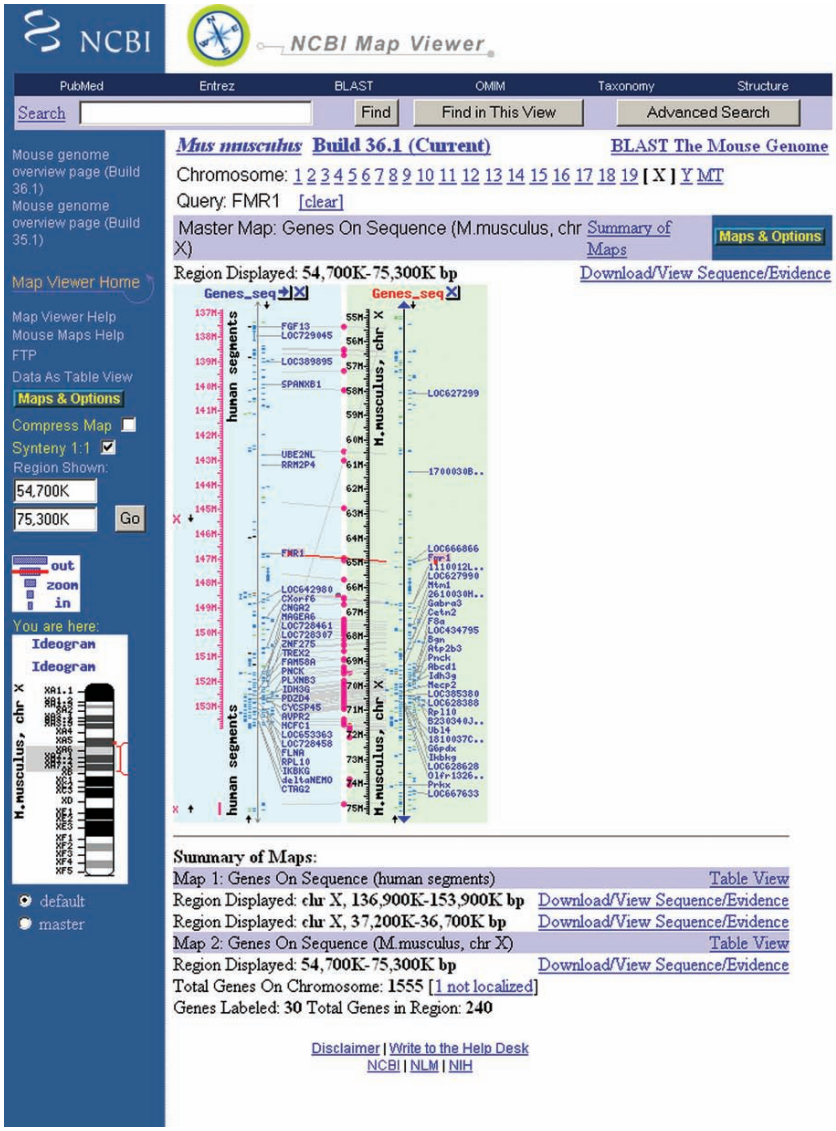


Fig. 8.2. Homology map of the X chromosome of man and mouse as taken from NCBI. Shown are a part of the detailed map for the X chromosome of the mouse (as a reference) and the human. The syntenic genes of mouse and human FMR1 are indicated in color and connected by a red line. Other syntenic genes on this part of the chromosome are indicated by gray connecting lines. (Printed with permission of the NCBI)

size of 12,000 kb, and the genome of the multicellular worm, *Caenorhabditis elegans,* contains approximately 19,000 genes at a genome size of 97,000 kb. Phylogenetically speaking, the human genome is very young and shows an enormous difference between the number of genes and its genome size; 30,000–35,000 genes at a total size of approx. 2.9 gigabases.

8.3.3 Noncoding Regions

The comparative analysis of noncoding regions, which in humans and other mammals can account for more than 97% of the genome, is still one of the greatest challenges of bioinformatics. Still, this area of genome analyses has received much attention in the last few years in the hope of identifying genomic regulatory units. For instance, it has already been shown bioinformatically that conserved noncoding regions have an accumulation of transcription factor binding sites. Furthermore, the probability of identifying such regulatory areas in noncoding regions increases when more than two genomes of closely related organisms are compared. It has already been shown that half of the noncoding regions identified in a comparison of the human and mouse genome are also conserved in the genome of the dog.

8.4 Comparative Metabolic Analyses

For gene prediction, special emphasis has been placed on those genes that encode proteins involved in metabolism. Using gene prediction it is possible to identify whether an organism possesses metabolic pathways, such as those in glycolysis or the citrate cycle, or whether alternative pathways are employed to generate energy. A comparison of two or more genomes at the level of their metabolic pathways can also be used to identify metabolic targets. This is particularly effective with prokaryotes

because many genomes have already been sequenced. There are a number of software technologies used to compare metabolomes: the Encyclopedia of *Escherichia coli* Genes and Metabolism (EcoCyc) [ecocyc], the Kyoto Encyclopedia of Genes and Genomes (KEGG) [kegg] (Fig. 8.3), and the database, Reactome [reactome], are among the best known.

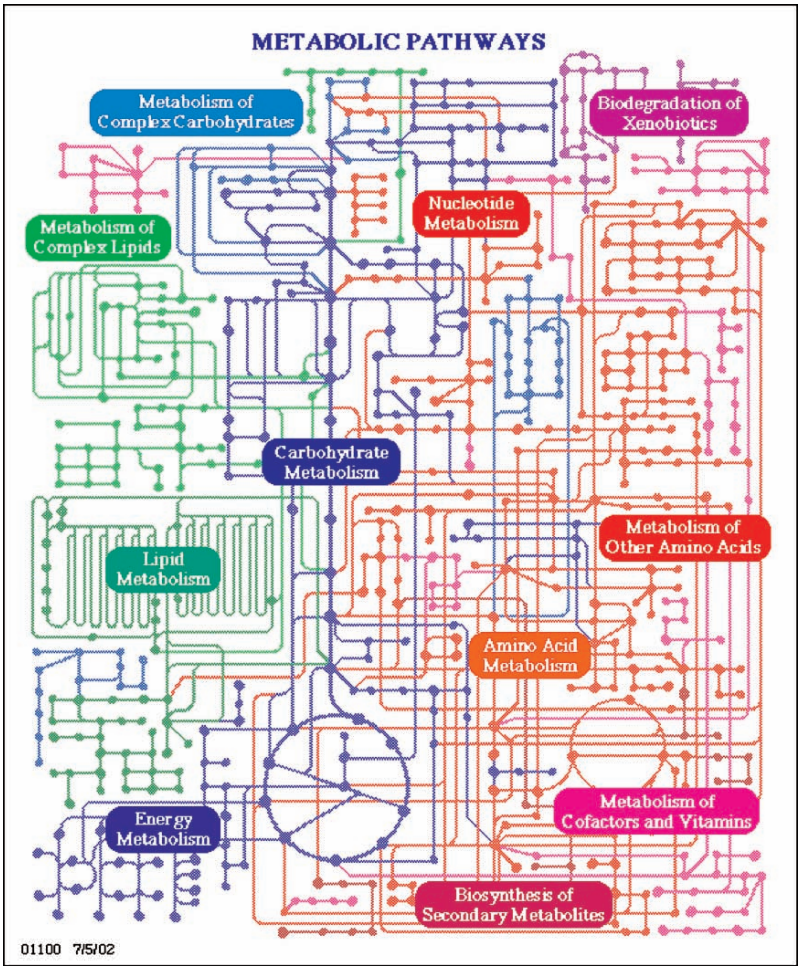


Fig. 8.3. Map of the metabolic pathways in the KEGG database (Printed with permission of the KEGG)

The methods include manual and semiautomatic analyses. So far, however, there is no fully automatic analysis software that can calculate all the metabolic pathways. Furthermore, such databases are not always complete. While initially the databases are dealt mostly with metabolic pathways, over time regulatory mechanisms such as membrane transport, gene regulation, and signal transduction have also been incorporated (Fig. 8.4).

In sequenced genomes, genes or proteins can be divided into orthologous groups. Accordingly, proteins that are either present or absent can be systematically identified and the resultant functional metabolic pathways constructed. If some required proteins are missing, either the corresponding metabolism is nonfunctional or others (including thus far unknown) are involved. During the analysis of the genome of *Helicobacter*

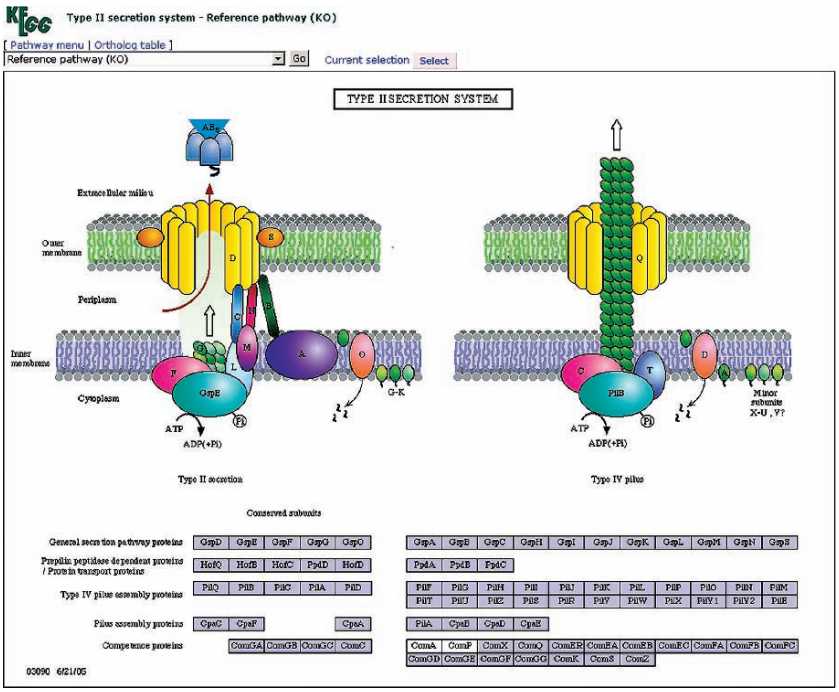


Fig. 8.4. Schematic representation of the type-II secretion pathway (Printed with permission of the KEGG)

pylori it was noticed that neither glycolysis nor pentose phosphate metabolism was operational due to the absence of the requisite enzymes. Because both metabolic pathways generate protons and, therefore, lower pH, their operation would lead to a further burden on an organism that already lives in the acidic environment of the stomach. In contrast, the genes coding for proteins that metabolize organic acids such as those involved in anabolic gluconeogenesis are present. Thus, *H. pylori*'s energy production seems to be fueled by amino acid degradation and the substrates necessary are probably directly derived from the gastrointestinal tract.

To find specific metabolic pathways in KEGG, the genome must be compared with a reference genome. If the gene exists it is highlighted in color. A sequence of colored rectangles therefore reflects the specific metabolic pathway in the studied organism (Fig. 8.5). To be successful with this strategy, however, all alternatives must be known. It is often the case that a metabolic pathway does not show all the genes or proteins and is, therefore, considered incomplete. Reasons include that gene predictions were incomplete or incorrect, or that current knowledge regarding the specific metabolic pathway is limited. It is also possible that one protein performs several functions and, thus, has a larger metabolic spectrum than originally anticipated. Finally, alternative metabolic pathways that lead to the same biological result can not be excluded.

8.4.1 Kyoto Encyclopedia of Genes and Genomes

KEGG is a product of the Japanese GenomeNet and is widely used for the analysis of metabolic pathways. Two of the three main databases, PATHWAY and LIGAND, deal with metabolic processes in cells and organisms. The third database, GENE, contains gene and protein information from sequencing projects and is comparable to other primary databases (Kanehisa et al. 2006). These databases are completed by BRITE, an ontology database for the description of biological relationships within

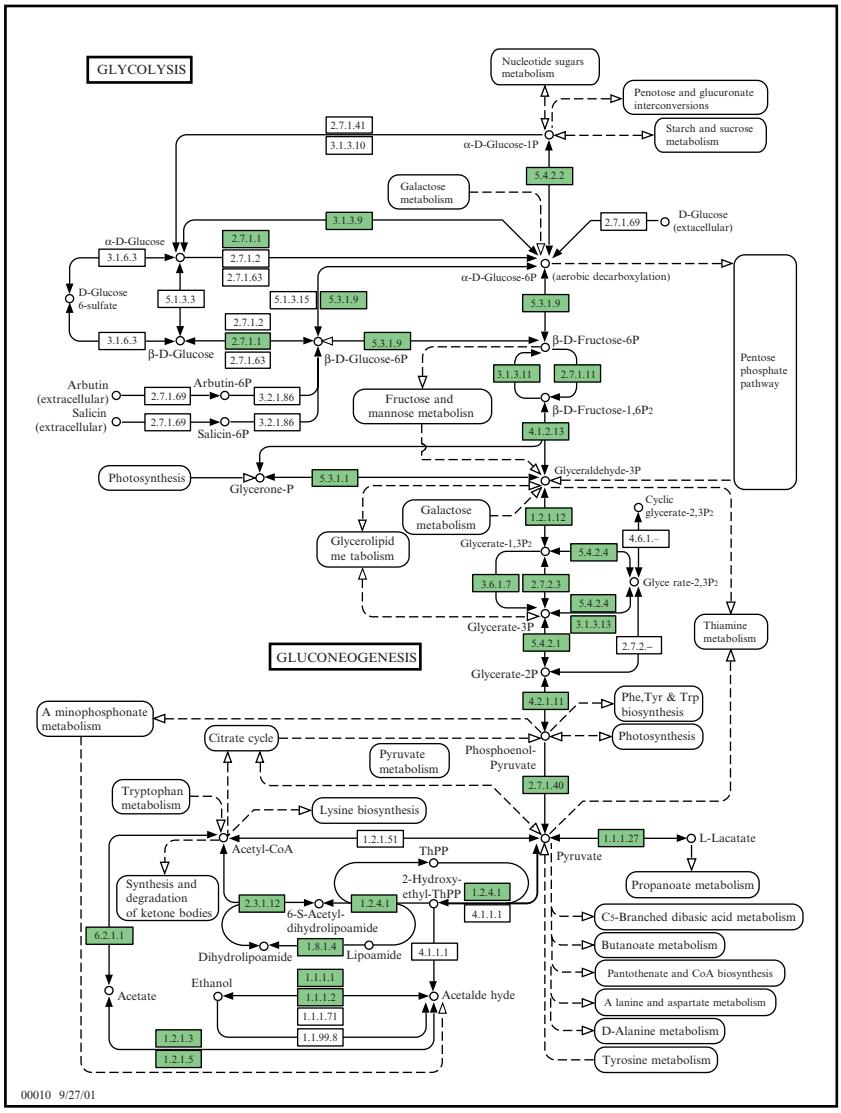


Fig. 8.5. Metabolic map of the glycolytic/gluconeogenic metabolism. The thus far known enzymes of this metabolic pathway in humans are highlighted in color (Printed with permission of the KEGG)

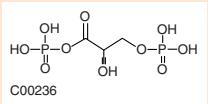
the pathways. Furthermore, KEGG offers information on experimental data from gene expression and yeast two-hybrid experiments (EXPRESSION). Another database, SSDB, contains information on groups of orthologous proteins.

The most interesting databases are undoubtedly PATHWAY and LIGAND. PATHWAY contains graphical representations of metabolic pathways from a number of organisms, mostly prokaryotes, and also eukaryotes. The representations of the metabolic pathways are similar to those in the Biochemical Pathways Chart from Boehringer Mannheim [biochem-pathway]. The individual maps can be selected from a list or chart sorted according to the main metabolic pathways (Fig. 8.3). The known enzymes in reference pathways can be highlighted in color. This facilitates the comparison of metabolic pathways between organisms. Figure 8.5 shows as an example the glycolysis/gluconeogenesis metabolism in humans. The enzymes drawn in green (small boxes) have already been described or are present in the human genome. The individual metabolic charts on the KEGG-WWW server are connected to the LIGAND database, a chemical database that contains the corresponding substances, enzymes, and reactions in the respective metabolic pathway. The small rectangular boxes with an enzyme number (NC-IUBMB 1992) [enzyme] are for cross-referencing. The EC number consists of four blocks of numbers, each separated by a period. The first number describes one of the six functional groups (oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases), the two blocks following refer to further subclasses within the main class. The last block is a consecutive number of each of the enzymes in the particular subclass. Further cross references are indicated by the circular symbols next to the substance names (e.g., β–D-glucose) as well as the rounded borders of other metabolic pathways. The latter do not lead to the LIGAND database, however, but to the detailed description of the respective metabolic path. In the case of glycolysis/gluconeogenesis metabolism, for example, this leads to the citrate cycle or pentose phosphate metabolism.

By clicking on the circle at *Glycerate-1,3P2* a new window opens with an entry from LIGAND (Fig. 8.6). In addition to a unique substance number, the substance name and the empirical and constitutional formulae of the substance are given. What follows are cross-references to entries of reactions in which 1,3-bisphospho-D-glycerate is involved, to the metabolic pathways

[LinkDB]

ENTRY C00236
NAME 3-Phospho-D-glyceroyl phosphate
1,3-Bishospho-D-glycerate
(R)-2-Hydroxy-3-(phoshonooxy)-1-monoanhydride
Swith phosphoric propanoic acid
FORMULA C3H8O10P2



REACTION R01061 R01063 R01512 R01515 R01517 R01660 R01662 R02188
PATHWAY PATH: MAP00010 Glycolysis / Gluconeogenesis
PATH: MAP00710 Carbon fixation
ENZYME 1.2.1.12 1.2.1.13 1.2.1.59 2.7.1.106
2.7.2.3 2.7.2.10 2.7.4.17 3.6.1.7
5.4.2.1 5.4.2.4
DBLINKS CAS: 38168-82-0
///

Option:

1.Launch ISIS/Draw ... See instructions for setup.

[KEGG | DBGET | GenomeNet]

Fig. 8.6. Database record in the LIGAND database for β-D-glucose (Printed with permission of the KEGG)

in which it operates, and to enzymes that are associated with the conversion of 1,3-bisphospho-D-glycerate. The CAS number in the field *DBLINKS* is a unique number given to every chemical substance by the Chemical Abstract Service [cas] upon first publication. The hyperlink *Launch ISIS/Draw* in the section *Option* allows one to load the constitutional formula directly into the structure drawing program ISIS/Draw. ISIS/Draw is a capable program for drawing chemical structural formulae and is available free of charge from the homepage of MDL Information Systems, Inc. [mdl].

In addition to database query via the graphical representation of the metabolic pathways, LIGAND facilitates text searches for reactants or enzymes and searches for the substructures of more complex chemical structures.

8.5 Groups of Orthologous Proteins

Upon completion of a genome sequencing project, attention is turned to the analysis and classification of the predicted genes and the possible function of their gene products. The simplest approach is to compare unknown gene sequences with known genes and assign a function based on similarity. Some of the tools have already been described earlier in this book. Because the comparison of whole genomes or proteomes with conventional methods is very laborious, however, commercial software packages have been developed that allow the comparison of large sequence data sets and the identification of common sequences. One such program, Genlight, is freely available to the academic community [genlight] (Beckstette et al. 2004).

In those cases where larger phylogenetic distances exist between organisms, direct sequence comparison is difficult due to low sequence similarities. Another approach to phylogenetic classify proteins, therefore, is the generation of *Clusters of Orthologous Groups* (COG). In any given COG, orthologous sequences are combined, i.e., over the course of evolution, all proteins within a COG have evolved from a common precursor via species formation or gene duplication. The generation of COGs is done by pair-wise sequence comparisons of all the proteins under study and the subsequent analysis of the derived network of relationships.

The determination of orthologous proteins within a group of species is of great importance both for evolution research and identifying the function of unknown proteins, in addition to comparative genome analysis. Therefore, a number of complex systems for the classification of orthologous proteins have been developed. One of the best-known systems is the COG database at NCBI [cog] (Wheeler et al. 2007; Fig. 8.7). At present, the database contains 9,724 clusters of orthologous proteins from 73 completely sequenced genomes (May 2007). Besides text-based searches in the database, it is also possible to compare one's own sequences against the database and thereby predict protein

function using the program COGnitor. The quality of the COG database is high due to the manual curation of individual data entries. It is, however, a static system, i.e., the number and types of species making up the precomputed clusters in the database cannot be changed by the user.

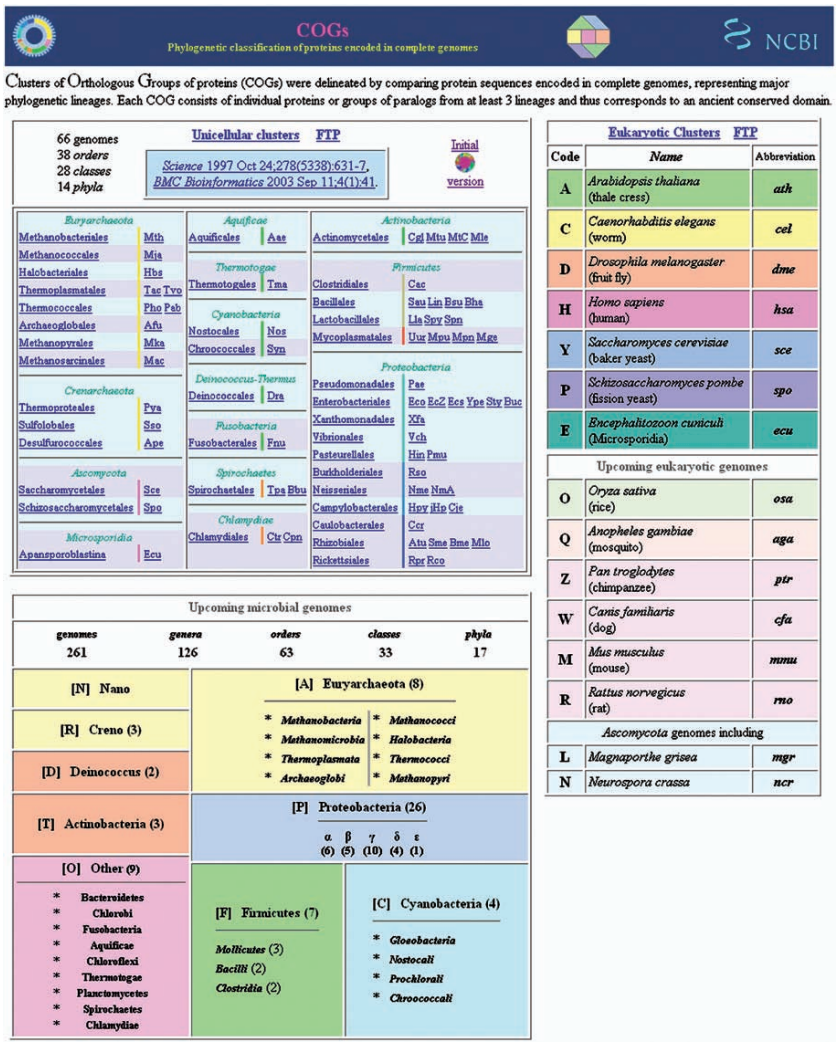


Fig. 8.7. Homepage of the COG database [cog] (Printed with permission of the NCBI)

In contrast to the COG database, the *Microbial Genome Database* (MBGD) facilitates the dynamic calculation of clusters according to the parameters set by the user [mbgd] (Uchiyama 2003; Fig. 8.8). This approach takes into account that the classification of proteins into orthologous clusters can depend on the choice of the organisms and that a static set of clusters may unintentionally influence the results. The MBGD database therefore provides a classification scheme rather than the static result of a classification. The cluster calculation depends on the user's parameter entries, either via orthology or homology criteria, and is based on precomputed similarity tables of all the proteins in the database. The same possibilities to query the MBGD database are as for the COG database. Besides text-based queries, MBGD offers a tool to evaluate and annotate one's own sequences.

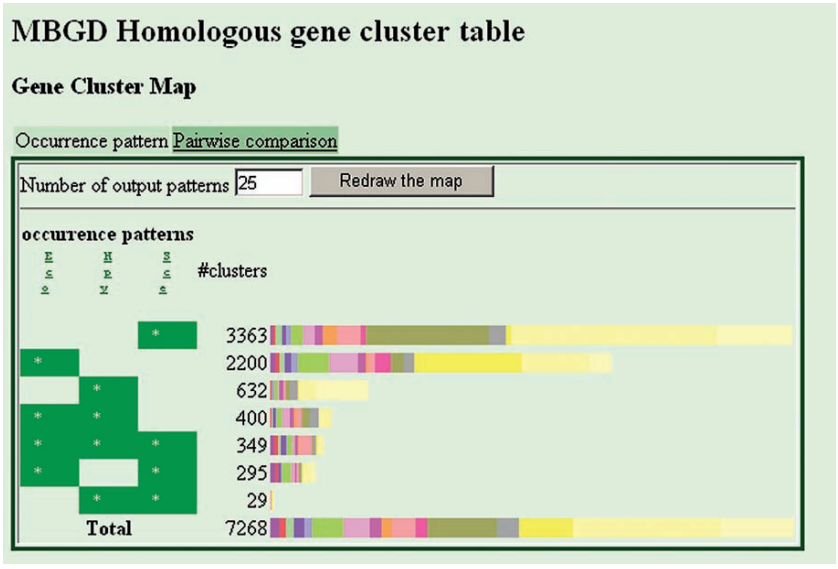


Fig. 8.8. Result of a *Cluster Analysis* of the MBGD database [mbgd]. The organisms *E. coli* (Ecs), *H. pylori* (Hpj),and *S. cerevisiae* (Sce) were selected for calculating the underlying *Cluster Table* (Printed with permission of the MBGD)

8.6 Exercises

1. How many genome sequencing projects are ongoing and how many have been completed?
2. Go to the KEGG home page (http://www.genome.ad.jp/kegg/) and display the metabolic map of the glycolysis/gluconeogenesis metabolism.
3. Which enzymes catalyze the conversion of L-lactate to pyruvate? Does this conversion take place in humans? Does *Saccharomyces cerevisiae* make use of this metabolic step?
4. How do the enzyme hyperlinks differ between the reference pathway and a species-specific map (e.g., *Homo sapiens*)?
5. Display the chart of the glycolysis/gluconeogenesis metabolism and compare the species-specific map of humans (abbr.: hsa) with that of *Helicobacter pylori 26695* (abbr.: hpy). What are the significant differences between the two maps? How can these differences be explained?
6. Go to the NCBI BLAST home page and perform a BLAST search with the sequence Q9ZK41 against the Microbial Genome Database (http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi) for the genomes of the following organisms or groups of organisms:
   Bacteria/Firmicutes/Bacillalis/Staphylococcus aureus RF122
   Bacteria/Firmicutes/Lactobacillalis/Streptococcus pneumoniae D39
   Bacteria/Proteobacteria/epsilon subdivision
   How many reasonable hits from each organism are acquired?
7. The Comprehensive Microbial Resource of the TIGR institute (Peterson et al. 2001, http://cmr.tigr.org/tigr-scripts/CMR/CmrHomePage.cgi) allows the simultaneous comparison of completely sequenced bacterial

genomes. Compare the genome of *H. pylori 26695* with the genomes of the following three *E. coli* strains: *E. coli K12-MG1655, E. coli O157:H7 EDL933*, and *E. coli O157: H7 VT2-Sakai*. How many proteins in *H. pylori* have no homologs in the *E. coli* strains when a similarity threshold of less than or equal to 30% is considered?

8. Go to the MBGD database (http://mbgd.genome.ad.jp/) and calculate a cluster table for the following organisms: *Staphylococcus aureus (RF122)*, *Escherichia coli (536)*, and *Saccharomyces cerevisiae (S288C)*.
9. From Exercise 8 how many clusters contain genes of all the selected organisms? Display these. To which functional category does the first cluster belong?
10. Go back to the start page of the MBGD database (http://mbgd.genome.ad.jp/). In the overview of organisms, only those previously selected will be marked in red. Make a keyword search for the keyword *fructokinase*. How many entries are found?

WWW Links

biochem-pathway: http://www.expasy.org/cgi-bin/search-biochem-index
cas: http://www.cas.org/
cmr: http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl
cog: http://www.ncbi.nlm.nih.gov/COG/
ecocyc: http://BioCyc.org/ecocyc/
enzyme: http://www.chem.qmw.ac.uk/iubmb/enzyme/
genbank: http://www.ncbi.nlm.nih.gov/Genbank/
genlight: http://piranha.techfak.uni-bielefeld.de/
gold: http://www.genomesonline.org/
kegg: http://www.genome.ad.jp/kegg/
mbgd: http://mbgd.genome.ad.jp/
mdl: http://www.mdli.com/
reactome: http://www.reactome.org/
uniprotkb: http://www.expasy.org/sprot/
whitehead: http://www-genome.wi.mit.edu/

Literature

Beckstette M, Mailänder JT, Marhöfer RJ, Sczyrba A, Ohlebusch E, Giegerich R, Selzer PM (2004) Genlight: interactive high-throughput sequence analysis and comparative genomics. J Integr Bioinform Yearbook 2004, 79–94 http://journal.imbio.de/index.php?paper_id = 8

Fleischmann RD, Adams MD, White O et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fraser CM, Gocayne JD, White O et al (1995) The minimal gene complement *Mycoplasma genitalium*. Science 270:397–403

Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC (2003) Cross-species sequence comparisons: a review of methods and available resources. Genome Res 13:1–12

Huynen M, Dandekar T, Bork P (1998) Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. FEBS Lett 426:1–5

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF et al (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34:D354–D357

NC-IUBMB (1992) Nomenclature committee of the international union of biochemistry and molecular biology, enzyme nomenclature 1992. Academic Press, Orlando, FL

Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O (2001) The comprehensive microbial resource. Nucleic Acids Res 29:123–125

Selzer PM, Brutsche S, Wiesner P, Schmid P, Müllner H (2000) Target-based drug discovery for the development of novel antiinfectives. Int J Med Microbiol 290:191–201

Uchiyama I (2006) MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. Nucleic Acids Res 35:D343–D346

Wei L, Liu Y, Dubchak I, Shon J, Park J (2002) Comparative genomics approaches to study organism similarities and differences. J Biomed Inform 35:142–150

Wheeler DL, Barrett T, Benson DA, Bryant SH et al (2007) Database resources of the national center for biotechnology. Nucleic Acids Res 35:D5–D12

9 Solutions to the Exercises

Exercise 1.1

Go to two different WWW servers that offer a free e-mail account. A directory of suppliers can be found on any web catalog (e.g., http://dir.yahoo.com/). Register at the login page. The registration is usually not complicated; only a user name, password, and the transmission of some personal data are required. After completion of the registration, e-mails may then be sent. Some providers (e.g., web.de) check the identity of new users via the postal service and activate full functionality only after this control step. This procedure is intended to diminish the abuse of free e-mail systems.

Exercise 1.2

Log into one of the two e-mail accounts and follow the hyperlink for writing new e-mails. Enter the address of your second account in the field *To* and a meaningful term in the field *Subject* of the e-mail form. You can then type the actual text into the corresponding text entry field and send the e-mail by a mouse click on the corresponding button. As email forms differ, only basic instructions can be given here; however, the forms provided are usually either self-explanatory or are provided with simple-to-understand instructions.

After sending the e-mail, in another browser window log into your second e-mail account and check whether the e-mail has arrived. It can take several minutes until the e-mail is delivered. When the e-mail has arrived, open it and send an answer using

the *Reply* button. The reply should arrive at the first e-mail account after a few minutes. If you have used the *Reply* button, the second e-mail contains the same *Subject* line as the first e-mail with additional information (e.g., *RE*:) that allows you to recognize it as an answer to the original e-mail. Both accounts can now be used for the exercises in the following chapters.

Exercise 1.3

For this and the following exercises dealing with UNIX you need an UNIX account. Many universities offer a computer pool that also contains Linux or UNIX computers. To obtain an account, consult the operator of the computer pool (usually the university's computing center). Computer pool staff may also introduce you to the UNIX environment.

Log into your account by entering your user name and password in the required fields. The login can be carried out either via a GUI, a command line, or a telnet gateway. If you use a GUI, then after a successful login open a shell. How a shell is opened differs between UNIX systems and the different GUIs. In case of difficulties consult your pool operator (computing center). If you use a telnet connection from a Windows PC to a UNIX computer then you can start in a *DOS* window or a command window with the following command: `telnet <computer.domain>`. For <*computer.domain*>, use the name of the corresponding UNIX computer, for example, `telnet unix1.somewhere.edu`

Exercise 1.4

After a successful login you are automatically in your home directory. Therefore, you can display the contents of the home directory with the following commands:

`ls`	Short form
`ls -l`	Long form
`ls $HOME`	Indicates the home directory via the environmental variable *$HOME*, which is set automatically at the login

Exercise 1.5

The path of the current directory is displayed with the following command:
`pwd`

Exercise 1.6

The command for copying under UNIX is *cp*. The following commands therefore copy the file */usr/motd* into your home directory:
```
cp /usr/motd .
cp /usr/motd $HOME
```
The dot (.) always indicates the current directory, i.e., the first command copies the respective file to your home directory only when you are actually in this directory. The second command uses the environment variable *$HOME*, which is set for every user at login and always points to the home directory. The content of this variable can be displayed on screen with the command `echo $HOME`.

Exercise 1.7

You can obtain information about options of UNIX commands from the *manual* page of the respective command. You can display the *manual* page of the copy command with the command `man cp`. The option *-i* prevents existing files with the same name from being overwritten. If such a file has to be copied, an automatic enquiry whether the file should be overwritten is generated.

Exercise 1.8

The command *mv* moves files between directories or renames files within a directory. The corresponding command is
```
mv motd current
```

Exercise 1.9

Directories are created with the command *mkdir*:
`mkdir message-of-today`

Exercise 1.10

In this exercise the command *mv* is used to move a file.
`mv current message-of-today`

The difference between the two commands for moving and renaming a file is simply that, in the first case, an existing directory is used as the destination. In the second case, a name not used so far in the current directory is given.

Exercise 1.11

Change with the command `cd message-of-today` to the directory created in Exercise 1.9. To display the file content the following commands can be used:
`more current` Displays the file content page-wise
`cat current` Displays the complete file content

Exercise 1.12

First change to your home directory with one of the following commands:

`cd ..`	After exercise 1.11 you are now in the directory *message-of-today*, i.e., one directory below your home directory. Two dots (..) denote the directory one step above the current directory, i.e., the command *cd ..* changes to the directory above
`cd`	The command *cd* without additional options or destination directories changes from any directory to the home directory of the user
`cd $HOME`	The command *cd $HOME* uses the environment variable *$HOME* as a destination directory and thus also changes to your home directory

After changing into your home directory you can create the new directory analogously to exercise 1.9 using the command `mkdir ftp-download`. Alternatively, you can create a directory by specification of the complete path from every current directory:
`mkdir $HOME/ftp-download`

Then change to the new directory with one of the following commands:

`cd ftp-download`	When your current directory is your home directory
`cd $HOME/ftp-download`	Changes from every current directory into the directory *ftp-download* in your home directory

Exercise 1.13

Before you download files make sure that such downloads are permitted on the computer you are working on.

Start the FTP connection with the command `ftp ftp.ebi.ac.uk`. At the login prompt enter the user name `ftp` or `anonymous`. As a password give your full e-mail address. Change to the EMBL database directory on the FTP server with the command `cd /pub/databanks/embl/release`. Display the directory on the server with the command `ls`. Switch the server to the ASCII transfer mode with the command `ascii` and download three arbitrary files with the extension *.dat*. The files on the EMBL server are compressed and therefore carry the extension *.dat.gz*. Omit the extension *.gz* in the download command so that the files are unpacked automatically.

`get <file>.dat`	For <*file*> use the base name, i.e., the part of the file name in front of the extension *.dat* and execute this command three times with different files

Stop the FTP session by typing either `bye` or `quit`.

Exercise 1.14

For display of the first lines of a file the command *head* is used. The corresponding command for display of the first 35 lines is therefore
`head -35 <file name>.dat`

Exercise 1.15

With the command *grep* you can search for a string of characters in a file. The standard output of the *grep* command contains the lines of the file that contain the sought-after string:

`grep contig*.dat`	**.dat* refers to all files of the current directory with the extension *.dat*
`grep -i contig*.dat`	The option *-i* turns off the distinction between upper and lower case

Exercise 1.16

First display the number of lines of all files with the command `wc -l *.dat`. The output contains the number of lines of each of the three files as well as the sum of these lines. To find out how many lines contain the term *sequence*, a *grep* command has to be combined with a *wc* command. The combination is carried out via the *pipe* character (|). The *pipe* character directs the output of the command on the left of the *pipe* to the input of the command on the right of the *pipe*:

```
grep Sequence *.dat | wc -l
```

Exercise 1.17

The command *rmdir* deletes empty directories. With the command *cd* change directly to your home directory and use the command `rmdir ftp-download`. Without having first deleted all of the contents of the directory you will receive an error message. Accordingly, you must first delete all contents in the directory *ftp-download* with the command `rm *.dat`. You can then delete the empty *ftp-download* directory from the home directory with the command `rmdir ftp-download`.

Exercise 2.1

DNA and RNA differ in the composition of their nucleotides. While in DNA deoxyribose is found as sugar residue, in RNA

this is replaced by ribose. Furthermore, in RNA, the base uracil replaces thymine. DNA is present as a complementary double-strand whereas RNA is single-stranded.

Exercise 2.2

In DNA, the base pairings A-T and C-G are seen. A purine ring is paired with a corresponding pyrimidine. Two hydrogen bonds are formed in the base pairing of A-T, whereas three such bonds are formed in the pairing of C-G.

Exercise 2.3

Genome describes all genomic DNA, transcriptome all mature mRNA, and proteome all proteins in an organism.

Exercise 2.4

The amino acid sequence of proteins is determined by the genetic code. There are 20 naturally occurring amino acids, but only four bases in the DNA to encode them. Consequently, amino acids must be encoded by combinations of bases. A base doublet of four bases allows the encoding of 4^2 or 16 amino acids and is, therefore, insufficient to code for 20 amino acids. However, a base triplet allows 4^3 or 64 combinations. Consequently, several triplets encode the same amino acid and the genetic code is, therefore, referred to as being degenerate.

Exercise 2.5

The name *CRICK* represents the amino acids cysteine, arginine, isoleucine, cysteine, and lysine. Cysteine is encoded by the base triplets UGU or UGC; arginine by CGU, CGC, CGA, or CGG; isoleucine by AUU, AUC, or AUA; and lysine by AAA or AAG. Thus, one possible genetic code encoding an amino acid sequence for

which the one-letter sequence would be *CRICK* is UGU CGU AUU UGU AAA.

Exercise 2.6

The central dogma of molecular biology was coined by Francis Crick and describes the relation between DNA, RNA, and proteins. The information of DNA is transcribed into RNA in the process of transcription, which is subsequently converted into proteins in the process of translation. This flow of information always proceeds in this direction in nature, with the exception of some RNA viruses that replicate RNA and transcribe RNA into DNA.

Exercise 2.7

Splicing describes the removal of introns from premature mRNA. The process of alternative splicing refers to varying possibilities for cutting and joining introns and exons. In this way one gene can code for several proteins. This is one explanation why there is a smaller number of genes in the human genome relative to the number of proteins.

Exercise 2.8

The Venn diagram (Fig. 2.5) displays the properties of the amino acids. The amino acids threonine and cysteine are indicated as hydrophobic, polar, and small. Isoleucine, leucine, and valine are hydrophobic and aliphatic.

Exercise 2.9

By definition the primary structure of proteins is read from the N-terminus to the C-terminus.

Exercise 2.10

Three structural building blocks are found in the secondary structure of proteins: the α-helix, the β-strand, and nonrepetitive structures or loops that connect α-helices and/or the β-strands.

Exercise 3.1

Go to the start page at NCBI (http://www.ncbi.nlm.nih.gov/). Select the term `Protein` in the pull down menu *Search* on the top left. Then enter the search terms in the desired combination into the text entry field next to the pull down menu on the right. To start the database search, click on the button `Go` on the right, next to the text entry field. Depending on the combinations of search terms different results will be obtained. For example, using `hydrolysis AND non-reducing AND arabinofuranoside AND bacillus AND subtilis`, three database records (as of May 2007) will appear: α-L-arabinofuranosidases 1 and 2 from *Bacillus subtilis* as well as the α-L-arabinofuranosidase from *Bacillus halodurans*. The latter entry also contains the words *Bacillus subtilis* in the quotation of the original article. Since a plain text search was performed, this entry is also shown in the result. However, if you restrict the term *Bacillus subtilis* to the database field *organism* then only the two entries of the α-L-arabinofuranosidases 1 and 2 from *Bacillus subtilis* will be found. In this case the query is `Bacillus subtilis[ORGN] AND terminal AND non-reducing AND arabinofuranoside`.

Exercise 3.2

To find the nucleotide sequence of the corresponding gene for ABF2_BACSU you must select the term `Nucleotide` in the pull down menu *Search* on the NCBI start page (http://www.ncbi.nlm.nih.gov/). The same search terms as in Exercise 3.1 will not

yield any entry. One can, however, find the name of the respective gene in the database record of the protein (see Exercise 3.1) in the *features* section, which is divided into subsections. In the subsections *gene* and *protein* next to the keyword */genes* =, the name of the corresponding gene `(XSA)` is mentioned.

Now enter the gene name XSA into the text entry field on the NCBI start page. Check that the term `Nucleotide` is selected in the pull down menu. In addition, combine this word with the term *Bacillus subtilis* and restrict this search term to the organism field. It should look as follows: `XSA AND Bacillus subtilis[ORGN]`. *AND* operators can be skipped – several terms are connected automatically with *AND* as long as no other operator is used.

Several database entries of the bacterium will be found, including the complete genome of *B. subtilis*. Clicking the respective hyperlink will display the complete genome of the bacterium. The information for the corresponding gene is again found in the *features* section. It is best to use the text search function of your browser to look for the gene name `XSA`. Above the gene name, next to the keywords for the subsection (*gene* and *CDS*), are found the number of the first and last base of the nucelotide sequence. If the keyword *complement* is indicated next to the numbers of the first and last base, then the gene is localized on the complementary DNA strand.

Exercise 3.3

Entrez is the database query system at NCBI. Therefore, go to the start page at http://www.ncbi.nlm.nih.gov/Entrez. Querying the system is done as in Exercise 3.1. Enter the accession number `P94552` into the text entry field and then click `Go`. Take care that `Protein` is selected in the *Search* pull down menu. Alternatively, click the hyperlink *All Databases* (highlighted in dark blue above the text entry field) on the NCBI start page to go to Entrez. Choose the query page by selecting the link `Protein`. Enter the accession number `P94552` into the text entry field and click `Go`. In both cases the entry for the protein ABF2_BACSU will be shown.

Exercise 3.4

Go to the start page of the EBI (http://www.ebi.ac.uk) and enter the AN `P94552` into the text entry field *Search Database for*. Then, in the pull down menu *in,* choose `protein sequences` and click `go`. The database record of the protein ABF2_BACSU will be found. At first sight, the entry appears different from the corresponding entry at NCBI. As mentioned earlier in Chap. 3, the standard view at the EBI WWW server for the Uniprot database, from which this record arises, is graphical. The original database record can be seen upon following the hyperlink `Viewers: Flat file` which is located in the blue bar directly above the database record. There one can also find hyperlinks to representations of the information in other formats.

Exercise 3.5

Go to the start page of EBI-SRS (http://srs.ebi.ac.uk/) and click on the tab `Library Page`. Then, in the section *Uniprot Universal Protein Resource*, check the boxes next to the databases `UniprotKB` and `UniprotKB/Swiss-Prot`. Enter the search term `ABF2_BACSU` into the text entry field above and then click `Quick Search`. Again, the same database record is found as in the previous exercise. Both the accession number P94552 and the entry name ABF2_BACSU are unique identifiers for this entry. `Quick Search` will carry out a simple full text search in the database records. It can happen that several database records are found despite the use of an explicit identifier. This will be the case if database records contain references to the desired entry in the text. In the present example, two identical entries are found. The reason why the entry is shown twice is because UniprotKB/Swiss-Prot is a subset of UniprotKB, as described in Chap. 3.

Exercise 3.6

In the graphically formatted *SwissEntry* view, the database record is divided into nine sections. In the first section, *General Information,* the two explicit identifiers *Entry name* and

Accession number are given. Also, the dates of the first entry and the most recent changes to the sequence information or annotations are listed. In the second section, *Description and origin of the Protein,* a short functional description and the name of the accompanying gene is given, as is a description of the organism from which the protein is derived, which includes the organism's taxonomical classification and a hyperlink to the *Taxonomy* database at NCBI.

In Sect. 3, *References*, the references of the respective original articles with the corresponding hyperlinks to Medline and PubMed are listed. Section 4, *Comments*, contains comments on the function of the protein as well as its affiliation to a protein family. Section 5, *Copyright*, describes any copyright information regarding the database record.

In Sect. 6, *Database Cross-references*, hyperlinks to other databases that contain entries for this protein are given. A mouse click on one of these hyperlinks queries the relevant database and displays the database record. Section 7, *Keywords,* lists the number of keywords that are given in the database record. These keywords can be used in a database query to search for records. In Sect. 8, a graphical overview of the protein features is given. In the last section, *Sequence Information,* the sequence information is displayed.

Exercise 3.7

Go to the *SwissEntry* view of the database record from Exercise 3.6 and follow one of the two hyperlinks (*Medline* or *PubMed*) of reference 1. The hyperlink provides a bibliography and a summary of the corresponding publication. The EBI SRS mirrors the PubMed database of NCBI and can be used as an alternative in case of poor accessibility to the NCBI server.

Exercise 3.8

Two genes, arf1 and arf2, of an unknown species that are homologous to the α-L-arabinofuranosidase 1 or 2 of *Bacillus*

subtilis are sought. To solve the problem, a short literature search will be performed. Return to the start page of NCBI and query PubMed by choosing `PubMed` in the pull down menu *Search* and entering the search terms into the text entry field. Using a combination of the terms `bacillus subtilis AND arabinofuranosidase` a number of publications is found. The solution to the question is hidden in the publication of Kim et al. [Kim KS, Lilburn TG, Renner MJ, Breznak JA 1998, arfI and arfII, two genes encoding α-L-arabinofuranosidases in *Cytophaga xylanolytica*. Appl Environ Microbiol 64, 1919–1923]. Arf1 and arf2 are from *Cytophaga xylanolytica* and are homologous to proteins in *Bacteroides ovatus* and *Clostridium stercorarium*.

Exercise 3.9

One can search for a publication by an author in different ways. The simplest way is to type the last name of the author into the text entry field on the NCBI start page and then click `Go`. Because a full text search is performed, all publications that contain the author's name in the text will be displayed. To restrict the search to authors only, upon typing the name, specify the database field to be searched. To do this, enter the *identifier* of the appropriate database field in square brackets (without any blanks) immediately after the search term. For this example, the search string is `Blobel[au]` and only those publications that contain the name Blobel in the author list are found. However, there are many authors with the last name Blobel, besides Günther Blobel. To retrieve only Günther Blobel's publications, `Blobel G` can be entered as a search term. Using this syntax, Entrez autonomously recognizes the search for an author's name and restricts the search accordingly. First and middle name letters must be written without any blanks behind one other (e.g., `Edison TA` for Thomas Alva Edison). To restrict the search to the author field, adding `[au]` can be used again. In the tutorial for the PubMed database (http://www.nlm.nih.gov/bsd/disted/pubmed.html), further useful information about how to restrict search results can be found.

Exercise 3.10

Go to the Prosite web page (http://www.expasy.org/prosite/) and enter the sequence (Raw or FASTA format) into the text entry field in the section, *PROSITE Tools*, via a cut & paste approach. Alternatively, the Swissprot accession number `P94552` or the Swissprot-ID `ABF2_BACSU` can be entered. Click on the button `Scan` to start the search.

Unless, the box *Exclude Patterns with a high probability of occurrence* is checked, 30 hits will be found with the following five motifs: N-glycosylation site, protein kinase C phosphorylation site, casein kinase II phosphorylation site, and N-myristoylation site (as of May 2007). All five motifs carry the warning *pattern with a high probability of occurrence*, which means that they frequently occur in sequences and might lead to an incorrect functional annotation. Next to each motif is placed a hyperlink to the respective entry in the Prosite database.

Exercise 3.11

Go to the start page of the Prints WWW server (http://bioinf.man.ac.uk/dbbrowser/PRINTS/) and follow the hyperlink *FingerPRINTScan* in the section *PRINTS search*. On the following page choose the link *FPScan* and in the text entry field enter the sequence of the entry ABF2_BACSU (Raw format) via cut & paste. After clicking `Query`, the results page should not show any significant hits for the chosen sequence. Repeat the same search with the sequence ADA1B_HUMAN in the UniprotKB/SwissProt database. To do this load the respective database record from UniprotKB/SwissProt and enter the sequence in Raw format via cut & paste. The results page should show the three highest scoring fingerprints in the first section. The two following sections list the ten best fingerprints. Each of the three highest scoring fingerprints has three links one each to the Prints database, to a graphical representation of the motif's distribution along the sequence, and to a 3D representation of the motif in the protein structure. The example sequence is a

human adrenergic G-protein-coupled receptor, which is confirmed by the three fingerprints.

Exercise 3.12

Go to the start page of the Blocks WWW server (http://blocks.fhcrc.org/) and follow the hyperlink *Blocks Searcher*. P35368 is the AN of the sequence A1AB_HUMAN from exercise 3.11. In case the corresponding browser window has been closed, download the sequence again from Swissprot and enter the sequence via cut & paste into the corresponding text entry field of the *Blocks Searcher*. Also, enter your e-mail address into the appropriate field to receive the search results by e-mail. Then submit the query by clicking on `Perform Search`. After a few minutes an e-mail in HTML format will arrive. If the e-mail program cannot display HTML, it can be saved and opened from within a browser.

The actual result of the search is found below a short description of the organization of the result page. The first section contains a summary of the search followed by a list of the possible hits. For ADA1B_HUMAN, eight possible hits should be listed. The first hit (*alpha-1B adrenergic receptor signature*) with an *E-Value* of 3.2e-123 can be regarded as statistically significant, and all seven corresponding motifs are found. The *E-Value* is a measure of the chance of finding a hit of the same quality within a random amino acid sequence. The value should be as small as possible according to its mathematical definition (see also Chap. 4). The following five hits show decreasing statistical significances. Also, not all motifs of the respective class are found in most sequences and suggests, therefore, that these receptors are part of a superfamily. The remaining hits are not statistically significant and can be disregarded. The lower part of the results page contains detailed information on each of the possible hits.

Exercise 3.13

Go to the start page of the Pfam WWW server (http://www.sanger.ac.uk/Software/Pfam/), click on the tab *Search by*, and

choose *Protein name or sequence*. Either enter the accession number (`P35368`) or ID (`ADA1B_HUMAN`) of the protein into the corresponding text entry field to query the previously compiled results. Alternatively, enter the sequence (FASTA format) into the respective text entry field via cut & paste. Start the search by clicking on the button of the respective method.

After a few seconds the result of the query is shown. The most probable hit is the Pfam protein family *7tm_1*. This designation stands for the rhodopsin family of G protein-coupled receptors with seven transmembrane helices. Both results pages contain hyperlinks to annotations of the protein family (precompiled and newly calculated).

Exercise 3.14

Go to the start page of the Interpro WWW server (http://www.ebi.ac.uk/interpro/) and follow the hyperlink *Sequence Search* on the left part of the page (highlighted in gray). Enter an e-mail address into the corresponding text field and select whether an interactive session or an e-mailed result is desired. An e-mail address is required for both operations. Now enter the sequence (FASTA format) into the text field via cut & paste or upload the sequence from a file. Start the search by clicking on `Submit Job`.
The results page displays each hit from the different member databases of Interpro in a graphical format. To see a tabular representation, click the `Table View` button. The result is identical to the results of the previous exercises, i.e., querying Interpro can frequently replace the searches of the individual databases.

Exercise 3.15

Go to the start page of the PDB database (http://www.pdb.org) and type in the search term `Bovine Rhodopsin` into the text field at the top of the page; then click `Search`. The search will return 20 hits in PDB. However, because a full text search was performed not all hits actually represent the 3D structure of bovine

rhodopsin. The structure with the highest crystallographic resolution of 2.2 Å has the PDB ID 1U19. Clicking on the title or the picture of the structure will display the database record. The *Structure Summary* lists the relevant reference, some information about the experimental method, the crystallographic unit cell, biological function, and co-crystallized ligands. Also, a ribbon representation of the asymmetric unit is shown. In the menu bar to the left (highlighted in blue) several hyperlinks to different representation methods are given together with an option for downloading the record. Detailed information regarding the individual points above can be found under the tabs *Biology & Chemistry, Material & Methods*, *Sequence Details*, and *Geometry*. Thus, the crystallization temperature of 283 K is found in the section *Material & Methods,* and the graph in the section, *Sequence Details*, details a cysteine bridge (dashed green line between two cysteine symbols that is highlighted in yellow).

Exercise 3.16

Go to the start page of Entrez (http://www.ncbi.nlm.nih.gov/entrez/) and in the *Search* menu select the database *PubChem BioAssay*. Then type `HERG channel activity` into the text field on the right and press `GO`. One *Assay* is found (as of May 2007). Click on the *Assay ID* (AID) 376. Under *Links* is indicated that 1947 compounds were tested of which 247 were active.

Exercise 3.17

Go to the start page of PubChem (http://pubchem.ncbi.nlm.nih.gov) or Entrez (http://www.ncbi.nlm.nih.gov/entrez/) and select the database *PubChem Compound* in the selection menu. Enter the word `Fenbendazole` in the text field to the right and click `GO`. Three *Compound* entries are found (as of May 2007). As can be seen in the overview, two compounds, fenbendazole and oxfendazole, differ by only a single oxygen atom. While fenbendazole has a phenylsulfide group (IUPAC: phenylsulfanyl-),

oxfendazole has a phenylsulfoxide group (IUPAC: benzenesufinyl-) at the respective site. Click on the *Compound ID* of the fenbendazole entry (CID: 3334). Under the point *Bioactivity* next to the molecular structure it is stated that Fenbendazole was tested in 47 assays (All) and active in five (Active). Under the point *Medical Subject Annotations,* some information regarding the medical use of the compound is found. Fenbendazole is used as a nematocide in veterinary medicine.

Exercise 3.18

Go to the start page of PhenomicDB (http://www.phenomicDB.de) and enter the search term `Coproporphyria` into the text field at the top of the page. In the menu *Select Organisms* choose the term *All* or restrict your search to humans with the term `Human`. Using the *Shift* and *Alt* keys in Windows, one can select several terms in this and the other selection menu, *Select data fields to show*. For the other parameters leave the standard settings untouched and click `Search`. Three genotypes and two phenotypes should result (as of May 2007). The first phenotype bears the name *Porphyria, Acute Intermittent* and is causally associated with a defect in the gene HMBS. Click the button `Orthologies` on the left next to the corresponding genotype. For *D. melanogaster,* six genotype entries in FlyBase are shown, of which the first five entries contain the keyword *lethal*. Therefore, a similar genotype–phenotype relationship also exists in the fruit fly.

Exercise 4.1

Open the *Needle* application and enter the two sequences. In the category *Required section* and in the field *Gap open penalty* enter `1.0`. Under *Matrix* in the category *Advanced section*, select `EBLOSUM62`, `EPAM250`, or `EPAM30`. Start the analysis by clicking *Run needle*. The results can be found in the file *outfile. align*. The scores for the best alignment of the two sequences will

be 31.0, 29.0, and 48.0 with the BLOSUM62, PAM250, and PAM30 matrices, respectively. The calculated alignments are quite different, however. For example, with PAM30 the introduction of several gaps is suggested. This shows that the choice of similarity matrix is important for the assessment of an alignment.

Exercise 4.2

Go to the NCBI page (http://www.ncbi.nlm.nih.gov) and select the protein database with the term `protein` under *Search* in the pull-down menu on the top left (highlighted in blue). Then enter the search string `5-hydroxytryptamine 2 A receptor` into the text field next to the pull-down menu on the right. Press *Go* next to the text field on the right. To limit the search further, you can combine the search string `homo sapiens[orgn]` with AND. Several entries for the human serotonin receptor are found. Check the box to the left of the Swissprot database record (Swissprot accession number, P28223; ID, 5 HT2A_HUMAN). Then select the data format `FASTA` in the pull-down menu above the results (highlighted in gray) and press `Send to`. Next to this button in the pull-down menu on the right, `file` should be selected. Upon these selections, the sequence information of the database record is saved directly onto the hard disk in FASTA format. In some cases, a browser dialogue window might open before saving, allowing the choice of either saving or opening the file directly with another program. Please save the file under a descriptive name onto your hard disk. Alternatively, examine the sequence in the browser and, if desired, copy and paste it into a note pad file. Other data formats can also be tried.

Exercise 4.3

Go to the NCBI-BLAST page (http://www.ncbi.nlm.nih.gov/blast) by either typing the URL or following the hyperlink `BLAST` in the links section on the start page of NCBI. Because

the sequence is that of a protein, the search must be executed using the program *blastp* against the nonredundant protein database of NCBI. Click on the hyperlink *protein blast* in the section *basic BLAST*. Then enter the sequence from exercise 4.1 via cut & paste into the text field *Enter Query Sequence*. Rather than the sequence text, the accession number (P28223) or the NCBI identifier (gi|543727) may only be used. However, this is a distinctive feature of the NCBI-BLAST server and not available for all servers on the WWW. Explanations about this text field and other fields or menu items can be found by following the respective hyperlink next to the entry field (e.g., *Enter Query Sequence*). In the section *Choose Search Set* select the nonredundant protein database called *nonredundant protein sequences (nr)*. Before starting the analysis by clicking the BLAST button please check the algorithm *blastp (protein-protein BLAST)* in the section *Program Selection*. Upon sending the query, a confirmation page is displayed that includes a multi-digit request ID. This ID allows the later retrieval of the result. Often the BLAST analysis takes a little time, e.g., due to a large number of concurrent queries of the server, but a self-updating status page will display until the analysis is finished. More than 100 hits are returned from the database (as of June 2007). The graphical overview provides a summary of the position and length of the hits with respect to the query sequence. The quality of the hits (alignment score) is color-coded.

Exercise 4.4

The program *blastn* is found by following the hyperlink *nucleotide blast* in the section *basic BLAST*. Likewise, the program *tblastx* is found by following the hyperlink *tblastx*. Execute both programs with the same nucleotide sequence (AB037513). The sequence can either be downloaded from the server, as detailed in Exercise 4.1, or its accession number entered into the text field *Enter Query Sequence* (see Exercise 4.2). In the section *Choose Search Set* check the box `Others (nr etc.)` in the category *Database*. Select

the database *Reference genomic sequences (refseq_genomic)* and limit the analysis to the organism *Drosophila melanogaster* by entering `Drosophila melanogaster` under *Organism*. Do not forget to check *blastn* in the section *Program Selection*. Using *blastn,* less than 10 database records are recovered for the human 5HT2 nucleotide sequence in the *Drosophila* genome database (as of June 2007). The quality of the hits is very low. With *tblastx*, however, more than 50 database records are found and some of these are significant. The discrepancy between the results is due to differences in how *blastn* and *tblastx* execute searches and the codon usage between the two species. While *blastn* performs a simple comparison at the nucleotide level, *tblastx* works at the protein level by first translating the query sequence into all six reading frames and then comparing these six theoretical proteins against the similarly translated *Drosophila* genome database. Because the genetic code is degenerate, an amino acid can be encoded by different codon triplets. The codon usage between *Drosophila melanogaster* and *Homo sapiens* is so different that no good agreement was found at the nucleotide level.

Exercise 4.5

Go to the *blast2seq* program at NCBI (http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi). The program can also be found under the hyperlink *Align two sequences using BLAST (bl2seq)* in the section *Specialized BLAST* on the NCBI-BLAST page. Enter the two accession numbers into the corresponding text fields in the sections *Sequence 1* and *Sequence 2*. Before starting the analysis and because protein sequences are involved, the program `blastp` must be selected (in the pull-down menu on the upper left). Press `Align`. The result shows that in the two sequences, two regions with an identity of over 40% are present. In the human serotonin receptor, the two regions lie close together whereas they are separated by more than 200 amino acids in the sequence of *Drosophila melanogaster*. The

graphical overview displays very well the spatial arrangement of these sequence regions. However, this overview should not be considered definitive as it contains little information regarding the alignment quality.

Exercise 4.6

The multiple alignment compares three protein sequences, each representing a biogenic amine receptor. Two sequences are from *Drosophila melanogaster* and the third is human. The alignment shows just a few matches between the three protein sequences because few amino acids are conserved. Those amino acids that are identical are frequently vital for protein structure and function. It is striking that for the proteins gi|543727 and gi|7296517 the conserved amino acids are present in the central region, whereas for gi|10726392 most conservation is found at the N-terminus.

Exercise 4.7

The multiple alignment makes it obvious that the sequences are very similar. The amino acids are either identical or conservatively exchanged over wide regions. Sequence gi|21245114 has an insertion of approx. 10 amino acids. Because of the high identity one can assume that they are homologous sequences. Indeed, the sequences are proteases of the cathepsin family from different species:

gi|2499874 Cathepsin L precursor *Sus scrofa* (pig)
gi|1705638 Cathepsin L precursor *Bos taurus* (cattle)
gi|19424144 Cathepsin 3 precursor *Mus musculus* (mouse)
gi|21245114 Cathepsin Q of *Rattus norvegicus* (rat)
gi|4503155 Cathepsin L preproprotein *Homo sapiens* (man)
gi|15214962 Similar to Cathepsin L preproprotein *Homo sapiens* (man)

The phylogenetic tree is shown in the lower part of the analysis. To visualize the tree in *Treeview,* save the file as a text file with the

ending .dnd by clicking the button *View DND file*. The phylogenetic tree points out the relationship between the six sequences. ClustalW calculates a close relationship between the two human sequences as well as between the sequences from mouse and rat. According to this analysis, therefore, the sequences from cattle and pig are more distantly related. If, for comparative purposes, the multiple alignment is to be performed with another algorithm called *Dialign* (http://bibiserv.techfak.uni-bielefeld.de/dialign/submission.html) then copy the *.dnd file* and view it in *Treeview*. A somewhat modified relationship between the sequences results. Again, the human sequences form one cluster; however, a closer relationship to the sequences from cattle and pig is calculated, whereas the sequences from mouse and rat are further apart. This example shows that different algorithms can produce diverse results that should be critically analyzed.

Exercise 4.8

Copy the sequence of the eukaryotic cosmid via cut & paste into the input field of the *Genscan* server (http://www.mit.edu/GENSCAN.html). If the sequence has been saved in FASTA format to the hard disk, the file can be sent to the *Genscan* server by *File upload*. Before starting the analysis, the organism from which the sequence is derived must be selected in the pull-down menu *Organism*. Because AC012088 is a human sequence `Vertebrate` must be chosen. Click `Run GENSCAN`. Optionally, a name for the sequence can be given; however, it will only be used for identification in the report. Depending on the settings (pull-down menu *Print options*) either just the proteins predicted to be present in the query sequence or the predicted proteins together with the corresponding nucleotide sequences will be displayed. It is also possible to display a graphic identifying the position of the predicted coding nucleotide sequences along the query sequence. For cosmid AC012088, two proteins are predicted; one of which is encoded by a single exon gene, meaning that the gene consists of a single exon without introns.

Exercise 5.1

Go to the home page of NCBI. Under *Molecular Databases* select the nucleotide database *dbEST*. Under *Information on the current release* click on *Number of ESTs*. More than 46 million ESTs are present in dbEST, of which almost one third (13 million) are of human or mouse origin (dbEST release 101207).

Exercise 5.2

At the start page of dbEST under *Search EST for,* enter the name `Mangifera indica`. The search will return 74 hits. In contrast, the query `Mangifera indica [ORGANISM]` yields 68 hits (dbEST release 101207) from the same database. The difference between the two queries is that for the first, all fields of a database record are searched for the term *Mangifera indica*. For example, if there were to be an entry such as "gene A similar to gene B of *Mangifera indica*," then this entry would also be returned even though the gene is from another organism. With the second query, only the field "organism" of a database record is searched. Thus only entries that are actually from *Mangifera indica* are returned.

Exercise 5.3

Under *Display* select the option *FASTA*. All sequences of the last query will now be shown in FASTA format. Save these sequences on the hard disk by selecting the option *File* under *Send to*. Give the file a name of your choice. The contents of the file can be viewed with any text editor (e.g., Notepad or Editor under Windows). Notepad is found under Start → All Programs → Accessories.

Exercise 5.4

Go to the CAP3 Sequence Assembly Program of the PBIL Institute. Copy the EST sequences of *Mangifera indica* into the

respective field and start the program by clicking `Submit`. View the results of the files *Contigs*, *Single Sequences*, and *Assembly Details*, and save the results as .txt files. Eight contigs are generated from the 68 *Mangifera indica* sequences by the sequence assembly (October 2007). One of the contigs consists of five ESTs. This shows that, within the 68 input sequences, some redundant sequences are present. In addition, many singletons are found. These do not show any similarities to other ESTs and, therefore, are not assigned to any contig.

Exercise 5.5

Analyze each of the eight contigs by copying their sequences into the respective fields of the BLASTx page at NCBI. Select the database *nonredundant protein sequences (nr)* and click BLAST. Some of the contigs show high similarity to known genes or proteins, e.g., the chloroplast enzyme, coproporphyrinogenase. However, not all contigs have reliable hits. These sequences are, therefore, new genes of unknown function.

Exercise 5.6

Go to the database search system Entrez at NCBI. Select *Search* `Nucleotide` and at *for* enter `AI590371`. Display the sequence in FASTA format by selecting the option *FASTA* under *Display*. Save the sequence to the hard disk by selecting the option *File* under *Send to*. The contents of the file can be viewed with any text editor.

Exercise 5.7

Switch over to the BLAST home page at NCBI and under *Basic BLAST* start a *blastn* search. Enter the FASTA sequence from the EST saved above via cut & paste into the box *Enter Query Sequence*. Select the database *Nucleotide collection (nr/nt)* and

click BLAST. Twenty-five hit sequences will be found in the nonredundant nucleotide database for this EST. These include 13 sequences from *Homo sapiens*, 10 from monkeys (*Pan troglodytes*, *Macaca mulatta*, *Gorilla gorilla*), and two from the domestic pig (*Sus scrofa*) (October 2007).

Exercise 5.8

Next to some human sequences (e.g., HSM805051) is found a link (letter U in a blue square) leading to the NCBI database UniGene. This hyperlink connects to the UniGene Cluster Hs.631993, which contains 31 sequences (October 2007). The nucleotide sequences that are merged in this cluster code for the protein DPCR1 (diffuse panbronchiolitis critical region). To learn more about the protein and its involvement in causing diseases, click on the hyperlink *Links* on the top right and select the database *OMIM* (*Online Mendelian Inheritance in Man*). Under the entry "diffuse panbronchiolitis" it can be read that the protein DPCR1 is involved in the onset of a chronic respiratory tract disease with the name diffuse panbronchiolitis that is exclusive to East Asian populations.

Exercise 5.9

Follow the hyperlink *Expression Profile* in the section *Gene Expression*. From the information regarding the origin of the ESTs it can be concluded that the protein is expressed in the stomach, colon, and pancreas. Moreover, the protein is found in various tumors.

Exercise 5.10

Follow the hyperlink *ProtEST* to connect to the ProtEST section of UniGene. Here, those nucleotide sequences showing hits with protein sequences are stored. For the cluster Hs.631993, one can

find eight nucleotide sequences, 4 cDNAs and 4 ESTs, all corresponding to the human protein DPCR1. If the magenta-colored bar is clicked, the alignments between the protein and translated nucleotide sequences are visualized. Four of these eight sequences, namely the cDNAs, align over the entire length of the protein. The reason why only four of the 27 ESTs of the cluster are listed in the ProtEST database is because the remaining 23 ESTs do not encode protein. They are from the untranslated regions of mRNA.

Exercise 5.11

Go to the database search system Entrez at NCBI. Under *Search* select `Protein` and at *for* enter `P01108`. Display the sequence in FASTA format by selecting the option *FASTA* under *Display*. Save the sequence on the hard disk by selecting the option *File* under *Send to*. The contents of the file can be viewed with any text editor.

Exercise 5.12

Change to the BLAST home page at NCBI and under *Basic BLAST* start a *tblastn* search. Enter the FASTA sequence from the protein c-myc saved above via cut & paste into the box *Enter Query Sequence*. Select the database *Expressed sequence tags (est)* and under *Organism* enter mouse. Then click BLAST to start the search. View the distribution of the ESTs using the graph *Distribution of BLAST Hits on the Query Sequence*. With the tBLASTn algorithm more than 100 mouse ESTs are found in the database, which show similarity to the proto-oncogene c-myc. It is striking when looking at the distribution of the ESTs that the majority of the EST sequences show high identity to either the 5′ or the 3′ end of the sequence. There are only a few ESTs that cover the middle region of the sequence. The reason for this distribution of ESTs lies in the method of EST production. ESTs are generated by end sequencing of cDNA clones.

Exercise 5.13

While the very good hits (alignment score > 200, red-colored bars), for the most part, show a 100% match with the mouse protein c-myc, those ESTs with alignment scores of 80–200 (magenta-colored bars) are only between 60 and 80% identical. This indicates that these ESTs code for a second, very similar protein. This can be verified by comparing these similar ESTs with the protein database Swissprot using the BLASTx algorithm. In this case, the best hit is the protein b-myc, which is very similar to c-myc. Thus, a similar gene can be identified by EST analysis.

Exercise 5.14

NCBI also offers an extensive collection of online textbooks. The *bookshelf* can be found at NCBI's home page under *Books* in the category *Literature Databases*. All the textbooks can be searched simultaneously with technical terms. To do this, select `Books` under *Search* in the database search system Entrez at NCBI and then enter the desired term under *for*. For example, enter `Genes and disease`. Then choose the textbook *Genes and Disease* in which information about a variety of genetically caused diseases can be found. There is a hyperlink *Phenylketonuria* under *Nutritional and Metabolic Diseases* that will lead to a page with detailed information about phenylketonuria. Here, information regarding the localization of the human phenylalanine hydroxylase gene can be found. The gene is in chromosome 12. Click on the hyperlink to the database Entrez Gene. Entrez Gene is a database that contains all available information about genes with hyperlinks to all the available databases. Accordingly, Entrez Gene is also an interesting starting point for database searches.

Exercise 5.15

Go to the NCBI database dbSNP. Under *Search by IDs* search for the *Reference Cluster* with the ID rs334. The *Single Nucleotide*

Polymorphism with the ID number rs334 is an SNP in the human genome. Information regarding the genetic variation can be found in the category *GeneView*. In the colored table, the type of mutation and its consequences are described. In this SNP, the nucleotide adenine is exchanged for a thymine in the hemoglobin beta gene. This mutation swaps the amino acid glutamate for valine. Click on the hyperlink *HBB* to be directed to the database Entrez Gene. There, more information about the gene and the disease caused by the mutation can be viewed. Individuals with this mutation suffer from sickle cell anemia, a condition most frequently found in areas that are endemic for malaria.

Exercise 6.1

Go to the home page of the PDB database (http://www.pdb.org/). The total number of solved structures is indicated on the upper edge of the page. At the time of writing (October 2007), 46,377 structures were stored in the database. The PDB database offers a RSS feed that announces on a weekly basis all newly available structures. This RSS feed can be subscribed to using a suitable web browser, e.g., Mozilla Firefox.

Exercise 6.2

In the menu on the left select the entry *Structural Genomics* and then follow the link *SG Home*. The desired information is found in the newly opened page if the hyperlink, *Worldwide Structural Genomics Initiatives,* in the section *Structural Genomics Initiatives* is followed. At present (October 2007), 25 national initiatives belong to the Structural Genomics Initiatives: 16 in North America (15 USA, 1 Canada), 8 in Europe (1 Germany, 3 England, 3 France, 1 Israel), and 1 in Asia (Japan). The national assignment seems somewhat arbitrarily chosen given that all initiatives have an international character and research groups from different continents and countries cooperate in a given initiative.

Exercise 6.3

Go to the Expasy page (http://www.expasy.org/) and enter the AN `P07801` or the ID `CHER_SALTY` into the text entry field on the upper side of the page. Press the `GO` button. The database record of the *Salmonella typhimurium* protein *chemotaxis protein methyltransferase* will be shown. Information about the tertiary structure of this protein can be found by following the hyperlinks to the PDB database in the section *cross references*. To do this, one can use a server from Expasy, go directly to the PDB database of the *Research Collaboratory for Structural Biology* (RCSB), or view the information on a server at EBI by following the corresponding hyperlinks (*ExPASy/RCSB/EBI*) next to the corresponding ID on the right. All three servers offer the possibility to download the database record and display it with a stand-alone visualization program (e.g., Rasmol, see Exercise 6.10). The structures may also be viewed using pre-computed pictures. The stored structures in the PDB database represent not only a single protein but often display complexes such as bound ligands, dimers, solvent environments, etc. It is, therefore, often the case that several database records exist in the PDB database for a single gene, e.g., CHER.

Exercise 6.4

Follow the hyperlink *RCSB* on the right next to the ID 1AF7. You will come to the *Structure Summary* of the database record 1AF7 in the RCSB PDB database. The structure summary presents an overview of the database record. In addition to the description of the stored structure and the original reference, information regarding the experimental method used for determination of the crystal structure is described (e.g., X-ray diffraction). Furthermore, the structure summary offers some references to other databases (CATH, SCOP, PFAM, etc.). Various options for displaying the 3D-structure are found on the top right in the box *Images and Visualization*. There are also a number of

program options for displaying the secondary structure. Using the KiNG-Viewer, the secondary structure will immediately be displayed as a cartoon representation with recognizable α-helices, β-sheets, and loops.

Exercise 6.5

The various representation modes of the QuickPDB Viewer can be selected in the control window on the left. To choose the secondary structure view, click *Secondary Structure* in the upper pull-down menu. Then select an amino acid from two neighboring β-sheets by double-clicking on a C_α atom in the structure window on the right. In both the structure and primary sequence windows the corresponding amino acids are colored cyan. It is obvious that neighboring amino acids in the primary sequence are not necessarily neighbors in the three-dimensional structure.

The QuickPDB Viewer offers other options to color the amino acids according to certain properties. These are the b-factor (only for structures that were examined by X-ray structure determination), the Exposure according to Lee and Richards, and the amino acids properties according to Taylor among others. These options can be selected in the two pull-down menus on the left.

The pull-down menu *Mouse* allows adjustment of mouse functions (rotate, translate, and zoom). The color of selected amino acids can be changed in the primary sequence and structure window using the colors in the pull-down menu *Color*. If the option *Stereo* is selected then two stereographic projections of the structure will be drawn. Representations of ligands, DNA, or RNA structures are not possible with the QuickPDB Viewer.

Exercise 6.6

Go to the Swissprot database of the Expasy server and search for the database record of the protein CHER_SALTY, as described

in Exercise 6.3. Display the protein sequence in FASTA format by clicking on the hyperlink *P07801 in FASTA format* in the section *Sequence information*. Then open a second browser window and go to the start page of the Expasy server. Follow the hyperlink *Secondary* on the line *Secondary and tertiary structure prediction* in the section *Tools and software packages*. From the list of servers that offers secondary structure prediction (section *Secondary structure prediction*), choose several and then enter the saved sequence of CHER_SALTY into the input field. As in the previous exercises, this can be done by cut & paste. Once the fields have been completed, start the analysis. Some servers will return the results via e-mail so take care to enter a valid e-mail address.

The predicted secondary structures agree to a greater or less degree with the actual secondary structure depending on the prediction program used. The actual secondary structure is available from the Swissprot database record. In the section *features* below the keywords *helix, strand*, and *turn*, the numbers of those amino acids that mark the start and the end of the structural elements can be found.

The mode of operation of each server influences the quality of the prediction. A distinction is made between methods that align the query sequence with those of known secondary structure and then apply this information for the prediction and methods that do not require such an alignment. If an alignment can be done with the query sequence then a significantly better prediction is to be expected.

Exercise 6.7

The protein CHER_SALTY is a methyltransferase that is not secreted. Consequently, it is not expected to contain a signal peptide. To verify this go to the SignalP server (http://www.cbs.dtu.dk/services/SignalP/) and enter the sequence into the input field either by cut & paste or by file upload. Then select `Gram negative bacteria` in the section *Organism Group*. The

remaining options do not need to be changed. Click `Submit`. The analysis takes just a few seconds before the results appear. If the other settings were left unchanged, both the text output and graphical display of the analysis are displayed. It is obvious that no signal peptide exists.

Exercise 6.8

Enter the sequence of ABPE_SALTY (AN P41780) into the input field of the SignalP server as described in Exercise 6.7. Because ABPE_SALTY is also a *Salmonella typhimurium* protein, select `Gram negative bacteria` in the section *Organism Group* and submit the job. Both neural network and HMM prediction algorithms predict the presence of a signal peptide. While the neural network predicts the cleavage site to be between amino acids 23 and 24, with HMM the cleavage site is calculated to be between amino acids 19 and 20. A third possible cleavage site is also predicted to be between amino acids 23 and 24 with a probability of approximately 18%.

Exercise 6.9

Go to the entry page of the *Center for Biological Sequence Analysis* (http://www.cbs.dtu.dk/services/) and follow the hyperlink *TMHMM*. Enter the saved amino acid sequence of the Swissprot database record Q99527 via cut & paste or by file upload into the input field of the TMHMM server. Press `Submit`. Before doing so, however, one of several output formats may be chosen. For this exercise select the format `Extensive, with graphics`. After a brief status page the results of the analysis are displayed. With the selected settings the results page contains both a text output and a graphical representation. The first few lines of the text output summarize the analysis and these are followed by lines referring to individual segments of the protein. Each segment is described numerically by the first and last amino acid. In

addition, the localization of each segment is given. The keywords, *inside, outside*, and *Tmhelix* indicate that the corresponding segment lies within the cytosol, extracellular matrix, or as a transmembrane helix within the lipid bilayer, respectively. These are also displayed in the graphical overview of the results.

The TMHMM server identifies seven transmembrane helices for the protein CML2-HUMAN. Seven transmembrane helices are typical for G protein-coupled receptors. Depending on the program used for secondary structure prediction the seven transmembrane helices also coincide with the predicted secondary structure.

Exercise 6.10

Go to the start page of the *Swiss-Model* server (http://swissmodel.expasy.org/swissmod/) and follow the hyperlink *First Approach Mode* in the section *Modeling requests* (left frame). The input field for the *First Approach Mode* is shown in the right frame. In the field *Your E-mail address*, enter a valid e-mail address in order to eventually receive the modeling result. Enter a name in the field below *Your name*. In the third field *Request title*, an optional name for the analysis can be entered – useful in the case of more than one analysis. In the text entry field *Provide a sequence or a SWISS-PROT AC code* enter the sequence by cut & paste. Alternatively, simply type the Swissprot Accession Number, `P29619`, and then click `Send Request` to start the analysis.

Within a short time, an e-mail confirming the receipt of the modeling query arrives. Depending on the load of the Swiss-Model server two further e-mails arrive, one containing the constructed model and the other a trace file that describes which sequences were recognized as homologous and which were used for the modeling process as templates.

Open the e-mail that contains the model (Subject: Swiss-Model-Model…) and save the enclosed file with the file ending *.pdb*. Then open the *Deep View – Swiss PDB viewer*. This viewer

is available for download on the Expasy server free of charge. If it is not possible to install the *Deep View – Swiss PDB viewer,* another program capable of displaying files in the Brookhaven Protein Data Bank format (PDB format), e.g., Rasmol (http://www.umass.edu/microbio/rasmol/index2.htm), may be used. If the *Deep View – Swiss PDB viewer* is not employed before starting the analysis, the output format on the start page can be changed to `Normal Mode` as not all programs can read the modified PDB format that is used as standard. The options for the output format are located at the end of the start page in the section *Results options*.

With the *Deep View - Swiss PDB viewer*, open the structures with *File – Open*. Should a message appear about missing or incorrect heteroatoms (HETATM) this can be confirmed with OK. Both the model and the underlying templates will be shown in the graphical display window. Control of the viewer is carried out via the main window as well as the control panel. Operating instructions and a tutorial can be found at http://www.expasy.org/spdbv/text/main.htm.

Exercise 7.1

Go to NCBI and choose *GEO Profiles* by selecting the term in the pull down menu on the upper left under *Search* (highlighted in blue). Then enter the search term `CG15848` into the text field on the right next to the pull down menu and press `Go` on the right next to the text field. Several entries are found for this gene. Observe the entry *GDS191*. The annotation of the gene is *Scp1*. Follow its link and learn more about the gene's function. This is particularly true for the link to the database *Flybase* (ID FBgn0020908). *Scp1* stands for *Sarcoplasmic calcium-binding protein 1* and encodes a subunit of a calcium-binding protein in *Drosophila melanogaster*.

Go back to the search results for the gene *CG15848* and the study *GDS191*. The graph (*Value/Rank Plot*) illustrates how much the gene is expressed in the different developmental stages of

the fly. Click on the graph to see the details. It is striking that *Scp1* is strongly expressed only towards the end of the pupal stage and that the expression then declines quickly in the adult stage. Thus, a logical hypothesis is that the protein plays functions in the transition from the pupa to the adult fly.

Exercise 7.2

Proceed as for Exercise 7.1 but enter the search term `CG3557 AND GDS191`. By using a combination of the two terms only one result will be shown. The *Value/Rank Plot* displays the expression profile of the gene. You can see that the gene is not expressed in the embryo or the larva. Expression only starts in the pupal stage and is eventually expressed in male adult flies only. Therefore, gene *CG3557* is a sex-specific gene. Genes with similar expression profiles can be found by clicking on the link *Profile Neighbors* above the *Value/Rank plot*. Altogether, 200 genes that are exclusively expressed in adult male flies can be found.
Go back to the NCBI page and in the pull down menu on the upper left (highlighted in blue) under *Search* select the term *Unigene*. Then enter `CG3557` into the text field next to the pull down menu on the right and click `Go`. In the category *Gene expression* under *Expression Profile,* the tissues in which *CG3557* are expressed using the EST analysis are found. Expression in the head and testes is indicated. In contrast, there is no expression in the ovaries. This is an independent confirmation of the results of the microarray analysis indicating that *CG3557* is a sex-specific gene.

Exercise 7.3

Select the entry GEO Datasets in the field *Search* and enter `GDS1399`. Then select the set GDS1399 entitled DNAadenine methyltransferase and mismatch repair mutants [*Escherichia*

coli] by clicking on the link GDS1399 record. In Table 4 assigned subsets, section Subset, and Sample Info and column Samples, the number of replicates is given. For the wild-type and dam mutant three replicates each were used. Clicking and removing the check mark will display at the end of the page those replicates and their descriptions that were used for the each strain.

Exercise 7.4

The result is 3,289 or 3,157 genes that are up- or down-regulated, respectively, in the *dam* mutant.

Exercise 7.5

In Exercise 7.4 the mean values from the corresponding three replicates of the wild-type and *dam* mutant are compared. The variation within the three replicates is not considered at this point. Therefore, this kind of calculation is statistically not supported. To obtain statistically significant results, a *t-test* is used by default for the analysis of microarray data. This well-known test asks the question whether the observed differences in the mean values of the wild-type and *dam* mutant are due to the mutation or just chance. In such a case, there is no difference in the expression of the gene between the wild-type and *dam* mutant.

In this data set there are 441 genes that are significantly up- or down-regulated in the *dam* mutant as compared to the wild-type. On the right side of the results page, all expression data in all replicates for each gene are displayed. Check at random whether the *t-test* shows genes with the desired expression profile.

Exercise 7.6

In this example, there is a significant up-regulation of 132 genes and weaker expression of 309 genes compared to wild-type.

Note the difference between the *two-tailed* (Exercise 7.5) and *one-tailed t-test* (Exercise 7.6). The former measures the change in gene expression in both the directions (up- and down-regulation), whereas the latter evaluates either one or the other direction. Therefore, the result is 132 + 309 = 441 genes.

Exercise 7.7

The tutorials of the SMD are an excellent introduction to the field of microarray analysis, including, for example, how data normalization and clustering algorithms work.

Exercise 7.8

Using the algorithms *Euclidian distance, Euclidian distance squared, Average distance*, and *Square Root of Average distance,* gene 04 does not form clusters. The conclusion, therefore, is that the expression profile of gene 04 does not correlate with any other gene. In contrast, the *Manhattan distance* algorithm indicates a cluster of genes 04 and 05. If the expression profile of genes 04 and 05 is compared, then the expression in experiments 1, 2, and 3 is very similar. Only in experiment 4 are differences found. Using the algorithm *Number of attributes with opposite sign,* gene 04 builds a cluster with genes 09 and 05. Thus, it is clear that the choice of algorithm influences the results. It is left to the scientist the choice of algorithm; there is no standard as all algorithms have advantages and disadvantages.

Exercise 7.9

The program *GenePattern* offers many functions for analysis and visualization that allow a comprehensive analysis of microarray experiments. *GenePattern* offers many individual software modules and has a clear and easy to use interface.

Exercise 7.10

The 2D gel of the HepG2 cells shows five spots that represent *HSP60*. All these spots have the same molecular weight (approx. 60 kDa) but differ in their pI values. The differences are probably due to post-translational modifications, e.g., phosphorylation, which affect the pI value. The phosphate group changes the charge of the protein and thus also the pI value. *HSP60* can be phosphorylated at several sites simultaneously, which explains why there are several spots for *HSP60*.

Exercise 7.11

The 2D gel of liver tissue reveals only three spots for *HSP60*, unlike HepG2 cells. Thus, there seems to be less post-translational modification of *HSP60* in the liver compared to HepG2 cells.

Exercise 7.12

In the 2D gel of secreted proteins from HepG2 cells there are no spots for *HSP60*. This indicates that the protein is not secreted.

Exercise 7.13

The protein is human *S100-A4*. This is an abbreviation for *S100 calcium-binding protein A4*. The protein has a molecular weight of 14.4 kDa and two alternative names, *CAPL* and *MTS1*.

Exercise 7.14

To identify the proteins three methods were employed. (1) *Gel matching* whereby existing 2D gels are compared. If spots with the same molecular weight or pI value are found, and the proteins are already known, then it is assumed that these proteins

are in fact identical. (2) *Immunodetection* whereby specific antibodies are used for unequivocal identification. (3) *Microsequencing* whereby protein spots are excised from the gel, eluted from the gel slices, trypsin-digested, and then sequenced.

Exercise 7.15

Microsequencing identified the sequence L V K K Q T Y H I.

Exercise 7.16

Enter the accession number `P12931` into the search field, select the enzyme *Trypsin*, and choose *1000* under *Display the peptide with a mass bigger than*. After clicking `Perform`, 21 peptides with a mass >1.000 Dalton are shown. They are all the results of a tryptic digest of the human protein kinase src. The largest peptide has a mass of 5,072 Daltons.

Exercise 7.17

In the top panel open *All* and under *Sample* enter the molecular mass *Mw* `38000`, the iso-electric point *pI* `7.0` and, in *Peak list*, the masses of the identified peptides (`1433 1422 1088 1030`). In the category *Protein* select the database *UniProt/SwissProt* and as taxonomy *Bos taurus*. Under *Peptide* the enzyme *trypsin* is selected by default. Under *Thresholds* select the allowed measurement deviation (*Spectrometer shift max* ±*0.5*). Execute the analysis by clicking `Submit`. The program will identify a bovine protein in the database that after in silico digestion yields four peptides with the desired masses. It is the protein *annexin II* with the accession number P04272. Because the four peptides, the molecular mass, and the pI are in good agreement, the protein isolated from the polyacrylamide gel can be identified.

Exercise 7.18

After clicking *enter as guest*, follow the hyperlink *Browse all complexes*. The *YEAST protein complex database* contains 232 multi-protein complexes from *Saccharomyces cerevisiae*. Complex 116 consists of 24 proteins. The function of the complex is categorized as being involved in transcription/DNA maintenance/chromatin structure.

Exercise 7.19

The protein *NHP10* is not only present in complex 116 but also complex 137. The function of complex 137 also falls into the category transcription/DNA maintenance/chromatin structure.

Exercise 8.1

Go to the GOLD Genomes Online Database (http://www.genomesonline.org/gold.cgi). At the time of writing (October 2007), the first table lists 3,030 genome sequencing projects and 660 genomes that are completed. The buttons in the fields of the table lead to listings of the corresponding genome sequencing projects that contain further information regarding the individual projects. Similar statistics and listings can be found at TIGR (http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl) and at the NCBI microbial genomes resources (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html).

Exercise 8.2

Go to the KEGG home page (http://www.genome.ad.jp/kegg/) and follow the hyperlink *PATHWAY* to the *PATHWAY* database. To find a metabolic pathway follow the hyperlink *1. Metabolism*. The glycolysis/gluconeogenesis metabolism is part of carbohydrate

metabolism and, therefore, the corresponding metabolic chart is found in the section *Carbohydrate Metabolism*. Select the hyperlink *Glycolysis/Gluconeogenesis* to display the metabolic map. Alternatively, the map can be found by following the hyperlink *1.1 Carbohydrate Metabolism*. In the carbohydrate metabolism map, click on the hyperlink *Glycolysis/Gluconeogenesis*.

Exercise 8.3

The entry *Pyruvate* is in the lower third of the metabolic map and *L-Lactate* is to the right of it. A double arrow connects the two entries. An enzyme (EC 1.1.1.27) is written on this arrow that catalyzes the conversion of L-lactate to pyruvate. By clicking on the EC number the corresponding enzyme entry can be found. EC 1.1.1.27 is an oxidoreductase (L-lactate dehydrogenase).

To see whether this conversion takes place in humans go back to the KEGG home page (http://www.genome.ad.jp/kegg/) and follow the hyperlink *KEGG Organisms*. On the new page the organism of interest can be found by clicking on the appropriate abbreviation. For *Homo sapiens* the abbreviation is *hsa*. Follow the hyperlink *Pathway maps* on the next page and then select the chart for the glycolysis/gluconeogenesis metabolism. The same metabolic map will be displayed and those enzymes, including enzyme EC 1.1.1.27, that function in the human metabolic pathway are marked in green.

Instead of selecting the desired organism in the organism table, the appropriate organism abbreviation can be entered directly on the KEGG start page into the text field on the right next to the hyperlink *KEGG Organisms*. Click Go. If the organism's abbreviation is not known, a search function on the left of the text field can be opened using the button `Organism`. Enter the first letters of the respective name, e.g., `Sacch` for *Saccharomyces cerevisiae* in the search field and then choose the corresponding organism below this field by clicking `Select`. The abbreviation of the name will be entered into the text field.

Go to the specific metabolic map of *Saccharomyces cerevisiae* by any of the two above-described approaches. L-lactate

dehydrogenase is no longer marked in green on this metabolic chart, i.e., *S. cerevisiae* does not have the gene for this protein and, therefore, does not use this metabolic pathway.

Exercise 8.4

Follow the hyperlink to EC 1.1.1.27 in the metabolic map from Exercise 8.3 (glycolysis/gluconeogenesis metabolism in humans). The entries for LDHA, LDHB, LDHC, LDHAL6A, and LDHAL6B from the GENES database are shown. This means that in species-specific metabolic charts, the hyperlinks of these enzymes lead to individual records in the GENES database. In the reference map, however, the enzyme hyperlinks lead to entries in the LIGAND database.

Exercise 8.5

Go to the KEGG home page (http://www.genome.ad.jp/kegg/) and open the chart for the human glycolysis/gluconeogenesis metabolism (organism abbr.: hsa) as described in Exercise 8.3. In a second browser window, display the species-specific metabolic pathway of *Helicobacter pylori* (organism abbr.: hpy). A direct comparison of the two pathways indicates that enzymes EC2.7.1.11 and EC2.7.1.40 are missing within *H. pylori*. Both proteins are kinases, i.e., they are phosphate group-transferring enzymes. Information about the function of both can be found by following the respective hyperlink (EC number) to the LIGAND database. Phosphofructokinase (EC 2.7.1.11) catalyzes the conversion of fructose-6-phosphate to fructose-1.6-bisphosphate via an irreversible reaction. In a subsequent irreversible reaction, pyruvate kinase (EC 2.7.1.40) catalyzes the conversion of phosphoenolpyruvate to pyruvate – the last step of glycolysis. By directly comparing both metabolic maps the conclusion is that *H. pylori* lacks two important enzymes of glycolysis. Consequently, *H. pylori* has an incomplete glycolysis pathway. This is not difficult to understand when one considers that the natural

habitat of *H. pylori* is the acidic environment of the stomach of mammals. Therefore, if pyruvate were to be produced, a further acid burden would result. Consequently, the bacterium does not utilize this metabolic step.

Exercise 8.6

Go to the BLAST-Homepage of the NCBI and follow the hyperlink Microbes in the BLAST Assembled Genomes section. The resulting special BLAST page can be used to BLAST against microbial genomes. Enter the accession number `Q9ZK41` into the text field and choose the type of Query and Database. Because a BLASTP is desired, choose Protein for Query and Database. Alternatively, the program blastp can be chosen in the pull-down menu *Blast-program*. Check the desired organisms in the taxonomy tree. Click on the green plus button in front of the category name if all organisms in a category for which protein sequences are available are desired (marked by the letter P in front of the organism name). To do this, the correct types of Query and Database must have been chosen already. Start the analysis by clicking on `BLAST` at the top or bottom of the page. On the following page, simply press `View report` to display the BLAST-result page. The page is self-refreshing until the analysis is finished.

Relevant database hits for the following organisms are found: *Helicobacter pylori, Campylobacter jejuni, Thiomicrospira denitrificans*, and *Campylobacter coli*. Obviously, the sequence with the accession number Q9ZK41 is the glucose/galactose transporter of *H. pylori* that is encoded by the gene *gluP*. *Campylobacter jejuni* possesses a homologous protein that was annotated by its homology to COG0738 (*fucose permease*). No homologous proteins were found in the genera *Staphylococcus* and *Streptococcus*.

Exercise 8.7

Go to the start page of the Comprehensive Microbial Resource (http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl)

and from there to *Multi-Genome Homology* in the menu bar *Comparative Tools → Protein Homology Tools*. Under *Select a Reference Genome* choose the genome of *H. pylori 26695*. Then, under *Make your Comparison Genome,* select the three *E. coli* genomes one at a time (*E. coli K12-MG1655, E. coli O157:H7 EDL933*, and *E. coli O157:H7 VT2-Sakai*). After each choice press the button `Add>>`. This way the selected genomes will be added to the *Final Comparison Genome List*. Once all three *E. coli* genomes have been selected, choose *Minimum Percent Similarity* and enter the number `30` in the text box on the right. Then start the analysis by clicking `Submit`.

The results of the analysis are graphically displayed and represent the four selected genomes as concentric rings. The outermost circle corresponds to the reference genome and the inner rings to the respective comparison genomes. Upon moving the mouse over the graph, the name of the corresponding genome is indicated on the right next to the graph. Furthermore, only homologous sequences are shown on the reference genome. If a gene is present only in one or two of the three selected *E. coli* genomes, then the corresponding color-coded part of the gene is not displayed for the reference genome.

The table *Summary statistics for Reference DNA molecule* indicates that *H. pylori 26695* has 800 genes that are not present in any of the three *E. coli* genomes. These genes, therefore, could be subjected to more in-depth analysis in order to identify potential target proteins that might be suitable for the development of a selective drug against *H. pylori*.

Exercise 8.8

Go to the home page of the MBGD database (http://mbgd.genome.ad.jp/) and follow the hyperlink *Create/View Ortholog Table* on the left side of the page. Choose the desired organisms in the selection box. Make sure to press the <*Ctrl*> button while selecting the second and third organisms. Then click `Create/View Cluster Table`. The calculation of the cluster can take a few

minutes. While the analysis is running, a self-refreshing HTML page is shown. Once the calculation is completed the cluster table is displayed.

Exercise 8.9

The cluster table of Exercise 8.8 contains the phylogenetic profiles of the selected organisms. The columns of the table represent the organisms, the rows the individual profiles. If an organism contributes proteins to a cluster, a mark (green block) is drawn in the table at the position of the organism. Therefore, the sought-after phylogenetic pattern corresponds to a closed green bar because all selected organisms contribute proteins to the cluster. A total of 464 clusters fit this phylogenetic pattern. Click on the colored bar to the right of the phylogenetic pattern to display the individual clusters. The clusters actually shown depend on that region of the bar clicked. The colors correspond to functional categories. To view the first cluster, click the first region of the color bar (purple). The purple color indicates that this cluster contains proteins of the functional category amino acid biosynthesis. The legend to the color code can be found by following the hyperlink *function category* on the *Cluster Table* page.

Exercise 8.10

Go to the start page of the MBGD database (http://mbgd.genome.ad.jp/). If the selected organisms are not highlighted in the *Organism selection* window, press *Reload/Refresh*. Then enter the search term `fructokinase` into the text entry field to the left of the *Organism selection* window and press `<RETURN>`. Three entries are found in the cluster table.

Glossary

@ – In 1972, the engineer Ray Tomlinson wrote the first e-mail program (Bolt Beranek and Newman, Inc.). He needed a character that would separate the first part of an e-mail address from the host's designation or domain. In addition, the required character should not appear in any name. Tomlinson decided to use the @ character on the keyboard of his teletypewriter model 33. The character had been known from writings and prints of the baroque period (17th century) and represented the Latin *ad*. Today @ is read as *at* and is an essential component of every e-mail address

Accession Number – A unique identification number for a database record in a sequence database. Accession numbers are static, i.e., they do not change even after database updates

Account – Access rights on a computer system

ADSL – Asynchronous Digital Subscriber Line. A DSL technology that possesses higher bandwidth for downloads from the net compared to uploads

Affinity chromatography – A technique for the purification of proteins that makes use of the affinity of a protein for a distinct substrate or ligand (e.g., antibodies for antigens)

Algorithm – Derived from Al-Khowarizmi (Arab mathematician, 825 AD). A logical sequence of steps for solving mathematical problems

Alias – Alias or alias names are names that stand for another name. For example, under UNIX operating systems, complicated command lines can be accessed more rapidly via aliases. Complicated user identifications, e-mail addresses, etc. are more easily remembered by using short alias names. Example:

the command *mount -t msdos/dev/fd0/floppy* can be executed by typing the command *diskmount* upon entering once into a system file: *alias diskmount mount -t msdos/dev/fd0/floppy*. See also Mail alias

Alignment – Adjustment of two (pairwise alignment) or several (multiple alignment) sequences so that similar or identical amino acids or nucleotides are arranged vertically to produce matches

Alpha (α) helix – An ordered folding pattern of the secondary structure of proteins. The α-helix displays a pitch of 0.54 nm with 3.6 amino acid residues per turn

Alternative splicing – Generation of different mRNA transcripts from one pre-RNA using different splice sites

Amino acids – The building blocks of proteins. Proteins are built from the 20 naturally occurring amino acids

Analogy – A classification according to common features of the structure and/or the function considered to be essential (e.g., proteins that have similar folds or functional centers yet cannot be grouped by a common ancestral protein; e.g., head and mouthparts of arthropods, such as insects, compared to those of vertebrates). See also Homology, Character, Relationship, Phylogeny

Annotation – Information on possible relationships and the derivation of possible biological functions

Antigens – Compounds that activate the immune system to generate antibodies. An antigen, for example, is a surface protein of a bacterium

Antibodies – Antibodies are proteins (also referred to as immunoglobulins) that bind to an antigen and consequently allowing cells of the immune system to neutralize the antigen

Applet – A small computer program that is downloaded via HTML from a server and executed on one's own computer by HTML. Applets are usually written in the programming language JAVA

Array – See Microarray

Array express – A database at EBI where the results of microarray experiments can be stored and are accessible at any time

ASCII – American Standard Code for Information Interchange. Code table for the encoding of 128 accent-free characters (a-z, A-Z, 0–9 as well as special and control characters). ASCII files are often referred to as *Plain text* or *Flat file*

Assembly – See Sequence assembly

Bases – Basic building blocks of DNA and RNA. The sequence of the bases (nucleotide sequence) forms the blueprint for the gene product

Base pair – Pairing between two bases on opposite nucleotide strands of DNA or RNA. In DNA, adenine pairs with thymine and, in RNA, with uracil; cytosine always pairs with guanine

Beta (β) sheet – Regular folding pattern of the secondary structure of proteins. β-strands are built from two amino acid chains. The peptide chains can be orientated either in the same or opposite direction leading to parallel or antiparallel sheets. Successive amino acid residues are on opposite sides of the plane of the sheet with a repetition unit of two residues and a distance of 0.7 nm

Binary file – A file that includes nonreadable text, such as, for example, executable programs, videos, and sound files

Biochip – See Oligonucleotide array

Bioinformatics (applied) – Application of informatic and mathematical concepts to large sets of biological data in order to accelerate and improve biological research. Applied bioinformatics is important in the fields of molecular biology, biochemistry, chemistry, and medicine

Bioinformatics (theoretical) – The development of computer-based databases, algorithms, and programs to accelerate and improve biological research. Theoretical bioinformatics is important in the field of computer science

BLAST – Basic Local Alignment Search Tool. Heuristic algorithm to search for sequences in sequence databases

BLOSUM – *BLO*cks *SU*bstitution *M*atrix, substitution matrices for the alignment of protein sequences. BLOSUM matrices were introduced by Henikoff and Henikoff in 1992 and are well suited to the alignment of remotely related protein sequences. BLOSUM matrices are characterized by an affixed

number that indicates the sequence identity of the sequences used to derive the matrix. Accordingly, the BLOSUM62 matrix is based on the observed substitution patterns of sequences that share 62% identity

Broad-spectrum antibiotic – A substance that kills many microorganisms, such as bacteria, and for which the mode of action is based on an ubiquitous target (e.g., a protein)

CAP3 – A sequence assembly program based on the Smith–Waterman algorithm

CATH – A structural protein database that hierarchically classifies protein domains into four groups: Class (C), Architecture (A), Topology (T), and Homologous superfamily (H)

cDNA – Complementary DNA; a DNA that is produced from mRNA as template with the help of the viral enzyme, reverse transcriptase. Like mRNA cDNA does not have introns

cDNA array – DNA microarray consisting of in vitro-amplified cDNA that is spotted onto a support material

cDNA library – A cDNA library contains all cDNA transcripts of a cell, tissue, or whole organism. Unlike a genomic library, it contains only coding DNA

CDS – See Coding sequence

Central dogma of molecular biology – DNA is transcribed in the process of transcription to mRNA, which is then translated into proteins during translation (Francis Crick, 1957)

CERN – Conceil Européen pour la Recherche Nucléaire, or Organisation Européenne pour la Recherche Nucléaire. European organization for nuclear research based in Geneva and with a research station nearby Meyrin. The development of the WWW started at CERN to organize research data in such a way as to make it available to researchers in other countries

Character – A property of a protein or species (motif, structure, function, morphology, physiological process, etc.) that distinguishes it from other proteins or species. Phylogenetic research always deals with either character pairs or character strings that can be resolved into character pairs. Such character pairs can be differentiated into relatively ancestral (plesiomorph) or relatively derived (apomorph)

character partners. See also Analogy, Homology, Relationship, and Phylogeny

Chemoinformatics – Analogous to the term bioinformatics, chemoinformatics includes all scientific disciplines that apply concepts from mathematics and informatics to large data sets facilitating chemical research

Chromatography – A method for the separation of substance mixtures involving stationary and mobile phases. The term chromatography was coined by the Russian botanist Michail S. Tswett (1872–1919) who used the method to isolate pigments from plant extracts

CIB – Center for Information Biology. Japanese bioinformatics institute that manages the nucleotide database DDBJ

Classical proteomics – Classical proteomics deals with the identification and quantification of proteins in cell lysates

Client – Computer program that communicates with a server. Browsers are classical clients that communicate with Web servers

Clone – A population of genetically identical organisms, cells, or bacteria that have a common origin. For example, a bacterial clone in a cDNA library consists of many thousand bacteria that all possess the same cloned DNA sequence on a plasmid. Another meaning for the term clone refers to a group of recombinant DNA molecules that are descended from an initial molecule (DNA clone)

Cloning – A specific DNA sequence is inserted into plasmids that serve as vectors. The DNA sequence, as part of the plasmid, is then propagated by transformation into bacteria

Cloning vector – See Vector

Cluster – A group that contains similar objects. Examples are EST sequences that are clustered due to sequence similarities or genes that are assigned to a cluster due to similar expression profiles

Cluster (comp.) – A number of computers that are merged into a single (private) network and are regarded as a single computer

Clustering – The process of grouping together objects into single clusters due to concurrences

Coding sequence – Part of the DNA that is transcribed into mRNA during transcription and then translated into protein

Codon – A set of three nucleotides (base triplet) of DNA or RNA that code for one of the 20 natural amino acids

Codon usage – Species-specific use of the different possible codons that encode amino acids

Command line – Lowest (text-based) level for communication between user and computer

Communication protocol – A number of fixed rules to communicate between computer programs. The communication between computers on the Internet is based on the communication protocol TCP/IP (Transmission Control Protocol/Internet Protocol)

Comparative genomics – Simultaneous comparison of two or more genomes with the aim of identifying similarities and differences between those genomes

Compiling – Assembly of a new complete database from a number of individual databases

Computer model – A mathematical model to simulate a biological system that allows the prediction of certain properties (e.g., the concentration of metabolites at a given time) and, because of its complexity, can only be solved with the aid of a computer

Consensus sequence – A single common DNA or protein sequence derived from a multiple alignment. Each position of the consensus sequence comprises that nucleotide or amino acid that occurs most often in the sequence alignment

Conserved sequence – Part of a DNA or protein sequence that has remained constant during evolution

Content provider – See Online services

Contig – Contiguous segment of a genome that was generated by joining overlapping sequences

CORBA – Common Object Request Broker Architecture. Industry standard that allows the connection of different objects and programs regardless of the programming language, machine architecture or whereabouts of the computers

Database – A collection of data that is organized to allow easy access to its content

dbEST – Publicly accessible database at NCBI that stores Expressed Sequence Tags (EST).

dbGSS – Publicly accessible database at NCBI that stores Genome Survey Sequences (GSS)

dbSNP – Publicly accessible database at NCBI that stores short genetic variations such as single nucleotide polymorphisms (SNPs)

DDBJ – DNA Data Bank of Japan. Together with the databases EMBL and GenBank, DDBJ forms the International Nucleotide Sequence Database

Deletion – A mutation in a nucleotide sequence where single nucleotides or whole regions are missing compared to the original sequence

DNA – Deoxyribonucleic acid. DNA carries the genetic information. It consists of a pair of nucleotide strands that wind around a common axis to form a double helix. The pairing of the nucleotide strands is via hydrogen bonds between specific base pairs

DNA denaturation – The conversion of double-stranded nucleotide sequences into single-stranded sequences. The hydrogen bonds between the single strands can be destroyed by strong heating, for example. The generation of single-stranded nucleotide sequences is a prerequisite to hybridize with complementary single-stranded sequences, e.g., in the assembly of a DNA microarray

DNA microarray – Miniaturized technology based on the method of nucleic acid hybridization. With DNA microarrays, gene expression profiles of cells can be analyzed, for example. One differentiates between oligonucleotide and cDNA microarrays

DNA sequence – Sequence of base pairs in a DNA fragment, gene, chromosome, or a complete genome

DNA sequencing – A method to determine the nucleotide sequence of a DNA molecule. A common method is the

Dideoxy-Chain-Termination method that was published by Frederick Sanger in 1977

Docking – Computer-assisted fitting of a ligand into the binding pocket of a protein

Domain (biol.) – A delimited functional unit of a protein with its own discrete folding. The complete functionality of a protein results from the combination of different domains

Domain (comp.) – Computer networks are subdivided into logical sections (domains). This division becomes obvious in the full-qualified domain name of the computer, e.g., ftp.ncbi.nih.gov. In this case the top-level domain is the domain .gov (government). Other known domains on the WWW are .com (private enterprise), .edu (educational institutions), .net (administrative networks), .de (geographical domain for Germany)

Download – Loading of a file from a remote server onto a local computer. The download can be carried out via FTP or HTTP from the WWW using a browser, for example

DSL – Digital Subscriber Line. Digital technology for the transmission of data that permits transfer rates using conventional copper lines and that are up to 100 times faster than ISDN

Dynamic methods – Breakdown of a problem into sub-problems and reuse of the solutions for sub-problems. To solve a problem of size n, all sub-problems of size 1, 2,…, $n - 1$ are solved. The solutions are saved in a table from which the solution for n is derived. Dynamic methods are usually very exact; however, they can be very slow (e.g., the Smith–Waterman algorithm)

EBI – European Bioinformatics Institute that is part of EMBL and is localized in Hinxton near Cambridge, GB

E-cell project – International research project with the aim of simulating biological phenomena on the computer and developing tools, technologies, and programs for the computational simulation of a complete cell

Edman degradation – Method to determine the sequence of polypeptides

EMBL – European Molecular Biology Laboratory. It was founded in 1974 and is funded by 16 European countries and Israel. Its

headquarters are in Heidelberg, Germany. Other sites are in Hamburg (D), Grenoble (F), Hinxton (GB), and Monterotondo (I)

ENTREZ – A general query system for all available databases at NCBI

Enzyme – A protein that works as a catalyst, i.e., to reduce the activation energy of a reaction and thereby influence reaction rate. Catalysts do not change the direction of a reaction

Epitope – Part of a protein bound by antibody

ESI – Electrospray ionization, in mass spectrometry a method to generate ions. Because of the gentle ionization of the analyte molecule the method is particularly suitable for the analysis of biomolecules

EST – Expressed Sequence Tag. Partial sequence of a cDNA clone

Ethernet – Technology to network computers

Eukaryotes – Organisms in which cells have a nucleus and other subcellular compartments, such as mitochondria. All organisms are eukaryotic with the exception of viruses, bacteria, cyanobacteria, and archaebacteria

Exon – Coding region of a eukaryotic gene. Exons may be separated from one another by noncoding introns

ExPASY – Expert Protein Analysis System. A WWW server of the Swiss Institute of Bioinformatics to analyze protein sequences. The Expasy server hosts the Swissprot database, amongst others

Expression profiling – Determination of the gene expression pattern of a cell or tissue with the aid of DNA microarrays

FASTA – Heuristic algorithm to search for sequences in databases

FASTA format – Simple database format to store sequence data. The FASTA format consists of a single header line that starts with the character >. It is directly followed (without a blank) by an identifier and optionally (separated by a blank) a short description. The subsequent lines contain the sequence information

Fingerprint – A number of sequence motifs that were derived from multiple alignments and form a characteristic signature for members of a protein family

Firewall – A mechanism to protect computers against attacks from the Internet. The firewall permits access of computers to the Internet yet blocks access from the Internet

Flat file – A flat file contains data that do not have any structural relationship to one other. Most biological databases consist of flat files

Frameshift – A deletion or insertion in a DNA sequence that leads to a shift in the reading frame of all subsequent codons. In nature, frameshifts can arise by accidental mutations. In DNA sequences, frameshifts are frequently observed due to reading errors by sequencing machines

FTP – File Transfer Protocol. Communication protocol for the download/upload of files between two computers

Functional genomics – Parallel analysis of genes of a given organism to identify the function of gene products. Methods used to identify gene function are, for example, DNA microarrays, Serial Analysis of Gene Expression (SAGE), and proteomics

Functional proteomics – The aim of functional proteomics is to identify the functions of proteins. An important aspect of functional proteomics is the identification of protein–protein interactions

Fusion protein – The product of a hybrid gene. Such hybrid genes are frequently produced experimentally so that the resulting fusion proteins can be purified or detected

Gap – A gap in a sequence alignment that arises from insertions or deletions

GCG – Genetics Computer Group. A number of bioinformatics programs to analyze DNA and protein sequences. GCG was founded in 1982 as a service of the University of Wisconsin and is, therefore, also known under the name Wisconsin Package. GCG became a commercial software in 1990 and is distributed worldwide by Accelrys Inc.

Gene – A DNA segment that contains genetic information encoding protein. A gene comprises several units, including exons and introns and flanking regions that mainly serve in

gene regulation. Genes are also described as the functional units of a genome

GenBank – A database localized at NCBI in which nucleotide sequences are stored

GeneChip – See Oligonucleotide array

Genetic code – Key for the translation of genetic information into proteins. Three bases (base triplet) encode an amino acid. Different base triplets can code for the same amino acid (degenerate code). With a few exceptions, e.g., in mitochondria or ciliates, the genetic code is universal for all living organisms

Gene expression – Process in which the information encoded by a gene is translated into functional structures. Expressed genes are those that are transcribed into RNA and then translated into protein, or those that are only transcribed into RNA (without translation)

Gene family – A group of related genes that result in similar protein products

Genome – All the genetic information of an organism. The genome represents the sum of all genes, those parts of the DNA that influence the expression of the genetic information and those areas yet to be functionally characterized

Genomics – Research field that deals with the analysis of the complete genome of an organism

Genomic library – A gene bank that consists of many clones with genomic DNA. Unlike a cDNA library, a genomic library also contains noncoding DNA, such as the gene introns, and DNA regions without genes

Genotype – Entirety of all genetically determined characteristics of an individual

Genotyping – Experimental determination of the genotype of an individual

GEO – Gene Expression Omnibus. A database at NCBI that stores a variety of gene expression data and can be queried. This includes the results of DNA microarray and SAGE experiments

Global alignment – An alignment over the entire length of two sequences

Glycosylation – Post-translational modification whereby sugar residues (under the release of water) are linked onto proteins after translation is completed. Other organic molecules such as lipids can also become glycosylated

Gopher – Internet service for the exchange of information. The gopher service can be considered as a predecessor of the WWW

GSS – Genome Survey Sequences. Like EST sequences, GSS are generated by single-pass sequencing of the end regions of DNA clones. In contrast to ESTs, to generate GSS sequences, clones from genomic libraries are sequenced. Therefore, GSS sequences can also contain regions that lie outside of genes

GUI – Graphical User Interface to operate a computer (e.g., Windows, X Window, etc.)

Heuristic methods – Procedures that are based on a sequence of approximations. Heuristic methods try to find optimal or at least nearly optimal solutions in an exponentially large space of solutions by problem-specific information. Though fast, heuristic methods may not find all possible solutions (e.g., BLAST algorithm)

HGVbase – A database at the Karolinska Institute in Sweden that records information regarding variations in the humane genome. HGVbase will be developed into a genotype/phenotype database in the near future

Hidden Markov Models – The Hidden Markov model (HMM) is named after the Russian mathematician A. A. Markov (1856–1922). It is a stochastic process (conjecturing, dependent on randomness) in which parameters that obey the system equations are not directly observable but can only be observed by derived quantities. HMMs consist of states, possible transitions between these states, and the state transition probabilities. In a specific state a result can be generated by taking into consideration all probabilities. The results, not the states, are visible to the external observer, i.e. the states are hidden. HMMs are used for the derivation of profiles from multiple protein alignments to identify new proteins, for example

HomoloGene – NCBI database of homologous proteins from different species

Homology – A classification based on the phylogenetic origin of structures. Characters that were inherited either unchanged or changed from common ancestors (e.g., specific kinases of mice and humans; or extremities of mice and humans) are considered homologous. See also Analogy, Character, Relationship, and Phylogeny

Homology map – Tabular overview of syntenic regions from the chromosomes of two species

Homology modeling – Development of a three-dimensional computer model (*in silico*) of a protein structure using, as a template, the structure of a similar protein that has been solved experimentally by X-ray analysis

Host (comp.) – Networked computer that permits access and provides different services or programs to computers to which it is connected. Also: The computer (or server) that a user logs onto to access the Internet. Also: Any computer on the Internet that can be accessed via an IP address

HTML – Hypertext Markup Language. Syntax for formatting documents on the WWW to allow their display by browser applications according to the WWW standard

HTTP – Hypertext Transportation Protocol. Communication protocol of the WWW. Specification of the communication between WWW servers and their users, such as browsers. Browsers can recognize HTML documents and display the contents with the help of this protocol

HTTPS – Hypertext Transfer Protocol Security. With HTTPS encrypted data are transferred on the WWW; e.g., commercial banks use this protocol

Hybridization – Pairing of two complementary and single-stranded DNA molecules to generate a double-stranded molecule through the formation of hydrogen bonds between complementary bases. For instance, hybridization is used to isolate complementary sequences in cDNA libraries

Hyperlink – A cross-reference on a HTML page that connects one document with another document on the WWW

Hypertext – Text that contains embedded cross-references (hyperlinks)

Identity – The number of identical sequence positions in an alignment

IMAGE consortium – Integrated Molecular Analysis of Genomes and their Expression. A consortium of academic groups that provides high-quality cDNA libraries to all interested

Immobilization – The covalent attachment of nucleic acids to solid supports. DNA can be immobilized onto nylon membranes by UV irradiation, for example

In silico – In silicon. Silicon is the material computer chips consist of. It means an experiment simulated on the computer

In vitro – Latin: with/in the glass; outside a living organism. Denotes the location where an experiment is performed or a compound tested, e.g., a drug

In vivo – Latin: with/in the living; within (the body of) a living organism. Denotes the location where an experiment is performed or a compound tested, e.g., a drug

Indexing – Process describing the contents of databases with the help of descriptors, informative keywords, catchphrases or text and, thus, allow for the efficient query of documents within a database

Insertion – Incorporation of single nucleotides or whole nucleotide blocks into a DNA strand

Interactome – The entirety of all interactions in a cell

Interactomics – Bioinformatics discipline that deals with the study of interactomes, i.e., the interaction of all proteins and other molecules in a cell

Internet – Worldwide networking of local networks by standardized data protocols

Internet service provider – Provides plain access to the Internet. Unlike online services, Internet service providers do not offer contents of their own

InterPro – Integrated protein motif database at the European Bioinformatics Institute that consists of several individual databases

Intranet – Computer network that is separated from the Internet by a firewall yet provides similar functions for the local users of the network

Intron – Noncoding part of a gene in eukaryotes. See also Exon

IP address – Internet Protocol Address. Industry standard for communication between open systems. The main task of the IP address is net-wide addressing. The protocol does not work with fixed routes to send data but with formatted blocks of data (packets). Datagrams find their way to the recipient via currently available connections. The IP address to identify individual computers is a unique 12-digit number of four three-digit blocks, each separated by a dot (e.g., 130.298.317.200)

ISDN – Integrated Services Digital Network. Digital telecommunication network to transmit language and data

Isoelectric focusing – Electrophoresis technique that separates proteins based on their individual pI values

JAVA – Object-oriented, hardware-independent programming language that was developed by Sun Microsystems, Inc. Java programs or applets can theoretically run on any computer that supports the Java run-time environment (JRE), independently of the respective computer architecture (PC, MAC, UNIX, etc.)

Knockdown – A method for elucidating the function of genes or proteins. For example, blocking transcription of a target gene by means of RNAi may result in phenotypic changes that can be analyzed. Because translation may not need to be 100% blocked to achieve the desired effect, the term knockdown applies rather than knockout, where translation is blocked completely

Knock-in – A method for elucidating the function of genes or proteins. To this end a transcribable gene is transfected into cells or organisms and the resulting phenotypic changes analyzed. Frequently a knock-in is used to reverse the change in phenotype caused by a knockout. If successful, then there is little doubt as to the function of the corresponding gene

Knockout – A method for elucidating the function of genes or proteins. With a knockout, the transcription of individual genes is entirely blocked. From the analysis of any resulting phenotype, conclusions can be drawn as to the function of the inhibited gene. Frequently, knockout experiments are combined with knock-in experiments

LAN – Local Area Network. Computer network connecting computers within a local area
Link – See Hyperlink
Local alignment – An alignment of sequences that does not take into account the entire sequence length
Locus – Position of a genetic marker or a gene on a chromosome
LocusLink – A database at NCBI that contains curated sequence data and descriptive information about genetic loci
Low complexity region – A region of DNA or protein that consists of one or few recurring bases or amino acids
Mail alias – Descriptive name of an e-mail account that can be easily remembered and used in an e-mail address instead of the real account name. See also Alias
MALDI-TOF – Matrix-assisted Laser Desorption/Ionization – Time of Flight. Mass spectroscopic technique that is frequently used to identify proteins
Mass spectroscopy – Spectroscopic technique that is used, for example, to determine the composition of peptides based on the masses of individual amino acids
Metabolite – Intermediate of a biochemical metabolic reaction
Metabolome – Entirety of all metabolites of an organism
Metabolomics – Scientific discipline that deals with the analysis of metabolites, i.e., the metabolic products of the cell
Microarray – See DNA microarray
Model organism – An organism that is used for the analysis of biological questions relevant also in more complex organisms (e.g. *D. melanogaster, C. elegans, M. musculus, D. rerio, A. thaliana*, *S. cerevisiae*, *E. coli*). However, the functional units being studied must be quite similar in the two organisms
Model system – See Model organism
Modem – Modulator/demodulator. Device to transmit digital signals through analog telecommunication systems
Motif – Conserved region within a group of related nucleotide or protein sequences
mRNA – Messenger RNA. RNA molecules that are synthesized during transcription and serve as templates for protein synthesis

Multiple alignment – Alignment of at least three sequences. See also Alignment

Mutation – Changes in the genome due to spontaneous events or triggered by mutagens such as UV light or chemicals. Leads to permanent loss or exchange of bases in the DNA sequence

Narrow spectrum antibiotic – Antibiotic with a mode of action limited to a species-specific target protein found within a small group of bacteria

NCBI – National Center for Biotechnology Information. The United States' contribution to the International Database Collaboration, which includes EMBL and CIB. NCBI is part of the U.S. National Library of Medicine, itself a part of the U.S. National Institutes of Health (NIH)

Needleman and Wunsch algorithm – Dynamic algorithm to compute a global alignment of two sequences

Nematodes – Roundworms or threadworms. Example: *Caenorhabditis elegans*

Neural network – A computational decision-making process to address complex problems that is analogous to the operation of the brain. A major characteristic of neural networks is their ability to adapt so that newly entered information can be recognized differentially

Newsgroups – Internet service to exchange information between many users. Newsgroups operate in a similar way to bulletin boards, i.e., news is published within a group and can be read by all users

NMR – Nuclear Magnetic Resonance. NMR is a spectroscopic technique to determine protein structures

Nonredundant database – A complete database composed of individual databases so that each database record is present only once, even if more than one component database contains the corresponding entry

Normalization – Correction of experimentally derived data to ensure accurate comparison between experiments. An example is the normalization of data that is necessary in expression profiling experiments

Northern blot – A method to detect mRNA. After electrophoretic separation in an agarose gel, the RNA is transferred onto a nylon or nitrocellulose membrane. On this membrane, individual mRNA transcripts can then be detected by hybridizing with labeled and complementary nucleic acids

Nucleic Acids Research – Molecular biological journal of the Oxford University Press. The first issue in January of each year is the database issue. All relevant biological databases are listed in this issue. In July 2003, a software issue was published for the first time that lists and describes freely available biological software

Nucleotide – Basic building block of DNA and RNA. Nucleotides consist of a base (C, A, T, G in DNA or C, A, U, G in RNA), a phosphate group, and a sugar residue (deoxyribose in DNA, ribose in RNA)

Oligonucleotide array – DNA microarray that consists of many thousands of single-stranded oligonucleotides. An oligonucleotide array is also called a GeneChip or BioChip

Oligonucleotides – Short DNA segments that consist only of a few nucleotides. These can act as starting points for PCR or they can be used in DNA microarrays as gene markers, for example

Online services – Providers of network services such as e-mail, chat, or bulletin boards. However, all these services run only on the computers of the providers, i.e., they are accessible only to their customers. The exchange of e-mails with customers of other providers is not possible. However, many online services also offer additional connections to the Internet

Open reading frame – ORF. A region within a DNA sequence that starts with a translation start codon (ATG) and ends with a translation stop codon (e.g. TAA)

Orthologous proteins – Homologous proteins that perform the same function in different organisms. Example: A serine protease in the digestive tract of humans and mice. The usage of ortholog and paralog in combination with the function of a protein is, however, controversially discussed. Although plenty of examples exist for which this evolutionary scenario

has played out, it is possible for orthologs to acquire different catalytic (or regulatory) properties and for paralogs to retain the same function (Jensen 2001). However, setting function aside, correct usage of ortholog and paralog requires knowledge of the details of the evolutionary pathways that produced the divergence of biological functions. Because insufficient information is available to accurately determine the timing of many of the speciation and gene duplication events, it can not be determined with certainty whether two contemporary proteins are orthologs or paralogs (Gerlt and Babbitt 2001). Therefore, we stick to the more manageable definition of orthology and parology as defined above because genome biologists need words to describe homologs encoded by different genomes and homolgs that have different functions. If necessary one could distinguish between isofunctional homologs that exhibit the same function(s) and heterofunctional homologs that exhibit different functions and isospecic homologs that are found in the same species and heterospecic homologs that are found in different species, as suggested by Gerlt and Babbitt (see also Chap. 4). See also Paralogous proteins

PAGE – Polyacrylamide gel electrophoresis. Analytical method to separate proteins based on their individual charges by applying an electric field across a polyacrylamide gel matrix

Palindrome – A DNA sequence that is inverse-complementary identical, i.e., that identical bases are present on complementary positions of the sense and antisense strand. For example, the complementary DNA sequence to GAATTC is CTTAAG, and the inverse-complementary to that again GAATTC. Such palindromes are frequently recognized by restriction enzymes

PAM – Percent Accepted Mutation. This matrix describes the amount of point mutations per 100 amino acids. The PAM matrix was developed by Margaret Dayhoff and is a substitution matrix used to generate sequence alignments

Paralogous proteins – Homologous proteins in the same organism that have similar, but nonidentical functions. Example: Two serine proteases in the mouse. See Orthologous proteins

Pathway – Metabolic route. Functional network between proteins

Patternhunter – A proprietary algorithm to perform homology searches in databases. Patternhunter is faster than gapped BLAST and also more sensitive. The program was successfully used to annotate the mouse genome, for example

PCR – See Polymerase chain reaction

PDB – Database containing 3D-structures of biological macromolecules, such as proteins

Pfam – A protein motif database based on Hidden Markov models

Phenotype – Appearance of a trait in an organism that is based on both a genetic disposition and on environmental influences. Examples of phenotypes are the eye color of humans or the association of certain diseases with families

Pharmacogenetics/genomics – Specific field that associates genetic disposition with the differing reactions individuals might have to drugs

Pharmaco-metabonomics – Method that analyzes those factors, e.g., genetics and environment, that influence the effects of drugs

Pharmacophore – Steric arrangement of functional groups in an active molecule that are responsible for the binding of the molecule to the receptor and thus its action. Also: The system of molecular properties responsible for the pharmacological action of the molecule

Pharmacophore modeling – Steric overlap of molecular structures of known drugs or inhibitor molecules and deduction of a pharmacophore from the analysis of congruent molecular properties

Pharmacophore screening – Search for molecules in a substance database that are similar to a calculated pharmacophore

PhenomicDB – Multiorganism genotype–phenotype database. PhenomicDB integrates data from a number of different genotype–phenotype databases thereby allowing cross-organism data comparisons

Phenome – The sum of all phenotypes of a cell, tissue, organ, organism, or species

Phenomics – Scientific discipline with the aim of understanding the function of proteins using phenotypes

Phosphorylation – An enzymatic process that involves the transfer of a phosphate group to proteins by a protein kinase

Phrap – Widely used sequence assembly program

Phylogenetic analysis – Analysis of the phylogenetic relationship between different organisms and their ancestors. Such analyses can include morphological, physiological, and genetic characters. See also Analogy, Homology, Relationship, Character, and Phylogeny

Phylogenetic tree – Graphical representation of phylogenetic relationships between organisms. Among others, phylogenetic trees can be derived from multiple sequence alignments of DNA or protein

Phylogeny – Phylogenetic evolution of living organisms and the origin of species over the course of the earth's history. See also Analogy, Homology, Relationship, and Character

pI value – pH value at which the positive and negative charges of a protein are neutralized and the net charge is zero. The pI value is also called the isoelectric point

Plasmid – Small ring-like DNA that can replicate independently of the chromosomal DNA of the cell. Plasmids are usually between 5,000 and 40,000 base pairs in length. They contain the information for building proteins, e.g., the antibiotic resistance genes. Bacteria can exchange plasmids. Because plasmids replicate quickly and are easily transferable between cells, they are used as vectors in genetic engineering to introduce and propagate genes in bacteria or yeast cells

Polymerase chain reaction – PCR. Reaction in which defined DNA fragments are exponentially amplified in vitro with the help of DNA polymerases. PCR was invented by Kary Mullis in 1983 who was awarded the Nobel Prize in chemistry in 1993

Polymorphism – A genetic variation in the DNA sequence of individuals within a population

Post-translational modification – Enzymatic modification of a protein upon the completion of translation. Examples are the phosphorylation and glycosylation of proteins

Primary database – A database that includes biological sequence data (DNA or protein) as well as accompanying annotation data

Primary structure – Linear sequence of the amino acids in a protein

Profiles – Position-specific assessment table to describe sequence information in a complete alignment. For each position in the sequence, profiles describe the appearance of certain amino acids, conserved positions, and deletions or insertions

Prokaryotes – Organisms that do not have a defined nucleus or other cellular compartments such as mitochondria. Bacteria belong to the prokaryotes

Promoter – A nucleotide sequence preceding a gene that determines where and when the gene is transcribed and to what extent. The enzyme RNA polymerase recognizes the promoter and binds to it thereby initiating gene transcription

Protease – An enzyme that processes or degrades other proteins or peptides. The term peptidase is also used

Protein array – Miniaturized technique with many thousands of proteins coupled to a solid support, thus allowing for their simultaneous functional analysis (e.g., for protein–protein interactions)

Protein profiling – Experimental technique that allows the understanding of a cell's profile based on the expressed proteins

Protein turnover – Time period between the synthesis of a protein and its degradation

Proteins – Proteins consist of one or several amino acid chains (polypeptides). Each amino acid is connected to the next by a peptide bond and a protein's sequence is determined by the nucleotide sequence of the corresponding gene. Proteins have various tasks in a cell (acting as enzymes, antibodies, hormones, etc.)

Protein families – Most proteins can be grouped into a protein family based on sequence similarities. Proteins or protein domains that are part of a protein family have similar functions and can be traced back to a common ancestral protein

Protein kinase – An enzyme that transfers phosphate groups onto proteins (phosphorylation). Phosphorylation frequently modulates the activity of target proteins

Protein lysate – A protein mixture that arises after the lysis of cells
Proteome – The entirety of all proteins of an organism
Proteomics – Scientific field that deals with the proteome of an organism by structural and functional analysis of proteins
ProtEST – Part of the NCBI database UniGene. ProtEST contains the EST sequences of a UniGene cluster that show a hit upon translation into a protein sequence
PSI-BLAST – Position-Specific Iterated BLAST. A program to find new members of a protein family within a protein database. PSI-BLAST also aids the identification of remotely related proteins
PubChem – A database at NCBI that contains information of small molecules and their biological activity
Point mutation – A single base change in a DNA molecule
Quality score – A measure that reflects the quality of each sequenced nucleotide of a DNA sequence as determined by DNA sequencers. Using the quality score, poor quality DNA regions can be removed from the final sequence
Quarternary structure – Association of several protein subunits to form a functional protein
Reading frame – Within a gene, groups of three nucleotides (codons) define an amino acid or a translation start or stop signal. Therefore, during protein translation, the reading frame corresponds to a sequence of consecutive "words" each with three "letters." If even a single nucleotide (letter) is inserted or lost within the gene, then the reading frame subsequent to the mutation will misalign resulting in the generation of a premature stop codon and a truncated, nonfunctional protein. On the other hand, the reading frame remains unchanged by the insertion or deletion of three nucleotides, resulting in either the gain or loss of one extra amino acid
Regular expression – Formalized description of a set of strings. Regular expressions allow the definition of a number of possible characters for every position in the string. The database Prosite uses regular expressions for the description of the characteristic signatures of protein families

Relationship – In a genealogical sense, an abbreviation represents a phylogenetic relationship. Unfortunately, the term is used very differently (e.g., also in terms of related forms = similarity). Two species or types or proteins (A and B) are regarded as more closely related compared to a third party C if they are descendants of a common ancestor not shared by C. Therefore, any ancestor that A and B share with C must be older than the common ancestor of A and B. Consequently, the degree of a phylogenetic relationship between species or proteins depends on how close common ancestors are to the present state. See also Analogy, Homology, Character, and Phylogeny

Reporter gene – A gene that encodes an easily detectable product. For instance, this can be an enzyme that converts a substrate resulting in a color (change) that can be measured

Restriction enzyme – Bacterial enzymes that cut DNA molecules at specific recognition sequences

Reverse transcriptase – An enzyme that catalyzes the transcription of RNA into DNA

RNA – Ribonucleic acid, a molecule chemically related to DNA that is central to protein synthesis. DNA is transcribed into mRNA that in turn is translated into proteins. Besides mRNA, a number of other RNA species exist (tRNA, rRNA, etc.)

RNAi – RNA interference. Naturally occurring mechanism in eukaryotic cells that blocks the expression of single genes. See also Knockdown

RT-PCR – A version of PCR that amplifies specific sequence regions in RNA. The RNA is first transcribed with the viral enzyme reverse transcriptase into cDNA and then specific sequences defined by primers are exponentially amplified by DNA polymerases

SAGE – Serial Analysis of Gene Expression. Experimental method to analyze gene expression in cells or tissues. SAGE, like DNA microarrays, is adaptable to the high-throughput production of expression data

SBML – Systems Biology Markup Language. A data interchange format for biochemical models based on XML

SCOP – Structural Classification of Proteins. A database that categorizes proteins with a known structure according to structural criteria

Score matrices – See Similarity matrices

SDS-PAGE – Sodium dodecyl sulfate polyacrylamide gel electrophoresis. See also PAGE

Secondary databases – Databases that contain information derived from primary databases. Fingerprint and motif databases such as Prosite, Blocks, and Pfam are secondary databases

Secondary structure – Ordered folding pattern of the polypeptide scaffold without consideration of the position of amino acid side chains. Example folding patterns are the α-helix, β-sheet, and loops

Sequence assembly – Generation of an alignment from overlapping short sequences of DNA followed by the assembly of a consensus sequence

Sequence retrieval system – SRS. Database management and query system to administer flat file databases. Amongst others, SRS is used on the EBI server to query biological databases

Sequence – Nucleotide or amino acid sequence

Sequencing – Determination of the nucleotide sequence in DNA or amino acid sequence in proteins. See also DNA sequencing

Server – A computer or a computer program that transfers information over a network, e.g., the Internet, to a client

Shell – Text-based input window to operate a computer, often referred to as command interpreter

SignalP – Computer program to estimate the N-terminal signal peptides of proteins

Signal peptide – Short N-terminal amino acid sequence (often between 15 and 30 amino acids) that serves as a signal for the cellular transport machinery

Similarity – Evaluation of the similarity of amino acid sequences. This implies the definition of similarity relationships between the 20 amino acids

Similarity matrices – Mathematical phrasing of similarity relationships between amino acids using a defined model

Significance – A significant result that does not occur by chance. The result is, therefore, assumed to be reliable with a high probability. Significance is calculated by a number of statistical tests

Singleton – EST sequences that do not show any overlap with other EST sequences and, therefore, cannot be grouped into contigs

siRNA – *S*mall *i*nterfering *RNA*. Small species of RNA (21–28 nucleotides in length) that are important in modulating transcription in eukaryotic cells

Six-frame translation – Translation of a DNA fragment into the six possible reading frames. This procedure is necessary when only uncharacterized DNA fragments are available and no details on the direction of the frame exist. See also Reading frame

SMD – Stanford Microarray Database. Database that allows the storage and retrieval of both raw data and normalized data from microarray experiments, and pictures of the corresponding arrays

Smith–Waterman algorithm – Dynamic algorithm to determine the optimal local alignment of two sequences. The Smith–Waterman algorithm can also be used to search databases. Though sensitive, the procedure is slow

SNP – Single Nucleotide Polymorphism. Genetic variation caused by a change in a single nucleotide

Spam – Unsolicited e-mails sent to a large number of recipients or undesired contributions to a large number of newsgroups. Spam is comparable to unwanted bulk mail

Splice variants – Proteins of different length originating from a process called alternative splicing

Spotting – Placing DNA spots onto a cDNA array with the help of a robot

SRS – See Sequence Retrieval System

Stackpack – A computer program developed to cluster EST sequences

Structural genomics – Worldwide initiative to automate the experimental analysis of the three-dimensional structures of as many proteins as possible

STS – Sequence Tagged Sites. Short unique DNA sequences that are used to tag genomes

Substitution matrices – See Similarity matrices

Swissprot – Curated high-quality protein sequence database of the Swiss Institute of Bioinformatics. See also Expasy

Synteny – Synteny refers to two or more genes lying on the same chromosome of a species

Syntenic regions – Chromosomal regions are syntenic if genes of orthologous proteins are in the corresponding chromosomal regions between two species, whereby the gene order is not considered

Systems biology – Scientific discipline with the aim of understanding biological organisms in their entirety. It involves the creation of an integrated picture of all regulatory processes from the genome to proteome and metabolome and on up to organelles, and the behavior of the entire organism

TAP – Tandem Affinity Purification. Method to identify multiprotein complexes.

Target – A protein that plays a central role in disease and whose activation or inhibition has a direct influence on the course of that disease

Target protein – See Target

Target-based approach – Modern search for drug targets that is carried out in vitro with a defined target protein

TCP/IP – Transmission Control Protocol/Internet Protocol. Protocol to communicate and transmit data on the Internet. An accepted industry standard to communicate between open systems. The transmission protocol defines the rules and conventions that control the flow of information in a communication system

Telnet – Teletype Network. The standard protocol for remote login on the Internet. Text-based method of communication between two computers that allows one to use a remote computer as if one were using its own terminal

Tertiary structure – The spatial organization (including conformation) of an entire protein molecule or other macromolecule consisting of a single chain

TIGR – The Institute for Genomic Research. A U.S. nonprofit genomics research institute. TIGR offers a number of databases as well as computer tools for sequence analysis

TMHMM – A computer program to determine the transmembrane domains of proteins using Hidden Markov models

Toxicogenomics – Scientific field that analyzes the effects of toxic substances on cellular gene expression

Transformation – The transfer of nucleic acids into living cells or bacteria (transfection). Also: The transformation of a normal cell into a tumor cell, for example by activation of oncogenes

Transcription – The act of producing an RNA copy of DNA using the enzyme RNA polymerase

Transcription factor – A protein that positively or negatively influences the transcription of genes, frequently by interacting with RNA polymerase

Transcriptome – The entirety of mRNA transcripts of an organism

Transcriptomics – Scientific discipline that performs global analyses of gene expression with the help of high-throughput techniques such as DNA microarrays

Translation – Synthesis of proteins at ribosomes using mRNA as the template

Transmembrane domain – Part of a protein that passes through a cell membrane

Twisted pair – Special type of cable that is frequently used to set up computer networks. The cable consists of several pairs of wires that are twisted around each other to guard against interference

Two-dimensional (2D) gel electrophoresis – Electrophoretic technique to separate protein mixtures. Proteins are initially separated in the first dimension according to their individual isoelectric points (pI value) and then in the second dimension according to their molecular weights

UniGene – Database at NCBI that contains all nucleotide sequences of a gene and describes them nonredundantly

Uniprot – Joint database of EBI, SIB, and Georgetown University that contains all the information of the Swissprot, TrEMBL,

and PIR databases and serves as a central repository of protein information

UniSTS – A nonredundant NCBI database containing STS markers from different sources

URL – Uniform Resource Locator. Address of an information source on the WWW. A URL consists of three parts, the protocol, the name of the server and the complete path, including file names (e.g., http://www.ncbi.nlm.nih.gov/genome/guide/zebrafish/index.html)

UTR – Untranslated Region. That part of RNA or cDNA that contains noncoding sequences. One distinguishes between 5′ UTR that is upstream of the translation start codon and contains important regulatory regions such as the ribosome binding site from the 3′ UTR that starts with the translation stop codon and often contains a terminal poly A-sequence

VDSL – Very high-speed digital subscriber line. A DSL technology incorporating glass fibers and permitting very high bandwidths

Vector – Usually plasmids (DNA rings) or phages (viruses that attack bacteria) to transfer genes between organisms. Vectors can be propagated in cells or bacteria as they include regulatory DNA fragments that are necessary for replication

Wildcard – A character used as placeholder that represents one or more arbitrary characters in the file name of a command

WWW – World Wide Web. Communication service on the Internet that mainly uses the HTTP protocol. See also CERN

X-ray crystallography – Technique to determine the three-dimensional structure of proteins based on protein crystals

Yeast two-hybrid system – In vivo method to identify protein–protein interactions in yeast cells

Index