This is the on-line version of our HGMP handbook, which is sent out to all our new users. A version, in PDF format, is available for you to browse locally, or to print off.

Brief Table Of Contents

This User Guide was written and produced by members of the UK MRC Human Genome Mapping Project Resource Centre.

First edition July, 1994; revised May, 1995; rewritten June 1996; revised July 1997; revised 1998; revised 1999.

UK HGMP-RC Contact Addresses

General Address:

UK HGMP Resource Centre,

Hinxton,

Cambridge,

CB10 1SB

UK

Tel: +44 1223 494500

Fax: +44 1223 494512

WWW: http://www.hgmp.mrc.ac.uk/

FTP: ftp.hgmp.mrc.ac.uk

Bioinformatics Help:

Tel: +44 1223 494520

email: support@hgmp.mrc.ac.uk

Biology Help:

Tel: +44 1223 494510

email: biohelp@hgmp.mrc.ac.uk

Administration:

Tel: +44 1223 494500

email: admin@hgmp.mrc.ac.uk

Training Courses:

Tel: +44 1223 494511

email: training@hgmp.mrc.ac.uk

Any Comments, Questions? Support@hgmp.mrc.ac.uk

1. Introduction to the HGMP-RC

I find that a great part of the information I have was acquired by looking up something and finding something else on the way.

Franklin P. Adams

1.1 General information

The UK Medical Research Council (MRC) established the Human Genome Mapping Project Resource Centre (HGMP-RC) in 1990. Its rôle is to provide biological materials, access to databases, information and the tools to analyse them, as well as training and help in the use of its facilities. It is primarily for scientists engaged in genome mapping and gene isolation studies.

The HGMP-RC's users now come from all over the world, and it has attracted significant funding from the EC. It is staffed by a team of software specialists and biologists dedicated to serving the needs of the research community in this rapidly evolving area. There are also other staff actively involved in research projects, whose close presence improves our service.

Other than the use of the GCG package, the services provided are free of charge to UK academics. Details of charges for other users are available from HGMP-RC Administration (see the contact sheet at the start of this guide), and they can also be found on our World Wide Web (WWW) pages starting at

http://www.hgmp.mrc.ac.uk/

1.1.1 Bioinformatics Services

The HGMP-RC aims to make access to information and services as simple and intuitive as possible. You can access the HGMP-RC via the internet from your computer, and then use the best tools and resources throughout the world. The HGMP-RC provides the latest versions of programs and data, often being aware of new developments before they have been publicly announced, and offers extensive online help and regular training courses.

Among the facilities provided are email, network news, protein and nucleic acid sequence analysis and manipulation, sequence database searching, genome data, linkage analysis, and databases of cell lines, clones and probes. These HGMP-RC facilities are made available via the WWW with convenient access to a vast array of international services. Details of how to access all these resources are given in Chapters 2 and 3 of this guide.

1.1.2 Biological Services

The HGMP-RC offers a variety of biological resources for research in human genetics and related areas. A centralised facility like the Resource Centre is a considerable asset to the scientific community, for the following reasons:

- It offers access to reliable and well-characterised materials which facilitate the comparison of data.
- It is in a good position to keep up-to-date with the latest developments and to import the latest available resources.
- Individual laboratories gain, as they are spared the cost of preparing and storing resources.

Details of the biological resources available, and an overview of how to make use of them, are given in Chapter 4 of this guide.

1.2 Registration

While much information is available to anyone who accesses our web site, the use of our biological services and full access to our computing services are restricted to registered users, who are issued with a unique username and password. Registered users also receive *Genome News* annually.

We welcome registrations from any individual or organisation with an interest in human genome mapping and sequencing. There is an application form at the back of this manual which you can photocopy, or you can register through our WWW server (see the contact sheet at the front of this manual). Registered users will be placed on the HGMP-RC newsletter mailing list and will be sent a username and password for their HGMP-RC computer account.

WARNING: Never share your password with anyone else, or allow them to use your computer account as a favour, even with constant supervision. They will be able to read your private mail, and delete all your data. No responsibility for any consequences of such actions will be accepted by the UK Human Genome Mapping Project Resource Centre.

If more than one person at any site requires access to the HGMP-RC, they MUST each obtain a separate account. Sharing of accounts is strictly forbidden.

The HGMP-RC reserves the right to deny access to its facilities to anyone found violating this rule or undertaking illegal activities which include copyright, libel, obscenity and "hacking" violations. Such restrictions are defined in the current registration form found at the end of this User Guide.

Once you are a registered user, please could you contact HGMP-RC Administration if you change your name or address. (See the contact address at the start of this manual).

1.3 How to Get Help

This manual is intended to be an introduction to the facilities offered by the HGMP-RC, and an explanation of how you can get to the information and tools you need to help you work more effectively.

1.3.1 Helpdesks

The HGMP-RC staff operate bioinformatics and biological helpdesks to answer any telephone or email queries that you may have. Details of these are given in the contact sheet at the front of this manual.

1.3.2 Help Documentation

Extensive help is available online, provided both by the developers of the tools we support and by the staff of the HGMP-RC. This is available along with each program and in the frequently asked questions on our WWW page. We also provide documentation that you can download to your own machine and print out locally.

1.3.3 Training Courses

Bioinformatics and biology training courses run by the HGMP-RC aim to show you how to use our resources effectively, and to illustrate particularly useful applications. Suggestions for new courses and locations are welcome. There are regular general introductory courses for bioinformatics, and specialised courses in particular subject areas and bioinformatics applications. For instance, if you wish to make best use of the phylogeny programs, it is essential to attend one of the phylogeny courses.

Details of course dates, locations, and content can be obtained from our WWW pages, or from the Training Course Administrator at the HGMP-RC. See the contact sheet at the front of this manual for details of how to reach us or look at the list of courses currently available at http://www.hgmp.mrc.ac.uk/About/Courses/

1.3.4 WWW Pages

To obtain detailed and up-to-date information on the services described above, use your WWW browser to look at the HGMP-RC WWW pages starting from http://www.hgmp.mrc.ac.uk/

---

2. General Computing Information

I think there's a world market for about five computers

Thomas Watson (Founder of IBM)

The HGMP-RC provides access to both bioinformatics and biological resources. We have bioinformatics programs to help you analyse your data and biological resources to help you produce it. Access to both is most readily achieved by connecting to our computing facilities.

Access to the bioinformatics applications (programs and databases) available at the HGMP-RC is through a series of menus, each containing options that may run programs, display data, or lead to sub-menus. Therefore this manual will not discuss in detail the arrangement of the menus, but will instead describe the general principles behind them and then look at the main types of programs and data to which they lead.

There are two methods that you can use to access the HGMP-RC applications. The first is access via the World Wide Web (WWW). The other, character-based route is via ***telnet***. Both methods of accessing the HGMP-RC, and the menus that they lead to, are discussed below.

2.1 Typographic Conventions

In this and subsequent sections, prompts and information sent to the screen by the HGMP-RC system are in **bold type**. What you type is in *italic type*. Comments in the middle of an example interaction between you and the computer are in normal type in brackets. For example:

	Prompt: *input*	(This is a comment)

A single key press is indicated by the name of the key in angled brackets; thus <*TAB*> indicates a press of the Tab key, and <*q*> means press the key marked **q**. A control character is indicated by the Control key name next to the character name, for example control-P is indicated by <*CTRL-P*>, which means, while holding down the key marked Control (or Ctrl), press **P**. < *RETURN*> means press the Return or Enter key. Program names are shown in bold italic type, for example ***telnet***.

2.2 Accessing the HGMP-RC Computing Facilities

In order to make use of the HGMP-RC computing facilities you will require access to the internet. Every University in the UK and most overseas are already connected to the internet, along with most research institutes. If you are an academic user in the UK you will probably have a connection to

JANET. JANET is part of the internet, and should provide high speed access to the HGMP-RC facilities.

If you are working from home you can use a dial-up service from a public provider or from your institution. We do not offer a dial-up facility at the HGMP-RC. If your institution does not offer this facility, there is a JANET national dial-up service for members of staff and students at academic institutions. The service currently costs [sterling]58 p.a. plus VAT and is operated by U-Net. If you are not eligible as an academic, then you can contact U-Net or one of the other service providers as an ordinary member of the public. See: http://www.u-net.net/ Alternatively you might explore some of the internet service providers offering free access for home users; many offers are currently available from supermarkets and high street shops.

To make full use of the facilities at the HGMP-RC you need a computer that has a Java enabled WWW browser. It is highly desirable that the machine is also able to display X-Windows. If you have neither WWW nor X-Windows then you will have to use our character-based menu accessed from a program called ***telnet***, or you could try connecting to us using ***VNC***. The discussion of X-Windows follows immediately, WWW access is described in Section 2.3, ***telnet*** access is described in Section 2.4 and ***VNC*** is described in Section 2.5

2.2.1 X-Windows and Point and Click

X-Windows programs run remotely (e.g. at the HGMP-RC) but display graphics output on your desktop computer. The graphics displayed often include a user-friendly point-and-click interface. X-Windows allows you to make the most effective use of the HGMP-RC. Some important programs and interfaces require or are best used with X-Windows.

X-Windows display software is available for PCs, Macintoshes, and OS/2, and it comes as standard on UNIX and VMS machines. PCs can use ***eXceed***,for which CHEST organise an academic deal in the UK. The CHEST web site can be found at http://www.chest.ac.uk/. Apple Macintoshes can run ***Mac-X*** or ***eXodus***.

Your desktop machine should be configured with a large colour screen with a high resolution (at least 1024x768 pixels is strongly recommended). A UNIX workstation (which may well be no more expensive than a PC) is likely to already be suitably configured, as is an X-terminal. The rest of the configuration will depend on the computing infrastructure to which the machine is connected.

You may find you need to configure your system to give our machines permission to generate X-Windows displays on your screen. To do this, issue the following command at the prompt on your local machine if it is a UNIX system:

```
unix%xhost menu.hgmp.mrc.ac.uk search.hgmp.mrc.ac.uk analysis.hgmp.mrc.ac.uk
link.hgmp.mrc.ac.uk services.hgmp.mrc.ac.uk
```

For details on configuring PC-compatibles and Macintoshes to display X-Windows, contact your local computer support personnel or the HGMP-RC Computing Helpdesk. You can find further information in the X-Windows FAQ (list of

Frequently Asked Questions) on our web pages at http://www.hgmp.mrc.ac.uk/MANUAL/faq

For the computer experts: we are able to handle ***Xauth***. Contact us for details.

2.3 Accessing HGMP-RC via the World Wide Web

In the same way that operating a computer can be easier when you can use a mouse to click on menus and buttons, rather than typing in obscure commands, so using the internet can be easier with a piece of software known as a web browser. The World Wide Web (WWW) is a collection of links between programs and data on computers all over the world, and you use a web browser to navigate the links.

There are a number of different browsers available, such as Netscape and Internet Explorer. Newer versions of these browsers tend to be larger than the earlier ones, and could run slowly on older computers, but they do offer additional features, such as Java & Javascript, which we are able to utilise to improve the service.

If you have a browser on your machine, you can use this now to connect to the HGMP-RC. We have constructed an elegant and powerful user interface that uses the WWW extensively, so that you can readily access your files and run programs. Simply point your browser at our WWW address , http://www.hgmp.mrc.ac.uk/ and you will see our welcome screen. Choosing any of the links on this screen will take you to further information on that subject.

2.3.1 The WWW Bioinformatics Applications Menu

Follow the ‘ **WWW menu**’ link under ‘ **Bioinformatics**’ from the HGMP-RC home page to access the available bioinformatics applications.

You can browse through the available programs, and search for anything you can’t find straight away. Each application has a description page, with details on what the application does and access to help on the application. To run it, select the **‘Run Now...**’ link. You will be prompted for your username and password, and then asked for how you want to access the application - using X, Java, or whatever. Where possible, these details are remembered so you shouldn’t be asked again in the same session.

Program options run in one of several ways:

- Some applications, such as ***NIX*** (see Section 3.8) and ***GLUE*** (see Section 3.24.1), are intended to provide an easy interface to an area of bioinformatics that users find difficult or tedious, or which is done frequently. Much effort has been put into making this integrated analysis service easy to use and well documented.
- They may display a WWW form for you to fill in, and will then either run and display their results, or they may inform you that they are running and you will receive the results later via electronic mail.
- They may be X-Windows programs, available only to those who have X-Windows set up.
- They may be programs that run in a UNIX session. These are usually programs that are interactive, making it difficult to specify all possible required parameters via a WWW form, or they may be a package of several interacting programs. You commonly have to type the names of the programs at the **unix%** prompt. You should refer to the help on each program for further details of running it.

The first three are displayed directly in your browser. To use the X-Windows programs you will need to have X installed on your machine, or if you have a reasonably up-to-date Java enabled browser then you can use ***VNC*** to display these programs. The ***VNC*** viewer in Java has been tested successfully on Windows 95 and NT and also works on Macs with Internet Explorer. See Section 2.5 for further

details.

If your computer has neither X-Windows nor a Java-enabled WWW browser, then any programs running in a UNIX session must log in to the HGMP-RC again. This will require your username and password before each program will start. We recommend that you ask your local computing department to set up X-Windows and a Java-enabled browser for you. In the meantime you should use the character based 'Telnet menu' (see Section 2.4.1), which only requires your username and pa ssword when you first start it.

2.4 Accessing the HGMP-RC via Telnet

If you do not have access to a WWW browser as described above, but you still have an internet connection, you can use ***telnet*** to access our menu of resources. The Telnet menu is a character based menu system. You may prefer to use it if you do not have X-Windows or if your internet connection is very slow.

The ***telnet*** program allows you to connect to other machines (such as the HGMP-RC's computers) from your own computer. To use it, you need to start the ***telnet*** program on your machine, and then connect to our site. On a UNIX machine, you would issue this command:

	unix% *telnet menu.hgmp.mrc.ac.uk*

On other machines, supply your program with the address above, and then instruct it to make the connection. If you have problems, either contact your local computer centre or the HGMP-RC Computing Helpdesk.

The following is an illustration of the user Fred Bloggs, who has a username of *fbloggs*, trying to log onto the HGMP-RC computing facilities via ***telnet***.

	unix%*telnet menu.hgmp.mrc.ac.uk*	
	Trying 193.62.192.50 ...	
	Connected to hgmp.mrc.ac.uk.	
	Escape character is '^]'.	
	Login: *fbloggs*	
	password:	(type your password, it is not echoed to the screen)

If the password entered is invalid the system will re-prompt for your username with **login:**, and you can try again up to three times before the network connection is closed.

You can now either simply press <*RETURN*> to accept the default terminal type that is suggested, or you can type in the terminal type that you are using. Some common types are: vt100, xterm, kermit. If you are unsure, then accept the default.

	Enter term = vt100

You will then be offered the opportunity to choose to use graphical or text-based versions of our programs:

	Do you wish to use X-Windows (y/n) > *y*

If you are not using X-Windows type <*n*> and then you will be logged in.

If you typed <*y*> for X-Windows, you will see:

	You may now enter your display name [or accept the default].
	Enter display name [somehost.redbrick.ac.uk:0.0] >

If you are unsure, accept the default display. The HGMP-RC menu system will then welcome you with the message of the day. If you have problems, please contact the computing helpdesk.

2.4.1 The Telnet Menu

Menu Options

		MOLECULAR BIOLOGY SOFTWARE FOR THE HGMP-RC MAIN MENU
	0)	Help
	1)	Exit
	2)	Electronic Mail
	3)	BIOSCI/Network News (Biologist's Bulletin Boards)
	4)	Information Services
	5)	Analysis and Manipulation of Sequences
	6)	Sequence Database Searching
	7)	Genome Data
	8)	Linkage Analysis
	9)	Cell Lines, Clones and Probes Databases
	10)	Other Molecular Data
	11)	Utilities (File Transfer & Management)
	12)	UNIX Operating System
	13)	Miscellaneous ('How to ...' etc)
	14)	Queries, Suggestions and Comments to User Support
	Enter a number, option-name or ? >	

Figure 2.1 HGMP-RC Telnet Main Menu for Registered Users

This table represents the main menu screen that you will see when you connect to the HGMP-RC by ***telnet.*** You can select an option by typing in the relevant number and then pressing <*RETURN*>. Option < *0*>will always give you some help about what sort of programs are available at the current menu. Option < *1*>will always take you back up a level to the previous menu; if you are at the top level (main) menu shown above, option < *1*> will check that you really did want to exit from the menu and will then log you off the HGMP-RC menu system and close your network connection to the HGMP-RC.

Menu Defaults

If you press <*RETURN*> at the menu prompt without having entered anything else (the default), the system will assume that you wanted option < *1*> (previous menu/exit). This is the general behaviour of the menu system and of many of the options presented to you.

However, this behaviour cannot be guaranteed once you have started a program that has been written by a third party. In many cases we have created a friendlier front end to prompt you for sensible information and this blurs the point at which the menu system finishes and the third party program starts. The prompts will generally respond in the way described here to defaults and to entering < *?*>, but in general, once an option has started a program, you should assume the worst. It is a very good idea to read the detailed help about an option in order to discover how to exit from the program before you run it.

Option Names

Options can be run by typing the name of the option as well as by typing the option number. Many of the option names are obvious: type ‘ *emboss*’ to run EMBOSS, ‘ *pine*’ to run the mail program Pine, ‘ *unix*’ to access the UNIX operating system, or ‘ *support*’ to send a request for help to the HGMP-RC Computing Helpdesk. All option names are in lower-case, and a full list of all of the option names can be seen by typing < *?*> at any menu.

An option may be run by typing its option name, even if it is not an option on the current menu. This allows you to run an option quickly without having to negotiate your way down the menu system, if you know which option you wish to run. Typing *‘help’* is the same as choosing option < *0*>. Typing *‘exit’* is the same as choosing option < *1*>.

UNIX Commands from the Menu

The menu prompt understands many UNIX commands for manipulating files. It also understands common aliases of these commands, so that you may either enter the UNIX command or its alias and the effect will be the same. For an introduction to the most commonly used UNIX commands, see Section 2.10.

2.4.2 Running an Option

Once you have selected an option, either by number or name, one of two things will happen. If an option leads to a submenu, you will then be able to make a selection from that. Otherwise, the program that you have selected will start to run. Most of the program options will display a page of brief help such as:

Do you want help on 'emboss' (y/n or quit) >

If you type <*q*> or ' *quit*', you will go straight back to the menu you have just come from. If you type <*y*>or ' *yes*', you will get further help on this program, and will then go back to the menu. If you type <*n*> or ' *no*', you will run the program. If you are confused by this question at any time, type < *?*> and you will get a help message.

Pressing <*RETURN*> here is the same as < *n*> - you will run the program. This is the only place in the menu where the default action will do something rather than getting out from wherever you are (unless you are running a program written by a third party). The reason for this is that most people will have chosen the menu option with the purpose of running the program it relates to, and want the default action to be to run it.

If you intend to use a program in your work, you should definitely read the full help presented for the option, plus any references given.

2.5 Accessing the HGMP-RC using VNC

VNC (Virtual Network Computing) is a system that allows you to display an X session running at the HGMP-RC on your desktop computer, even if that computer doesn't understand X. There are two ways of viewing the session; you can run it as a Java applet in a Java compatible browser such as Netscape or Internet Explorer, or you can use the standalone viewer available for some platforms.

You should note that our ***VNC*** service is still at an experimental stage. Nevertheless, it has such great potential that we feel it's worthwhile offering it. We would appreciate both positive and negative feedback. You can find information about system requirements and common problems at:

http://www.hgmp.mrc.ac.uk/Registered/Webapp/vnc/

The official ***VNC*** webpages can be found at http://www.uk.research.att.com/vnc

2.6 Accessing the HGMP-RC via ssh

It is possible to access the HGMP-RC computing facilities using the ***ssh*** (Secure Shell) protocol. ***ssh*** has two main advantages over ***telnet***:

- The connection is encrypted and secure.
- X-windows traffic can be securely carried over the ***ssh*** connection.

It is worth trying ***ssh*** if it is available to you. In addition, the ability of ***ssh*** to tunnel X-traffic across the network can be very useful at sites where the use of X-windows is restricted by a firewall.

If you are using a UNIX system with ***ssh*** installed then the command to connect to the HGMP-RC would be:

unix%*ssh -l fbloggs menu.hgmp.mrc.ac.uk*

where*fbloggs* is your HGMP-RC username.

The Telnet menu should start automatically once you have logged in.

2.7 General account information

2.7.1 Disk Quotas

When your account at the HGMP-RC is set up, you will be allocated a certain amount of disk space. If you see a message saying that your disk quota has been exceeded, this means that you have run out of room to store any more data, and you need to clear away some of your files. You can check your current disk usage at any time by typing ***pquota*** at a UNIX prompt. We can, within reason, increase your disk quota on request. Contact the HGMP-RC Computing Helpdesk for details.

2.7.2 Temporary File-space and Large Projects

If you require large amounts of file-space (> 50 Mb) to store files temporarily, you can use the directory */data/scratch*. You should create a subdirectory there (e.g. */data/scratch/fbloggs*) and do your work in that subdirectory. Files under */data/scratch* that are more than one month old are deleted automatically.

If you have a project that requires you to store (or even share with other users) large amounts of data for more than a couple of weeks, you should contact HGMP-RC User Support; we may be able to arrange a data directory that is not subject to the normal quota constraints. File space at the HGMP-RC is not unlimited and we will require at least some justification for doing this.

2.7.3 Passwords

Changing Your Password

To change your password when using the WWW Bioinformatics menu follow the ‘ **Password**’ link under the ‘ **Common Options**’ heading. If you are using the ***telnet*** character based menu then select the ‘ **Utilities**’ menu , followed by the ‘ **File Management**’ option and then select ‘ **Change your password**’. You will see the following:

	Old password:	(type your old password)
	New password:	(type your new password)
	Re-enter new password:	(type your new password again to check for errors)

If this has been accepted, the following will be displayed:

	NIS+ password information changed for user.
	NIS+ credential information changed for user.

Choosing a Password

The object when choosing a password is to make it as difficult as possible for a password cracker to make any guesses about what you've chosen. This leaves him no alternative but a brute-force search, trying every possible combination of letters, numbers, and punctuation. A search of this sort, even conducted on a machine that could try one million passwords per second, would take an average of over one hundred years.

If we guess your password using any of the available guessing programs your account will be disabled and you will have to contact Bioinformatics user support to have it reactivated.

It is therefore in your best interests to select a secure password. With this as our goal, a set of rules for password selection can be constructed:

- Don't use your username in any form (as-is, reversed, capitalised, doubled, etc.).
- Don't use your first or last name in any form.
- Don't use your partner's, child's, or pet's name.
- Don't use other information easily obtained about you. This includes your date of birth, car numbers, the make of your car, telephone numbers, national insurance numbers, the name of the street you live on, etc.
- Don't use a password of all digits, or all the same letter. This significantly decreases the search time for a cracker.
- Don't use a word contained in (English or foreign language) dictionaries, spelling lists, or other lists of words.
- Don't use a password shorter than eight characters.
- Don't write it down!
- Don't tell it to anyone!

So:

- Do use a password with mixed-case alphabetics.
- Do use a password with nonalphabetic characters, e.g., digits or punctuation.
- Do use a password that is easy to remember, so you don't have to write it down.
- Do use a password that you can type quickly, without having to look at the keyboard. This makes it harder for someone to steal your password by watching over your shoulder.

2.8 Email

Electronic mail (email) allows you to send messages to colleagues on networks all over the world. They cost nothing, they are fast (1 - 15 minutes), and they avoid problems like shouting down telephone lines to give your colleague in the States the fine details of your latest technique - you simply send a precise description that you type into the mail system. It is well worth asking people you meet for their email address as well as their telephone number and physical mail address.

The email address for your account at the HGMP-RC is formed by your username followed by the letters ' *@hgmp.mrc.ac.uk*'. Thus, Fred Joseph Bloggs' email address would be:
fbloggs@hgmp.mrc.ac.uk

(With some frequently occurring names we have had to use all the initials and possibly append a number: *fjbloggs2@hgmp.mrc.ac.uk*)

The recommended program to read and send mail is ***pine***, which has comprehensive online instructions. Alternatively, you may prefer ***dtmail.***It is best to find the one you prefer and then stick with it. Do not have two mail readers running simultaneously or you could start losing messages.

If you wish all of the mail sent to your HGMP-RC account to be automatically forwarded to your account at your local site, then use the forwarding option in the ‘ *Electronic Mail*’option of the telnet menu .Type in your local email address:

Enter forwarding email address: *fjb765@somehost.redbrick.ac.uk*

If you wish to change this in the future, choose the option again and type in your new address. You can also use http://www.hgmp.mrc.ac.uk/Registered/Webapp/forward/

to set up forwarding. In either case, you should get a confirmation test message. Check that you receive this, otherwise email messages sent to your account at the HGMP-RC may be lost.

2.9 Network News

Network News is a set of thousands of newsgroups, which are like electronic bulletin boards on which people can post messages about specific subjects for the world to read, allowing quick and efficient exchange of ideas. They are particularly useful for obtaining possible solutions to problems. By default you are subscribed to the newsgroups devoted to genome project related subjects.

You may be able to run a news reading program on your own computer, accessing the messages held at the HGMP-RC. There are several programs that can do this for you, including some versions of the more popular web browsers.

In order to use this facility, you will have to tell your news reader to talk to our news server (your program may refer to it as the ‘nntp’ server). It is called *newshost.hgmp.mrc.ac.uk.*If your program is able to connect to our news server, it will ask you for your HGMP-RC username and password. When these have been verified you will be able to continue and select the news group you wish to start reading.

However, you should be aware that there are some programs that are not designed for allowing an individual to authenticate themselves for access to a news server. If you do not have a news reader that understands this method of authentication, then you will have to run a news reader, such as ***trn*** or ***knews***, on the HGMP-RC computing facility. ***trn*** is a character based program while ***knews*** is point-and-click, requiring that you can display X-Windows on your computer.

Articles are expired after a period of time; there are often archives for newsgroups, allowing searching for topics or old articles.

2.10 UNIX

2.10.1 Introduction to UNIX at the HGMP-RC

This section provides introductory information on the UNIX operating system used at the HGMP-RC. The purpose of this section is to provide you with the basic commands required to use those applications programs which run in a UNIX environment. If you wish to learn more about UNIX then you should see the online help (http://www.hgmp.mrc.ac.uk/Documentation/Unixhelp), attend one of

the HGMP-RC training courses, or read an appropriate book. Please contact the HGMP-RC Computing Helpdesk if you are having difficulty using UNIX or any of the applications at the HGMP-RC. of

2.10.2 Getting out of UNIX

If you ever get into a program by mistake that you don't know how to get out, try (in increasing order of desperation):

	<RETURN>, exit, quit, x ,<CTRL-C>, <CTRL-D>, Q, q, ?, help, <CTRL-z>, <ESC>z, <ESC>:q!

If, while using the Telnet menu, you find yourself facing a **unix%** prompt and you don't know what to do, don't panic. You can get back to the menu by typing *'exit'*. You may sometimes see the statement **'There are stopped jobs'** displayed; don't worry about this, just type ' *exit*'again

2.10.3 Basic UNIX Commands - Files

Listing Files

The command that lists the files in a directory is called *ls*(Think **lis**t). There are various ways to use *ls*:

	unix% *ls*	Lists the files in your directory
	unix%*ls -l*	Lists the files with the time they were last edited, their size, and a few other useful features
	unix%*ls mydata**	Lists all the files whose names begin with "mydata"

You can look at the contents of a file using the command *more*:

	unix%*more data.txt*	Displays the contents of the file called *data.txt* one page at a time

Once you are running *more*, some commands you might like to use are:

	<SPACE>	Shows the next page
	<RETURN>	Shows the next line
	<b>	Goes back one page
	/text	Searches for ' *text'*
	<q>	Quits

Do not press the *<v>* key while using *more* - it starts an editor called ***vi*** (and you wouldn't like it!). Note: to get out of *vi* type < *ESC><:><q><!><RETURN>*.

Copying and Renaming Files

You can copy a file using *cp*(think **c**o**p**y). To rename files, move the original file to a file with the new name using the command *mv*(think **m**o**v**e):

	unix%*cp file1 file2*	Makes an exact copy of *file1* called *file2*
	unix% *mv file1 file2*	Renames *'file1'* as *'file2'*

N.B. When copying or renaming files, don't create files with spaces or the following punctuation characters in their name: * & ? < > $ ( ) ~ | \ /

Deleting Files

You can delete files using the command *rm* (think **r**e**m**ove):

	unix%*rm filename*	Removes a file called *'filename'*

The file is then removed from the system. (Yes - we may be able to get it back for you from the daily or weekly file backups if it is over a day old. But please be careful!).

2.10.4 Basic UNIX Commands - Directories

Directories are groups of files. They are organised in a hierarchy, so directories can have subdirectories and so on. Directories can be created, deleted, and renamed. Files can be moved or copied into specific directories. You can only work in one directory at a time - your "current working directory".

When you login, you are in your "home" directory. You can return to this at any time using the command *cd* with no further arguments (see below). To find out which directory you are currently in, use the command *pwd* (think **p**rint **w**orking **d**irectory).If you *cd* to your home directory and then type *pwd*, you will see something like this:

	unix%*cd*	
	unix%*pwd*	
	/people/fbloggs	(where **fbloggs** is replaced by your username)

fbloggs is the name of the home directory of Fred Bloggs, and it is a subdirectory of the *people*directory.

Changing Directories

You can move around between directories using the command *cd* (think **c**hange **d**irectory):

	unix%*cd*	Change to your home directory (regardless of where you are currently)
	unix%*cd ..*	Change to the directory above your current working directory:
	unix%*cd subdir*	Change to the subdirectory called *subdir*

Creating and Deleting Directories

You use the command *mkdir* to make a new directory (think **m**a**k**e **dir**ectory). You can remove empty directories with the command *rmdir*(think **r**e**m**ove **dir**ectory

	unix%*mkdir subdir*	Makes a new subdirectory called ‘ *subdir*’
	unix% *rmdir subdir*	Deletes an empty subdirectory called ‘ *subdir’*

Moving a File to a Directory

You can use *cp* and *mv* to **c**o**p**y and **m**o**v**e files between directories:

	unix%*cp file subdir*	Makes a copy of *file* in *subdir*
	unix% *mv file subdir*	Moves *file* from the current working directory into *subdir*

2.10.5 UNIX Commands - Quick Reference

You can find out more about any UNIX command by typing ‘ *man*’ and the name of the command. For example:

	unix%*man ls*

To help you out, here is a table of the commands you will most frequently use; remember, these are the commands you type at the

unix % prompt:

	pwd	print working directory
	ls	lists the files in the directory
	ls -l	lists file details eg date,size,owner,permissions
	cp file1 file2	copy ‘file1’ to the new ‘file2’
	cp file1 subdir	copy ‘file1’ into the directory ‘subdir’
	mv file1 file2	rename ‘file1’ to be ‘file2’
	mv file1 subdir	move ‘file1’ to the directory ‘subdir’
	rm filename	deletes the file permanently
	*rm -i file**	
	more filename	types out the file ‘filename’ one page at a time
	cat file1 >> file2	appends ‘file1’ to the end of ‘file2’
	cd gammaseqs	change to the directory ‘gammaseqs’
	cd	go back to your login directory
	mkdir gammaseqs	makes the ‘gammaseqs’ directory
	rmdir gammaseqs	removes the empty directory ‘gammaseqs’
	man command	online help for a UNIX command.

2.10.6 Creating Files with Pico

pico is the editor used in the ***pine*** mailer. It has extensive on-line help, and a list of available commands is constantly displayed at the bottom of the screen. The syntax ^X means < *CTRL-X*>, that is, press the key marked *X* while holding down the < *CTRL*> key. You should use this editor whenever possible. To start ***pico*** editing a file called *myfile.dat*, type the following:

	unix%*pico myfile.dat*

You may encounter other editors, including ***emacs***; if you find yourself in it and want to quit, type <*CTRL-X*><*CTRL-C*>, or to save the file and exit, type <*ESC*> <*Z*>

If you have X-Windows, you might like to try using the editors ***nedit***(straightforward to use) or ***xemacs***(more complicated but very powerful).

2.11 Transferring, Processing and Printing Files

Perhaps you have data from your latest sequencing run that needs to be searched against one of the nucleic acid databases - but the data are on your desktop PC and the databases are at the HGMP. Alternatively, maybe you have performed a complex analysis on some data in your account at the HGMP and now need to incorporate the results in your latest paper - but the Word document is on your PC at home. For many reasons, you will often need to transfer files between the HGMP and your

local computer, especially if you want to print them. We cannot print files out at the HGMP-RC for you. We will now discuss some of the options available to you for achieving this.

2.11.1 File Transfer Using the HGMP-RC Filemanager

This is the easiest way of transferring individual files from your HGMP-RC account to your local computer, printing such files on your local printer, or uploading files from your computer to the HGMP. You should follow the ‘ **Files**’ link from the WWW Menu at http://www.hgmp.mrc.ac.uk/Registered/Menu/ and run ***filemanager*** from there. There are several options that allow you to view your files and directories, and you can transfer or print any files you see using your WWW browser by using the browser’s built-in **‘Save As ...’** and **‘Print’** options respectively. You can also upload files from your desktop machine into your account at the HGMP by following the link marked ‘ **Upload to here’** from the ***filemanager*** page.

2.11.2 Submitting data to HGMP applications

Most of the applications that ask you for input give you three options for achieving this:

- Give the location of a file in your HGMP account - you can use the filemanager to find the path to the file you want
- Specify a file on your desktop machine - the "Browse" button pops up a menu to allow you to select the file you need
- Paste your data into a textbox; this is a very easy way to submit data to an application.

Either of the second two options will automatically do the data transfer for you, though you should note that if you want to use the data again later at the HGMP, you might be better advised to do a "proper" file transfer and save the data in your account.

2.11.3 File transfer using HGMP forms

We have written some useful forms based interfaces to allow you to transfer data from your local machine into a file in your HGMP account that you can later use in our applications. They can be found at:

http://www.hgmp.mrc.ac.uk/Registered/Menu/filetran.html

"Easy file transfer by WWW upload" allows you to browse for a file on your local machine to be copied into your home directory at the HGMP. You should specify a name for the HGMP file where prompted.

"Cut & Paste data to transfer it to the HGMP" provides a text box for you to paste information such as sequence data. The data will be copied into your home directory into a file given the name that you specify on the form.

2.11.4 File Transfer by Email

You can also transfer files by attaching them to mail messages For example, to send an attachment using ***pine***, start ***pine***and choose option *’c’*(compose a message). Fill in the destination and subject fields, and type any accompanying message, and put the name of the file to be transferred in the *Attchmnt* field:

	To	**:**fbloggs@somehost.redbrick.ac.uk
	Cc	**:**
	Attchmnt	**:***file.dat*
	Subject	**:***bacterial data*

If the recipient's email system understands about 'Attachment' files, they will receive email with the file attached, and they can save this file and then read the data.

If you receive an attached file via email at your HGMP-RC email address, you save the file in ***pine*** by giving the ' *v*' command to view the attached file and then you can save it.

2.11.5 File Transfer Using FTP (File Transfer Protocol)

This is a method available on most Macs, PCs and UNIX machines that allows you to transfer files between two computers. If a file is plain text, transfer it in ASCII mode, otherwise transfer it in binary mode (if you know what you're doing). This is necessary as different computers store files in different ways.

The HGMP machine to which you should connect to access your home directory via ***ftp*** is ***files.hgmp.mrc.ac.uk***

Most desktop computers now come with ***ftp*** software installed. There are a very large number of graphical PC interfaces available to you; your internet service provider will probably recommend one or more of these applications and will be able to advise you on installing them. We recommend that if you are not already familiar with command line ***ftp*** you try using one of the graphical interfaces as they make file transfer extremely straightforward. An alternative would be to use the ***ftp*** facility built into a browser such as Netscape.

When you open an ***ftp*** connection to the HGMP you will be prompted for your HGMP username and password; you will be connected to your home directory and can transfer files to and from your own account. Using ***ftp*** you are also able to copy data or programs available on machines world-wide. Many sites have a guest account allowing you access to data. In these cases, login in with the username *anonymous* and then enter your email address where it asks for a password. This is also the method used to get copies of the registration forms from the HGMP-RC if you do not access us via a Web browser.

The details of how to transfer files by ***ftp*** will differ depending upon the type of machine you are using, but the basic principles remain the same. If in doubt, you should consult your local computing support.

If you cannot find an ftp tool on your machine, there are a few provided for you at http://www.hgmp.mrc.ac.uk/Registered/Menu/filetran.html. You can start up a command line ftp session or use one of the graphical interfaces listed. PC users running Vista eXceed may find Exchanger the easiest option.

2.11.6 File Processing and Printing

Once you have transferred a file back to your local site, you may need to uncompress it before you can print it. Many of the manuals are compressed so they take up less room and transfer across the networks faster. You can tell whether a file needs uncompressing by its suffix (the letters on the end of its name). The following table illustrates the meanings of some of the more common suffixes.

Suffix	**File Type**	**Compressed?**	**Process with:**
.txt/.asc	Plain Text	No	Any text editor
.ps/.epsf	Postscript	No	Any Postscript viewer or printer
.tex	TeX/LaTeX	No	LaTeX
.tar	UNIX tape archive	No	tar
.Z/.z	UNIX compressed	Yes	uncompress/gunzip
.gz	UNIX gzip compressed	Yes	gunzip
.zip	PC archive	Yes	winzip, among others
.hqx	Macintosh archive	Yes	unstuffit
.sea	Macintosh archive	Yes	Self-extracting

If you have problems transferring, processing, or printing files, ask your local computing support or if this fails, contact the HGMP-RC User Support Helpdesk.

Any Comments, Questions? Support@hgmp.mrc.ac.uk

3. Bioinformatics Services

Don't expect your computer to tell you the truth.

Gunnar Von Heijne from "Sequence Analysis in Molecular Biology"

We offer a variety of bioinformatics training courses to help you make the most of our services. You can see a list of the training courses currently available at

http://www.hgmp.mrc.ac.uk/About/Courses/

You can also look at the course notes for our Introductory Biocomputing Course via http://www.hgmp.mrc.ac.uk/About/Courses/current/comp.intro.course.html. Access to the bioinformatics programs and databases available at the HGMP-RC is through a series of menus, each containing options that may run programs, display data, or lead to sub-menus. Use of the menu system via the WWW or ***telnet***was described in Chapter 2.

3.1 Applications

This chapter is a brief introduction to the various packages we have available to registered users, broken down into the types of task users most frequently want to perform. For detailed step-by-step instructions on how to use some of these programs, please see Chapter 6. Space constraints do not allow us to give detailed information here and we strongly recommend that you read the detailed help available for most of these applications. This can be found on the web page of each application, or from links thereon.

3.1.1 The Bioinformatics Applications Support Rating

The options available from the HGMP-RC menu system are rated by a star system in order to give you some indication of the support and help we can offer if you use the option. Three stars are given to fully-supported applications, while one star means there is no specific expertise on this application at the HGMP-RC.

3.1.2 Getting Started

The table overleaf gives an idea of which programs to use in order to carry out specific tasks.

Task	**Program**	**Comment**
Gene Identification	NIX	Integrates and displays many gene identification programs.
Searching sequence database with sequences	BLAST	Fairly fast, excellent for a first pass.
	FASTA	Slower than BLAST; sometimes more sensitive.
Search by keyword	SRS	Very quick, also provides links between databases.
	Entrez	Sequences and MEDLINE.
Sequence analysis packages	EMBOSS	Many useful programs.
	GCG/EGCG	Comprehensive and well documented.
	Staden	Many useful programs.
Genome databases	GDB	Human genomic information.
	OMIM	Catalogue of human genetic disorders.
	MGI	Mouse genomic information.
Bibliographic databases	ISI	Science citation index.
	EMBASE	Excerpta Medica Database; abstracts etc.
	Entrez	Integrated searching of MEDLINE,DNA and protein sequences.
Phylogeny	Phylip	Comprehensive, well documented.
	PIE	WWW front end to phylogeny.
Protein analysis	PIX	Integrates and displays results from many protein analysis programs.
Cosmid assembly	Staden	Used by many genome centres.
Multiple sequence alignment	Clustal	Popular alignment program.
	MAGI	WWW front end to multiple sequence alignment
Linkage analysis	GLUE	WWW front end to linkage analysis.
	Fastlink	Improved version of LINKAGE package.
	VITESSE	Extremely fast linkage programs.

3.2 Getting Help

There is a huge amount of help available. Every menu option, whether from the WWW menu or ***telnet*** menu has some accompanying help. This is usually based on what the authors provide, so will vary in quality and quantity.

We also have many program manuals and a frequently asked questions (FAQ) section that can be found at http://www.hgmp.mrc.ac.uk/MANUAL/faq/ or by following the ‘ **FAQ**’ link from the HGMP-RC home page.

If you cannot find out how to do something, or if you have complaints or suggestions, please contact the helpdesk. The more feedback you give us the better the service we can provide. You can contact us by:

- Clicking on the support@hgmp.mrc.ac.uk option at the bottom of every web page.
- Emailing support@hgmp.mrc.ac.uk
- Visiting http://www.hgmp.mrc.ac.uk/MailSupport.html
- Choosing the support option from the telnet menu.
- Telephoning the support desk: *01223 494520*

3.3 Databases

There are many biological databases available on the WWW in general and from the HGMP-RC in particular. Here we look at some of the most useful and important. Some of these are held at the HGMP-RC, but most are links to external sites.

There is often a substantial overlap between the types of data held in different database systems, and thus no hard and fast classification of databases can be made. Some of these databases are extremely complex and space does not permit us to illustrate all their capabilities; however, we hope to give you a useful introduction to them.

Most of the web based databases have links to related data in other databases that can save you time. You should be aware that sometimes the links between databases are incorrect and you will arrive at an inappropriate entry.

As a general starting point:

If you want to:	**you should use**:
- find a gene location	GDB, Genecards
- find a sequence or bibliographic reference	Entrez
- retrieve sequences	SRS
- rapidly access a large amount of information about specific human genes and their products	Genecards

3.3.1 Sequence databases

There are a number of sequence databases in widespread general use in the biology community:

Nucleic Acid Sequence Databases	**EMBL** (compiled at the EBI) **EPD** (Eukaryotic Promoter Database)
Peptide Sequence Databases	**SwissProt** (Compiled at the EBI & Switzerland) **TREMBL** (A translation of EMBL sequences) **PIR** (Protein Identification Resource) **OWL** (Non-redundant collection from many protein databases) **NRL_3D** (the sequences of PDB entries)
Other data	**REBASE** (Restriction enzymes) **PROSITE** (Protein motifs) **TRANSFAC** & **TFD** (Transcription factors)

EMBL is the most widely used nucleic acid database, while SwissProt and TREMBL are the most popular protein ones. Many of these databases are accessible using the EMBOSS package described in Section 3.5.1

Accession Numbers

Each sequence has a unique accession number permanently associated with it. Each sequence also has an ID name, but this is not guaranteed to be unique between databases. If the sequence is merged with another sequence another accession number may be assigned. The old accession number becomes the secondary accession number.

3.3.2 Genomic databases

GDB

http://www.hgmp.mrc.ac.uk/gdb/

GDB was devised as the ultimate repository of human mapping and genomic data. There are several ways to make powerful searches of GDB, though it is also very easy to search for information on genes, clones and so on. GDB does not hold any information on gene products.

There are links from GDB to Entrez, sequence databases, MGI, other genome databases, OMIM, GeneCards, HGMD etc.

ACeDB

AceDB started life as a repository for mapping and genomic data for the nematode *C. Elegans*. Data on several human chromosomes and on the genomes of several other organisms are now held in ACeDB-style databases. The human data can be found in links from

http://www.hgmp.mrc.ac.uk/Registered/Menu/human-gen-db.html

3.3.3 Clinical and Mutation

OMIM

http://www.hgmp.mrc.ac.uk/omim/

A very simple and very useful database of phenotypes of human diseases having a substantial genetic component. It has links to HGMD, GDB and other databases.

HGMD

http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html

A database of sequences and phenotypes of human disease-causing mutations. It has links to OMIM and GDB, and to locus-specific databases.

3.3.4 Integrated

GeneCards

http://bioinfo.weizmann.ac.il/cards/

GeneCards is a database that aims to address some of the problems of information overload and time-consuming data-mining by integrating biomedical information taken from several sources (GDB, MGI, OMIM, SwissProt, HGMD, Doctor's Guide to the internet etc.) and by presenting them concisely.

GeneCards is the best place to start searching for human genomic information, and has links to many other databases.

Entrez

http://www.ncbi.nlm.nih.gov/Entrez/

Entrez is a set of tightly linked databases including nucleic acid sequences, protein sequences and MEDLINE. It has a user friendly interface and is a very powerful system. It is self referential - for example, when you find an interesting nucleic acid sequence entry you can quickly find others like it, the corresponding protein entry and abstracts of papers describing it.

SRS - Sequence Retrieval System

http://srs.hgmp.mrc.ac.uk/

SRS is a system that holds sequence and other databases and allows you to search them for words in the annotation, such as keywords, author, title and so on. It holds each type of database individually, unlike Entrez which lumps all databases of the same type (e.g. nucleic acid) into one non-redundant database.

There are various other bibliographic databases and information services available, including:

	BIDS (Medical and science bibliographic databases)
	http://www.hgmp.mrc.ac.uk/Registered/Webapp/bids-isi/
	WISDOM (Wellcome Information Services)
	http://www.hgmp.mrc.ac.uk/Registered/Option/wisdom.html
	NISS (UK Academic National Information Services and Systems)
	http://www.niss.ac.uk/

3.4 Sequence formats

Most molecular biology packages and applications read sequence data from files. Unfortunately, the format of these files is not consistent between applications and programs may not understand your data if it is not in the preferred format; for example, GCG will only read in sequences that are in GCG format. Fortunately, converting your sequence to the format required by the program you wish to use is straightforward. There are many ways to do this, but one of the easiest is to use the program ***readseq***, available from the WWW menu at

http://www.hgmp.mrc.ac.uk/Registered/Webapp/readseq/

You can cut and paste your sequence into the ***readseq*** form or specify a file in which the sequence can be found. You then select your desired output format and filename, and your sequence will be converted when you press the ‘ **Reformat**’ button. From the telnet menu you can start ***readseq*** by typing ‘ *readseq*’ at the telnet prompt.

3.5 Sequence Analysis Packages

3.5.1 EMBOSS

http://www.sanger.ac.uk/Software/EMBOSS/

EMBOSS stands for European Molecular Biology Open Software Suite. EMBOSS is a collaboration of European biological software developers which aims to develop and integrate a range of currently available packages and tools for sequence analysis into a general, publicly available, suite of programs and libraries. You can run a beta version of EMBOSS from
http://www.hgmp.mrc.ac.uk/Registered/Option/emboss.html

As always, we recommend you consult the online documentation for this package before you try to use it.

Historical Background

Since 1988, the sequence analysis package EGCG has provided extensions to the market leading commercial sequence analysis package GCG. EGCG development was a collaboration of groups within EMBnet and elsewhere.

EGCG provided support for core sequence activities at the Sanger Centre, and has been the basis of new sequence analysis software for internal use, as well as providing advanced features in use at approximately 150 sites, and for more than 10,000 users of EMBnet national services.

That project has reached the limits of what we can achieve using the GCG package. As a result, the former EGCG developers have been designing a totally new generation of academic sequence analysis software. This has resulted in the EMBOSS project.

Overview

EMBOSS now proposes to develop a new suite of programs and libraries for sequence analysis and to integrate a range of currently available public packages and tools into a general, publicly available, suite. Applications will be in the general area of sequence analysis, though expansion into related areas is not ruled out. Specific targetted applications which will be in EMBOSS include:

- Rapid database searching with sequence patterns
- Rapid database searching for sequence overlaps
- Simple and species-specific repeat identification
- Nucleotide sequence pattern analysis, for example to identify CpG islands.
- Codon usage analysis for small genomes
- Gene identification tools for genomic sequencing
- Rapid identification of sequence patterns in large scale sequence sets.
- Protein motif identification, including domain analysis
- Presentation tools for publication
- EST clustering

More details on using various EMBOSS applications are given in subsequent sections (for example, Sections 3.9 and 3.10) and worked application examples can be found in Chapter 6.

How to specify sequences for use in EMBOSS

All EMBOSS applications use the Uniform Sequence Address, or USA, for sequence naming. The USA includes sequence files, database queries and external applications. Queries, and individual entries in files that have more than one sequence, use wildcards of "?" for any character and "*" for any string of characters. If these wildcards are used on the command line they need to be hidden in quotes or preceded by a backslash.

Sequence databases can be in a variety of formats, and accessed by a variety of methods defined through a set of control files. For example, EMBL entries could be read by:

- Original EMBL flatfiles using the CD-ROM or Staden indices
- Original EMBL flatfiles using SRS indices
- GCG 9 format using SRS indices
- A query to any SRS web server

Use the program ***showdb*** to see the available EMBOSS databases; an example of using ***showdb*** is given in Section 6.4.3

The USA syntax is one of:

- "file"
- "file:entry"
- "format::file"
- "format::file:entry"
- "dbname:entry"
- "@file"

The "::" and ":" syntax is to allow, for example, "embl" and "pir" to be both database names and formats. The following are valid USAs for sequences:

xxx.seq	A sequence file "xxx.seq" in any format
fasta::xxx.seq	A sequence file "xxx.seq" in fasta format
xxx.seq -sformat=fasta	A sequence file "xxx.seq" in fasta format
embl::paamir.em	A sequence file "paamir.em" in EMBL format
embl:paamir	EMBL entry PAAMIR, using whatever access method is defined locally for the EMBL database
embl:X13776	EMBL entry X13776, searching by accession number and entry name (X13776 is the accession number in this case)
embl-id:paamir	EMBL entry PAAMIR, searching by ID only
embl:paami*	EMBL entries PAAMIB, PAAMIE etc.
embl:*	All sequences in the EMBL database
@mylist	Reads file mylist and uses each line as a separate USA.. List files can contain references to other list files or other standard USA.
list::mylist	Same as "@mylist" above

How to find out what is in EMBOSS - wossname

There are a large number of applications within EMBOSS already, and the package is being actively developed. A list of applications is held at http://www.sanger.ac.uk/Software/EMBOSS/Apps/ You can also produce a list of available applications using the program ***wossname***. A worked example of using *wossname* is given in Section 6.4.2

3.5.2 GCG

http://www.hgmp.mrc.ac.uk/Registered/Option/gcg.html

GCG is an extensive package of applications for molecular biologists, containing programs for sequence editing, fragment assembly, sequence analysis, multiple sequence alignments, database similarity searching and so on. This section describes some essential points you will need to know when using GCG; some individual applications, for example database searching (Section 3.5.2) are discussed in later sections. If you intend to use GCG you should attend one of the HGMP-RC GCG training courses, as a detailed guide to the full range of programs available is beyond the scope of this

document.

Due to the large license fee imposed by GCG on people external to a site using this software, the HGMP-RC have been forced to impose an annual charge on the use of GCG. When you first run GCG each year, you will be asked to pay a fee of [sterling]100 + VAT (academic) or [sterling]500 + VAT (commercial) if you are not employed by the MRC. These figures were correct at the time of going to press but may change subsequently.

The standard GCG package runs in a UNIX session, although there are WWW and other graphical interfaces to it, such as ***W2H*** and ***SeqLab***. We realise that many people will prefer to use these interfaces, and we urge you to do so; however, the character based examples given in Chapter 6 will work for everyone on any machine and contain some fundamental concepts of GCG that are useful to know but are often hidden by the graphical interfaces.

When you start GCG either from the WWW or telnet menu a new UNIX window will appear if you are using X-windows. If you are telnetting to us without using X-Windows, the GCG connection will appear in your telnet window. You need to start the GCG environment before you can run ***SeqLab***, and you need to be in GCG before you can run any of the GCG programs described later. You can run ***SeqLab*** at the prompt by typing:

unix%*seqlab &*	run ***SeqLab***
unix%*seqlab -small &*	run ***SeqLab*** on a small screen

Help Using GCG

There are two main programs for obtaining help when using the GCG package:

genhelpand ***genman***.These can also be viewed from our web pages:

http://menu.hgmp.mrc.ac.uk/people/gcg10/gcghelp/html/unix/gcghelp.html

http://menu.hgmp.mrc.ac.uk/people/gcg10/gcghelp/html/unix/gcgmanual.html

genhelp lists all the standard GCG programs. ***genman***gives the programs grouped by function. They otherwise provide the same information. Two analogous programs, ***egenhelp*** and ***egenman*** describe the programs in the EGCG package, a comprehensive suite of public-domain programs that have been written to work in the GCG style. The help information in these programs is arranged hierarchically. Subtopics describe a program’s function and details of its input and output.

We also have a GCG frequently asked questions section on our web site:
http://www.hgmp.mrc.ac.uk/MANUAL/faq/faq-gcg.html

Another useful guide to using GCG can be found at:

http://www.cbc.med.umn.edu/MBsoftware/GCG/Unofficial_Guide/MolBio_man.html

3.5.3 Staden

http://www.hgmp.mrc.ac.uk/Registered/Option/staden.html

Staden is a suite of programs for performing some of the same functions as the GCG package, plus many useful gel assembly programs. A detailed list of all the programs available in the Staden package and their functions can be found at

http://www.mrc-lmb.cam.ac.uk/pubseq/overview.html

3.6 What can I do with my nucleotide sequence?

The various large scale sequencing projects are producing massive amounts of data. Sometimes it can be difficult to know where to start with the analysis of these sequences. In the following section we have tried to suggest various approaches you might take in studying your latest nucleotide sequence. We do not have space to show you everything that is possible and thus in many cases will point you to additional sources of information.

3.7 Sequence based database searches

3.7.1 BLAST

http://www.hgmp.mrc.ac.uk/Registered/Webapp/blast/

There are many different programs for searching sequence databases with a sequence. ***BLAST*** and ***FASTA*** are the most frequently used; the former is faster although the latter is sometimes more sensitive.

BLAST (Basic Local Alignment Search Tool) is the algorithm used in a family of programs that perform different types of search and use Karlin-Altschul statistics to ascribe significance to their results. The ***BLAST*** programs were designed for sequence similarity searching - for example to identify homologs to a query sequence. They are not generally useful for searching using very short sequences; ***fuzznuc, fuzzpro*** or ***findpatterns*** would be more useful in this case.

We have created a web based form interface to ***BLAST*** that is accessible from the URL given above. You can cut and paste the sequence you would like to use for searching, or alternatively specify a file from which to upload the sequence. You can choose the database(s) to be searched and can change some of the parameters to tailor your search. ***BLAST*** search requests are placed in a queue and run on the machines at the HGMP-RC as soon as possible.

Additional help and information about the ***BLAST*** programs is contained in a FAQ at http://www.hgmp.mrc.ac.uk/MANUAL/faq/faq-dbsearch.html

3.7.2 FASTA

http://www.hgmp.mrc.ac.uk/Registered/Webapp/fasta/

FASTA is a program to search a protein or nucleic acid database for similarity to your test sequence. You will be presented with a menu of databases.

We advise you to use the ***BLAST*** programs for your initial database search. ***FASTA*** takes a long time (hours) to run whereas ***BLAST*** takes minutes. If no good match is found with ***BLAST***, a more sensitive search can then be done with ***FASTA***.

3.7.3 fuzznuc and fuzzpro (EMBOSS)

BLAST and ***FASTA*** have problems searching for sequences less than about 30 bases long. ***fuzznuc*** and ***fuzzpro*** are designed to search for these short sequences - for example, searching primer pairs against a database to check for potential non-specific amplification, or searching for small amino acid patterns within protein sequences. Examples of using ***fuzznuc*** and ***fuzzpro*** are given in Section 6.4.11

3.7.4 findpatterns (GCG)

findpatterns is also useful for searching for ambiguous patterns in sequences or for searching for short sequences in databases. An example of using ***findpatterns*** is given in Section 6.5.7

Additional useful tools can be found at

http://www.hgmp.mrc.ac.uk/GenomeWeb/nuc-db.html

3.8 Gene Identification: NIX

http://www.hgmp.mrc.ac.uk/Registered/Webapp/nix/

NIX (**N**ucleotide **I**dentify **X**) is intended as a tool to aid the identification of interesting regions in genomic or transcribed nucleic acid sequences. There are many useful computer tools that can be used for this; however none of them is completely accurate and it is useful to be able to compare the results from many programs that use different methods and have differing strengths and weaknesses. ***NIX*** integrates many of these programs, runs them on a sequence and displays their results side by side. Such direct comparison allows us to see when many programs have a consensus about a feature.

Rather than giving you total control over how the programs are run by providing innumerable poorly understood choices of argument for each program, ***NIX*** selects reasonable defaults based on whether the sequence is genomic or transcribed, its size, and its species of origin. Repeat sequences are masked using Washington University's ***repeatmasker***program. ***BLAST*** searches are started using the masked sequences against *ecoli*, *est* and e *mbl* (minus *sts*, *est*, *gss* and *htg*), vertebrate complete mRNA, *trembl* and *swissprot* databases. The Expect value cutoff is set to 0.1 and up to a million alignments can be output. The ***BLAST*** results are compressed to save filespace. For transcribed sequences, ***Grail*** is run on the masked sequence to look for exons. For genomic sequences, exons are found using ***Grail***, ***Genefinder***, ***Genemark***, ***Fex***, ***Hexon***, ***Fgene***, ***Fgenes*** and ***Fgenesh***, and ***trnascan_SE*** is also run on the sequence.

This approach is not without problems. You cannot alter parameters to programs to give results under more or less stringent conditions; however, you can run the programs yourself from their web sites or from the options in the HGMP-RC menu, supplying the exact parameters you require. Additionally, many exon finding programs take a species as a parameter. ***NIX*** holds a compromise species list covering the taxonomic groups most frequently available as parameters to the various programs. The mapping from the list on the ***NIX*** form to the species supplied to the program is done as carefully as possible but there are often large mismatches. Please see the documentation on individual programs in the results display for details of this mapping.

3.9 Retrieving sequences from databases

There are many reasons for wanting to retrieve sequences from one of the sequence databases. You may have found the accession number of a sequence of interest from a literature search, or perhaps you want to retrieve the sequences of all members of a family in order to do a multiple sequence alignment. The easiest methods for doing this are offered by the EMBOSS programs***seqret, seqretset*** and***seqretall***. The methods for specifying sequences to be used by EMBOSS programs were described earlier.

3.9.1 Accessing database sequences using EMBOSS programs

The common sequence databases listed in Section 3.3.1 are regularly updated and can be accessed from within EMBOSS using the USA format described in Section 3.5.1

Each of the programs can use database sequences directly, or you can download the sequences yourself using the programs ***seqret*** and ***seqretall***. Examples of using ***seqret*** are given in Section 6.4.4

Sequence Formats

The native sequence databases all have their own distinctive formats. EMBOSS can understand many different sequence formats:

fasta	Input & Output
ncbi	Input & Output
pearson	Input & Output
gcg (gcg9.x)	Input & Output
gcg8 (gcg8.x)	Input
embl	Input & Output
swiss(SwissProt)	Input
genbank	Input & Output
ig (Intelligenetics)	Input & Output
msf (gcg MSF)	Input
clustal	Input
staden	Input & Output
text, plain, raw	Input & Output
asn1	Output
fitch	Output
nbrf (PIR)	Output
phylip	Input & Output
strider	Output
trace	Output
unknown	Input

EMBOSS will automatically recognise all input formats except text/plain/raw. You must specify that you sequence is in raw format by using the USA *raw::myseq*. By default, sequences fetched from the databases using ***seqret*** will be in fasta format, but you can specify your desired output format.

3.9.2 Accessing database sequences using GCG programs

Any sequence you obtain using GCG programs will automatically be in GCG format. Sequence files usually consist of header information followed by the sequence.

We have created a web based interface to get database entries, called ***fetch***: http://www.hgmp.mrc.ac.uk/Registered/Webapp/fetch/ In addition to entering search terms you can specify the output file format and the output file name. The web interface ***fetch*** is not a part of EMBOSS or GCG.

3.10 Graphical Sequence Comparison

One of the best ways to view the alignment of two sequences is using a dot plot. Applications that produce dotplots represent one sequence on the horizontal axis of the plot and the other on the vertical axis. Comparisons between the sequences are made using either window matching or word matching. With window matching, each group of nucleotides within a window on the vertical axis is compared with the corresponding sequence on the horizontal axis. If the comparison scores above a threshold a dot is drawn on the plot. The window then slides on until all comparisons between the sequences have been made; adjusting the window size will change the appearance of the plot. Comparisons based on word matching search for short perfect matches between the sequence; these matches are referred to as words. Word comparison is significantly faster than window searching but requires that the two sequences contain regions of sequence identity rather than sequence similarity.

If a sequence is compared to itself a central diagonal line appears on the dot plot, with any off-centre lines corresponding to internal repeats. When two distinct sequences are compared, any diagonal lines correspond to regions of homology. Breaks in the diagonals represent deletions or insertions.

You have a choice of programs for producing dotplots.

3.10.1 Dotter

http://www.hgmp.mrc.ac.uk/Registered/Option/dotter.html

Dotter is a dot-matrix program with interactive greyscale rendering for genomic DNA and protein sequence analysis. It draws a dotplot of two sequences on the screen, and allows you to adjust the score cutoffs for displaying dots, so the separation between noise and signal can be fine-tuned interactively. You can position a crosshair on the dotplot at a region of interest; the residue alignment of the two sequences at this position is displayed in a separate window.

3.10.2 Dotplots in EMBOSS

There are a few different EMBOSS programs for producing dotplots

- ***dottup*** produces a dotplots of two sequences using word matching
- ***dotmatcher*** produces dotplots of two sequences using window matching
- ***polydot*** compares two sets of sequences, draws a dotplot for each pair of sequences, and reports all identical matches of a specified length

3.10.3 Compare and dotplot (GCG)

These are two programs used in GCG for producing dotplots - ***compare*** creates a dot file which ***dotplot*** displays. In ***SeqLab*** these are run together by default.

3.11 Pairwise sequence alignments

There are three main categories of methods for comparing sequences:

- **Segment methods** compare all overlapping segments of a predetermined length from one sequence with all segments from the other. This is the approach taken with the dotplots described above.

- **Optimal global alignment** methods produce the best overall score for comparing two complete sequences, including a consideration of gaps.
- **Optimal local alignment** methods identify the best local similarities between two sequences, again including a consideration of gaps.

For both local and global approaches the score is affected by the introduction of gaps in the alignment. You should pay particular attention to the gap creation and extension penalties used as defaults in the various alignment programs. You should also consider the scoring matrix to be used in protein:protein comparisons. As with all of our applications, you are strongly advised to read the help available for these programs and/or to attend an HGMP-RC training course so that you will have a better understanding of how to select sensible values for the various parameters required.

3.11.1 EMBOSS

matcher and ***water***produce local alignments between sequences, while ***stretcher*** and ***needle*** produce global ones. The different programs use different algorithms to produce alignments - for database searches or aligning very long sequences, ***stretcher*** and ***matcher*** will probably produce results more rapidly, while ***needle*** and ***water*** produce a more rigorous alignment at the cost of increased computation time. Examples of using these programs are given in Sections 6.4.5 and 6.4.6

3.11.2 bestfit and gap (GCG)

GCG also offers programs for performing local and global alignments, called ***bestfit*** and ***gap***. Both programs align pairs of sequences but use different algorithms for computing the matches - ***bestfit***finds and aligns the best matching regions between the two sequences (local alignment) while ***gap*** finds the best match over the entire length of the sequences (global alignment). ***gap***is the program to use if you are looking for an alignment that considers the entire length of both sequences, while ***bestfit*** is more useful for finding the region of highest similarity between them. ***bestfit*** displays the single best match between two query sequences; that is, the alignment that gives the highest match score where each matching residue increases the overall score and each mismatch reduces the score. ***gap*** tries to align two sequences by introducing gaps to produce the best global alignment between them.

3.12 Multiple sequence alignment

If you have a group of sequences that you would like to align, perhaps to identify consensus regions shared between them, the easiest program to use is probably ***clustal.***

3.12.1 clustal

http://www.hgmp.mrc.ac.uk/Registered/Option/clustal.html

http://www.hgmp.mrc.ac.uk/Registered/Option/clustalx.html

You can use ***clustalw*** to align your sequences. This is a command line based interface to ***clustal***. If you have X-windows enabled on your machine, you can use ***clustalx*** which is a graphical interface to the ***clustal*** program.

3.12.2 emma (EMBOSS)

EMBOSS contains the program ***emma***, which is an EMBOSS-style interface to ***clustalw***. ***emma*** accepts sequences in a file, or a set of database sequences and performs a multiple alignment. The aligned sequences are output in fasta format.

3.12.3 MAGI

Alternatively, you can try ***MAGI***, our web interface to ***clustalw*** found at http://www.hgmp.mrc.ac.uk/Registered/Webapp/magi/. The purpose of ***MAGI*** is to make it easier to enter sequences to be aligned with ***clustalw***. You can enter the data as a single file containing many sequences (either aligned or unaligned) in one of various formats: fasta, MSF (GCG), clustal, PIR or Phylip. The sequences can alternatively be held as individual sequence files or entries in the databases and the required sequences are then specified by entering a list of the names of the files or database entries. They are then aligned using ***clustalw***, and you can look at the results later using a range of multiple sequence viewing and analysis programs.

3.12.4 pileup (GCG)

pileup is a GCG program that takes two or more input sequences (protein or nucleotide) and produces a multiple sequence alignment using progressive pairwise alignments. The program can align up to 500 sequences providing no single sequence in the final alignment exceeds 7000 characters (including gaps). If you use sequences longer than this, ***pileup*** can align fewer of them.

Your input data can be in the form of:

- a GCG list, such as that produced by ***lookup***
- a sequence specification with an asterisk (*) wildcard such as GenEMBL:*
- a file in GCG MSF or RSF format. Details of this format can be found in the GCG manual: http://menu.hgmp.mrc.ac.uk/people/gcg10/gcghelp/html/unix/using_sequences.html

As usual, you should read the documentation that accompanies ***pileup*** and should not blindly accept the default values it offers as you could easily produce incorrect alignments if you do.

3.13 Producing a restriction map

Several hundred restriction enzymes have been isolated and their recognition sites characterised. These recognition motifs are mostly palindromes and vary in length from tetramers to 26-mers. The currently available restriction enzyme recognition sequences are stored in a public database, called REBASE. Various programs can produce restriction maps of nucleotide sequences; these programs use the REBASE entries to scan the sequence of interest. Here we mention two of the most useful restriction map programs that we have available here at the HGMP-RC.

3.13.1 restrict (EMBOSS)

restrict is available from http://www.hgmp.mrc.ac.uk/Registered/Option/emboss.html. It produces a restriction map from an input seqence and currently gives text based output. An example of using ***restrict*** is given in Section 6.4.10

3.13.2 tacg

tacg is available from http://www.hgmp.mrc.ac.uk/Registered/Option/tacg.html. ***tacg*** accepts sequence data in different formats including GCG, fasta, *EMBL* and plain text. The sequence is stripped of non nucleotide symbols, filtered through REBASE to generate restriction sites and outputs cut-frequency summary tables, a linear map (with or without translations), tables of site and fragment data and various other options.

3.13.3 map (GCG)

This program finds restriction sites in an input sequence in GCG format and displays the result as both strands of the sequence together with restriction enzyme sites and conceptual translation products. An example of running ***map*** is given in Section 6.5.5

Additional useful tools can be found in

http://www.hgmp.mrc.ac.uk/GenomeWeb/nuc-anal.html

3.14 Translating nucleotide sequences

If you have a nucleotide sequence you can translate it into a peptide, not least because searching databases with peptide sequences might give you more useful results than searching with nucleotide sequences. Of course, you could use one of the ***BLAST*** family of search algorithms to conceptually translate your sequence before performing the search; indeed, our HGMP-RC web based BLAST interface will use the appropriate BLAST algorithm to match your sequence data with the databases you want to search - see Section 3.7.1. However, if you do want to do the translation yourself, there are programs available to help you.

3.14.1 transeq (EMBOSS)

transeq will translate nucleotide sequence into peptide sequence using one of a variety of codon tables. You can specify particular regions of the sequence to be translated; for example, if you have a genomic sequence you can tell ***transeq*** where the exon:intron boundaries are and it will produce the corresponding protein product. An example of using ***transeq*** is given in Section 6.4.9

3.14.2 backtranseq (EMBOSS)

For phylogenenetic analyses nucleotide sequences may be more useful than protein sequences. Alternatively, you might want to design oligo probes to recognise the coding region for your favourite protein sequence. ***backtranseq***allows you to produce nucleotide sequence from peptide sequence, and offers you the choice of various codon usage tables.

3.14.3 translate (GCG)

The input to ***translate*** is one or more nucleotide sequences from a file in GCG format, a GCG list file, a GCG MSF file or database sequences. ***translate*** can also combine several exon sequences into a single sequence for translation. An example of using ***translate*** can be found in Section 6.5.3

3.14.4 backtranslate (GCG)

backtranslate will convert a peptide sequence into a nucleotide sequence. There are various options for the output from ***backtranslate***:

- you can generate a table of possible backtranslations based on a codon usage table. The possible codons are ordered by their frequency in that table and can help you to make a set of possible oligos to compensate for ambiguous regions.
- you can calculate the most probable back-translation
- you can calculate the most ambiguous back-translation

Additional useful tools can be found in

http://www.hgmp.mrc.ac.uk/GenomeWeb/nuc-anal.html

3.15 Designing primers

We have a list of frequently asked questions about primer design at http://www.hgmp.mrc.ac.uk/MANUAL/faq/faq-primers.html. Here you will find links to programs we have here at the HGMP-RC to help you design primer pairs suitable for your project, including ***prime*** from GCG.

You will find a selection of programs for designing appropriate PCR primers that have web based interfaces at http://www.hgmp.mrc.ac.uk/Registered/Menu/nuc-primer.html

and in the primer FAQ mentioned above. The intention of all these programs is to help you to avoid designing bad primer pairs - for example, a pair that have very different melting temperatures, or form internal secondary structure, or are not sufficiently specific for your purposes. Most ask you to provide your sequence and other parameters such as whether you want to use the primers for PCR or for sequencing. The calculations that these programs make for primer melting temperature are probably more accurate than you would calculate at the bench. The EMBOSS program ***prima*** can also be used to design primers.

Once you have designed your primers it is a good idea to check them against known sequences so that you can reduce the likelihood of non-specific hybridisation. A range of primer databases are listed in the primer FAQ, or you could use ***fuzznuc*** from EMBOSS to screen against EMBL. An example of using ***fuzznuc*** can be found in Section 6.4.11

3.16 What can I do with my protein sequence?

In the following sections we attempt to give you some ideas of useful analyses you can perform with a new protein sequence. Many of these mirror the analyses possible with nucleic acid sequences and where this is the case we will refer you back to the earlier sections. Additionally you may be interested in attending our training courses; details can be found at http://www.hgmp.mrc.ac.uk/About/Courses/

3.17 Searching sequence databases

You can use various programs, including ***BLAST*** and the more sensitive ***FASTA***, to search sequence databases with protein sequences. This topic is discussed in detail for nucleic acids in Section 3.7; the principle is the same for protein sequences.

3.18 Sequence alignment

Once you have performed your homology searches you can gain insights into features that are important for your family of proteins by making pairwise and/or multiple sequence alignments. Again, these were earlier described in detail for nucleic acids; we refer you to Sections 3.10, 3.11, and 3.12 for these discussions.

3.19 Finding motifs and domains in protein sequences

Frequently one of the first things you might do with an unknown sequence is look for known protein motifs. Many databases of protein motifs are available, and are listed at http://www.hgmp.mrc.ac.uk/GenomeWeb/prot-domain.html Proteins can be grouped into families based on their sequence; some sequence motifs have been well conserved during evolution and may therefore be important for protein structure and function. By identifying these motifs it may be possible to produce a signature for a family of proteins. Similarly, the presence of known motifs in a protein may give clues to its structure and function.

One of the more popular of these is PROSITE, found at http://www.expasy.ch/prosite/ PROSITE is a database of protein families and domains containing motifs for a large number of protein families. There are a variety of tools available for scanning PROSITE, including ***ScanProsite*** and ***ProfileScan*** from the PROSITE web page; these allow you to scan a sequence against PROSITE or search a pattern against SWISS-PROT. The EMBOSS program ***patmatmotifs*** also searches for known PROSITE motifs in your sequence as does the GCG program ***motifs*** .

A fingerprint is a group of conserved motifs used to characterise a protein family. Usually the motifs are separated along a sequence, though they may be contiguous in 3D-space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs. ***pscan*** compares your sequence against the PRINTS protein fingerprints database and is a useful complement to ***patmatmotifs.***An example of using ***pscan*** can be found in Section 6.4.8

Pfam, found at http://www.sanger.ac.uk/Software/Pfam/help/faq.shtml, is another popular alternative. This is a database of multiple alignments of protein domains or conserved regions. The latest release of ***Pfam*** contains nearly 1400 families; it is estimated that over half the proteins in SwissProt 35 and TrEMBL-5 have at least one match to a ***Pfam*** family. You can search with your sequence using the form based interface at the ***Pfam*** website. You can also browse the families currently available in ***Pfam***, or can find the ***Pfam*** organisation of any SwissProt entry.

3.19.1 Profile based searching

An alternative to looking for motifs in your protein sequence is to perform a profile based search. A profile is a position specific scoring table that encapsulates features of the family of proteins; profiles can be useful for finding sequences similar to those of the alignment as a whole rather than to individual sequences. Profile based searches tend to be computationally expensive.

EMBOSS's ***prophecy*** creates a profile for you from a group of aligned sequences. The profile contains as many rows as there are positions in the aligned sequences and as many columns as there are residues. Each row (ie position) has a score for each residue that is a measure of the likelihood of the residue appearing in this position. Thus if alanine is conserved amongst all members of a group at position 17, the score for alanine in row 17 will be high and for all other residues will be low.

The profile produced by ***prophecy*** can be used as the input for ***prophet*** (for Gribskov or Henikoff profiles) or***profit*** (for frequency matrices) to search through databases for sequences similar to the profile, or to align individual sequences to the profile. In this way more distantly related proteins may be discovered and the family of proteins extended. Similar functions are performed by GCG's ***profilemake*** and ***profilesearch***.

If you are interested in performing full database searches using profiles you should also look at PSI-BLAST (http://www.ncbi.nlm.nih.gov/blast/psiblast.cgi) and Pfam (http://www.sanger.ac.uk/Software/Pfam)

3.20 Predicting protein secondary structure

The prediction of protein secondary structure from amino acid sequences is not easy. Some of the many programs available to help you to do this can be found at http://www.hgmp.mrc.ac.uk/Registered/Menu/prot-2-struct.html

and http://www.hgmp.mrc.ac.uk/GenomeWeb/prot-2-struct.html

Some programs (for example, ***pepinfo*** in EMBOSS) measure the likelihood of protein secondary structure (alpha helices, beta sheets, turns etc) and hydrophobicity for a single sequence. Others (for example ***DSC*** or ***PsiPred***, available from the web pages above) take a set of aligned sequences as input. There are also programs that attempt to predict transmembrane regions (eg ***tmap***) or subcellular localisation based on primary structure. The best approach is to run a selection of programs over your sequence and compare the results from each. One resource that does this for you is ***Jpred***, found at http://barton.ebi.ac.uk/servers/jpred.html. ***Jpred*** takes either a single protein sequence or a multiple alignment predicts secondary structure by running ***DSC***, ***PHD***, ***PREDATOR***, and ***NNSSP*** over the sequences and combining the results with predictions from ***Mulpred*** and ***Zpred*** to form a consensus result. The consensus approach is more accurate than any individual method but ***Jpred*** allows you to make the final decision on where the most sensible consensus is. The biggest disadvantage of ***Jpred*** is that it has become an extremely popular service and as such analyses may take a long time if the server is busy - your analysis will be placed in a queue behind previously submitted jobs.

3.21 PIX - an integrated approach to protein analysis

http://www.hgmp.mrc.ac.uk/Registered/Webapp/pix/

PIX is a WWW tool written at the HGMP. It runs a selection of protein analysis programs on an amino acid sequence and presents a graphical overview of the results. ***PIX*** runs programs to perform the following analyses:

- Sequence database searches (BLAST searches of SPTR, ecoli, vector and NRL3D)
- Domain database searches (BLAST searches of ProDom and SBASE; ProSearch and PFScan searches of PROSITE; hmmpfam searches of Pfam; Blimps searches of Blocks and PRINTS)
- Cellular localisation (PSORT)

- Secondary structure predictions (PREDATOR, DSC)
- Signal peptide predictions (signal, sigcleave)
- Transmembrane region predictions (TMAP, Tmpred, DAS)
- Coiled coils prediction (Coils)
- Helix-turn-helix prediction (HTH)
- Antigenic sites predicted (antigenic)
- Enzyme digestion (digest)

PIX is intended as a tool to aid in the identification of features in the query amino acid sequence. No prediction algorithm is 100% accurate so it can be very helpful to compare the output of several programs which use different algorithms to predict the same feature. This is analogous to the approach taken by NIX (described in section 3.8) in predicting gene structure. The graphical overview of the output produced by ***PIX*** is an ideal way to compare the outputs of programs side by side allowing you to quickly identify consensus features.

Another advantage of using ***PIX*** is the simple input interface. You do not have to worry about learning the different command line options to the many different programs run by ***PIX***; the parameters to each application are set to sensible defaults by ***PIX***. The disadvantage of this approach is that you have no control over the individual program parameters and cannot tailor them to get the best results for your particular input sequence. The individual programs run within ***PIX*** are all available at the HGMP so that you can run them yourself from the command line and adjust the parameters as you require.

3.22 Predicting protein tertiary structure

The prediction of protein 3D structure from sequence information is an important step towards unravelling protein function and the design of experiments for fully understanding the protein's role. However, it is an extremely difficult problem; sometimes even two proteins that are closely related by sequence may not have similar structures or functions. As yet it is not possible to accurately predict the 3D structure of a protein from its sequence. That said, there are some approaches you can try that might be helpful. homology

3.22.1 Homology modelling

It may be useful to discover whether your protein has significant to one whose 3D structure is already known. The PDB (Protein Data Bank) holds 3D structure data determined by X-ray crystallography and NMR. You can visit the PDB at http://www.rcsb.org/ and search through its holdings. It is advisable to try one of their online tutorials in order to get the most out of this site.

Alternatively you can use the cross referencing facility in some of the sequence databases; for example, if the 3D structure is known for a protein entry in SwissProt, there will be a reference to the PDB entry for that protein.

3.22.2 Fold prediction

You might also try predicting folds; protein are defined as having a common fold if they have same major secondary structures in same arrangement and with the same topological connections. This is something you could try after predicting secondary structure. The SCOP database holds entries for known fold families produced by visually inspecting and comparing known 3D structures and can be accessed at http://scop.mrc-lmb.cam.ac.uk/scop/. Additionally you might try the CATH database at http://www.biochem.ucl.ac.uk/bsm/cath/

Caution should be used as proteins can adopt similar folds despite having no significant sequence or functional similarity and proteins placed in the same fold category may not have a common evolutionary origin.

3.22.3 Additional links

This is an enormous subject and we cannot hope to cover it here. We recommend you look at some of the web pages listed at http://www.hgmp.mrc.ac.uk/Registered/Menu/prot-3-struct.html and http://www.hgmp.mrc.ac.uk/GenomeWeb/prot-3-struct.html.

Additionally, you could attend our course in Protein Structure Prediction. Details can be found at http://www.hgmp.mrc.ac.uk/About/Courses/

You should always remember that wet biology is essential in confirming functional predictions made with the computer.

3.23 Phylogeny

Various programs for performing phylogenetic analyses are available at the HGMP-RC and are accessible from http://www.hgmp.mrc.ac.uk/Registered/Menu/phylo.html

PHYLIP is a package of programs written by Joe Felsenstein for inferring phylogenies and carrying out certain related tasks. It contains many programs for applying various algorithms to different kinds of data and is extremely powerful. PUZZLE is a PHYLIP compatible program that implements the quartet puzzling method for reconstructing tree topologies from character state data. MOLPHY is another package for molecular phylogenetic analysis. In particular, it includes the program PROTML for inferring evolutionary trees from amino acid sequences using the Maximum Likelihood method.

For many users, PIE offers a convenient web based interface for phylogenetics and is available at http://www.hgmp.mrc.ac.uk/Registered/Webapp/pie/ Developed at the HGMP-RC, it provides a front end to various programs for performing phylogenetic analysis including programs from PHYLIP and other phylogeny packages. PIE aims to simplify the generation of phylogenetic trees from a multiple sequence alignment, and also offers the ability to perform techniques such as bootstrapping more easily. If you wish to perform more advanced analyses then the individual phylogenetic programs are accessible from the menus as usual.

Although PIE makes it easier to generate phylogenetic trees it is still crucial that you understand the limitations of phylogenetic analysis. Don’t just perform a single analysis and assume that this must be the correct tree.

3.24 Linkage Analysis

The various programs available at the HGMP-RC for performing linkage analysis can be found at http://www.hgmp.mrc.ac.uk/Registered/Menu/linkage.html

We recommend that you attend one of our Linkage training courses if you plan to do a lot of linkage analysis. For details, see http://www.hgmp.mrc.ac.uk/About/Courses/

3.24.1 GLUE

Many of the programs available for linkage analysis can be difficult to use. A good place to start is GLUE (Genetic Linkage User Environment) which provides a user friendly forms based interface to the Fastlink linkage programs. You can find GLUE at
http://www.hgmp.mrc.ac.uk/Registered/Webapp/glue/

The interface replaces the linkage utilities ***makeped***, ***preplink*** and ***lcp*** and automates the submission of your data to the batch queues for longer jobs. Suggestions for improvements to GLUE are welcomed.

3.25 RHyME - Radiation Hybrid Mapping Environment

http://www.hgmp.mrc.ac.uk/Registered/Webapp/rhyme/

RHyME takes as input the results of a marker typed against the Genebridge4 panel. The marker is analysed in relation to the human 1998 International Gene Map. RHyME uses the RADMAP program to assign your marker to chromosomes and find their best positions on the framework and find the markers that have the most similar vector. It is advisable to also use the Sanger Centre rhmapper facility for further evidence.

You will be emailed when your analysis has completed and you results will be placed in your directory. The results can be viewed using our RHyME results viewer. To improve the radiation hybrid mapping resources, please do submit relevant RH data to the EBI's RHdb.

3.26 PINT - Sequence assembly

http://www.hgmp.mrc.ac.uk/Registered/Webapp/pint/

PINT is a WWW tool for assembling sequence data. It takes the output from sequencers and produces assemblies that can be viewed and edited with popular assembly manipulation programs, using a set of tried and tested default values. It provides a simple interface to a complicated process, using a "data-rich" assembly strategy. It uses ***Phred*** for base-calling, ***Cross_match*** for vector screening, and ***Phrap*** for the assembly. If you want to assemble a set of sequences but you don't have the original chromatograph files, then you should be using something like the

TIGR-Assembler instead of PINT.

---

4. Biological Services

4.1 Biological Resources available from the HGMP-RC

This is not a complete list of all the biological resources available from the HGMP-RC. The collection is constantly changing; resources which are no longer relevant may be discontinued, and new resources are always being added. For an up to date list, please see our World Wide Web page: http://www.hgmp.mrc.ac.uk/Biology/

4.1.1 Genomic Libraries

The HGMP-RC offers a variety of genomic libraries in different vectors as PCR pools and/or high density filters. They include several total human genomic YAC libraries, a human PAC library and CpG island libraries. Single chromosome cosmid libraries are also available as high density filters. Libraries for other species include Drosophila, Fugu (pufferfish), chicken, dog, pig and rat. .

4.1.2 cDNA Libraries

The HGMP-RC is an authorised distributor for the I.M.A.G.E. consortium cDNA clones The number of clones in this collection is at the time of writing nearly 2,500,000 and constantly increasing.

The Centre also supplies mouse embryonic cDNA libraries.

4.1.3 Hybrid Panels

Several different types of hybrid panels are available from the Resource Centre, including the GeneBridge4 radiation hybrid panel and a human/rodent monochromosomal panel.

4.2 Using HGMP-RC Biological Resources: an Overview

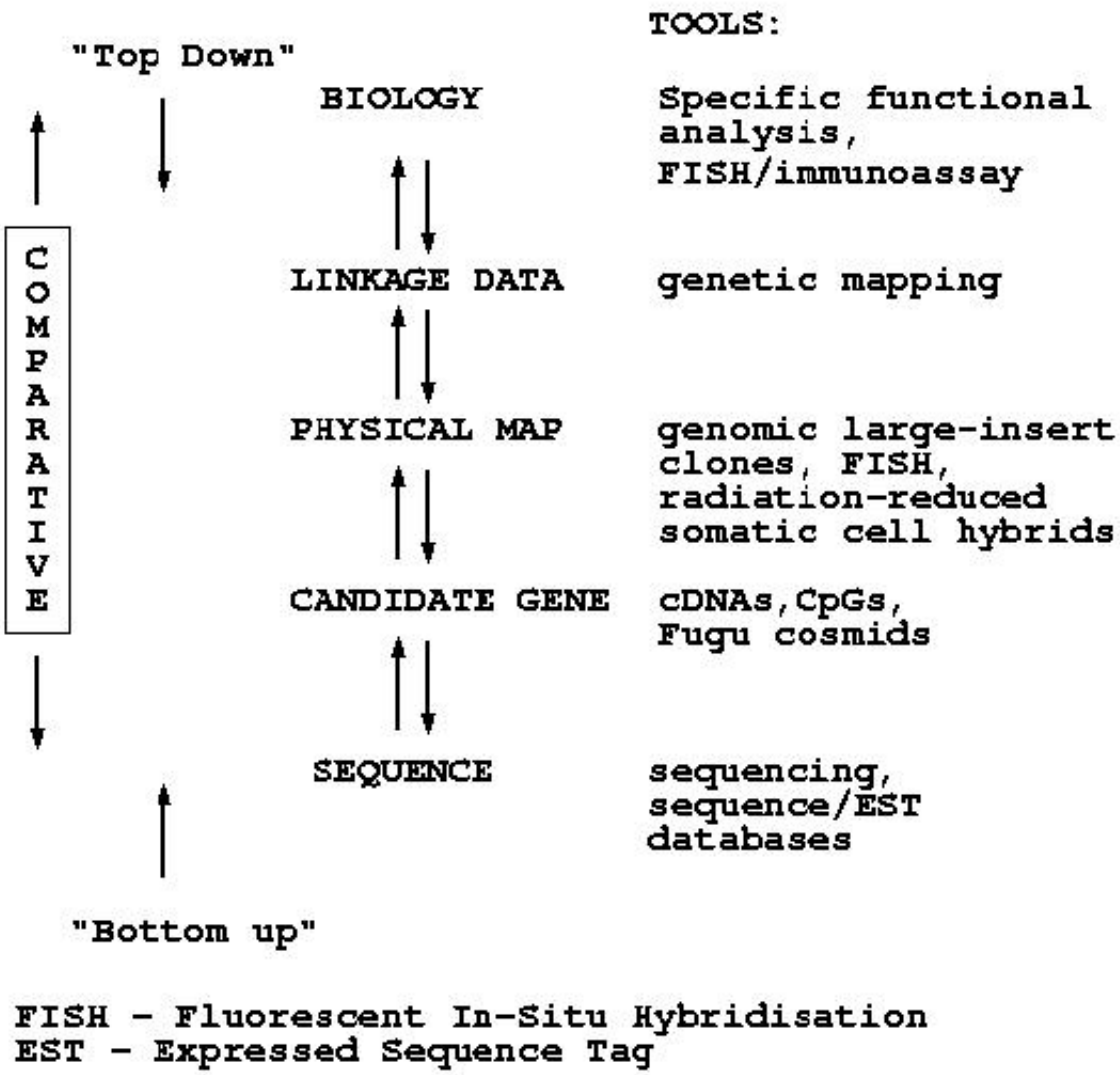

Fig 4.1. *A schematic representation of different research strategies. "Bottom-up" approaches to gene finding make use of sequence data and coding sequence similarities between genes with similar functions, including between different organisms. The "top-down" approach to gene finding relies on association of the biological effect of interest with a genetic map position, followed by physical mapping and candidate gene identification (positional cloning).*

4.2.1 cDNA and Genomic Libraries

Genomic and cDNA libraries play a key role in genetic analysis, regardless of the approach chosen. (See Fig.3.1 for a schematic representation of different research strategies) Together with the development of highly informative genetic markers, the large-insert genomic libraries form the basis of modern genetic maps. They are also used for functional analysis and contig building, with the YAC libraries being of particular importance in the study of gene function. The cDNA libraries form a different and complementary resource; expressed sequences can be identified without the interference of non-coding DNA.

A full list of all the libraries currently stocked by the Resource Centre can be found at http://www.hgmp.mrc.ac.uk/Biology/resources_index.html

4.2.2 cDNA Libraries

cDNA libraries have been constructed by isolating messenger RNA from the cells of interest and synthesising the corresponding cDNA. The cDNA molecules are then cloned into an appropriate vector, making a library representing the particular tissue and developmental stage. The complexity of the resulting library will be considerably lower than a genomic library as it only contains DNA corresponding to expressed genes, although representation of different sequences will differ very widely. The cloned cDNAs can be expressed *in vitro*, facilitating analysis of gene function.

The UK HGMP Resource Centre is an authorised distributor of I.M.A.G.E. clones. There are already almost 2,500,000 clones in our collection, and the number is constantly increasing as new clones become available.

The members of the I.M.A.G.E. (**I**ntegrated **M**olecular **A**nalysis of **G**enomes and their **E**xpression) Consortium have created a high quality cDNA resource from individual libraries, and made it available to all scientists. Sequence, map and expression data for the I.M.A.G.E. clones can be found in public databases. On the basis of this information, the clones will eventually be rearrayed to form a "master array" which ultimately should contain a representative cDNA from every gene in the genome.

Each clone has an IMAGE ID which must be used when an order is placed. This ID can be obtained via dbEST. A materials transfer agreement has to be signed when ordering IMAGE clones for the first time. The agreement form can be printed out from our WWW pages, and need only be signed once, with your first order. The signed contract must be posted to The Resource Centre - **we cannot accept faxed copies**.

The Resource Centre also supplies two mouse embryonic and ten Fugu cDNA libraries (http://www.hgmp.mrc.ac.uk/Biology/descriptions/cdna_resources.html).

cDNA resources are usually screened by hybridisation of high density filters. PCR screening to identify an individual clone is unusual, although PCR can be used e.g. to identify in which of a range of tissue-specific libraries a particular cDNA species can be found.

4.2.3 Genomic Libraries

Genomic libraries, in contrast to cDNA, are constructed by fragmenting genomic DNA and cloning it into a vector capable of accommodating large inserts of DNA. The libraries usually give more than a single coverage of the genome in question which increases the chance of finding the DNA fragment of interest. Only a small proportion (approx. 5%) of the human genome corresponds to genes; thus much of the library is non-expressed.

Systems for cloning large inserts include the yeast artificial chromosome libraries (YACs; Burke *et al*. (1987) Science *236*, 806-812), P1 and P1-derived systems (Sternberg *et al*. (1990) New Biol. *2*, 151-162; Ioannou *et al*. (1994) Nature Genetics *6*, 84-89) and bacterial artificial chromosomes (BACs; Shizuya *et al*. (1992) Proc. Natl. Acad. Sci. USA., *89*, 8794-8797.). Cosmids, with a capacity for inserts of up to 40 kb, are also used.

Genomic libraries are screened by two different methods: PCR based screening and screening by filter hybridisation. PCR screening is convenient and sensitive, and does not usually involve radioactivity. One disadvantage is a high number of both false positives and false negatives. One report (Chumakov *et al*. (1992) Nature *359*, 380-387) claims to have found a rate of false negatives of approx. 20%. For hybridisation screening, the main advantage is that there is no need to know the sequence of the region screened for. Any DNA fragment can be used as a probe. The process can also be faster as the whole library is screened in one step, whereas PCR screening requires multiple steps. A disadvantage, especially with YAC libraries, is that the low yield of insert DNA (see more about this below) causes signals to be weak and problems with both false positives and negatives. False positives may be reduced by duplicate spotting, which has been done with all of the genomic library filters supplied by the Resource Centre.

YAC Libraries

Yeast artificial chromosomes (YACs) have been instrumental in building up the initial physical maps of the human genome. They enable very large fragments to be cloned, which in turn has assisted the building of contigs in a variety of species. A YAC contig map reliably covering about 75% of the human genome was published in the autumn of 1995 (Chumakov *et al* (1995) 'Nature' *377* (*suppl*.), 175-183).

The use of YACs have made it possible to isolate genes responsible for inherited disorders on the basis of their chromosomal location when nothing was known about the defective protein (positional cloning). When a linked marker is found through genetic analysis it is possible to isolate the gene by first isolating a YAC containing the marker and building a contig around it. YACs can also be used in functional analysis by transferring them into mammalian cells or transgenic mice.

There are however disadvantages with YACs, such as chimaerism and instability. A chimaeric YAC clone will contain DNA from different chromosome regions and consequently not represent a contiguous section of DNA. The chimaerism is particularly high in libraries with very large inserts, such as the CEPH 'mega-YAC' library. A further problem is the low yield of insert DNA already referred to in the description of screening methods. On average, only between 2-8% of the total yield of DNA from a yeast culture would represent YAC DNA. This is further complicated by the fact that the YAC DNA is structurally very similar to the yeast DNA, so the usual methods for separation, such as caesium chloride gradients don't work.

The problem with instability can be very frustrating, especially with older libraries, when a clone which was previously isolated as positive for a particular DNA fragment can lose significant portions of its insert irretrievably. Alternative systems have been developed, such as the PAC (discussed below).

YAC libraries can be screened in two different ways: by PCR amplification of DNA pools or by hybridisation. For PCR screening, the DNA is isolated from clone pools of different complexity and embedded into agarose. The positive clone can usually be identified in two rounds of screening.

High density gridded membranes for screening by hybridisation are available for many of the genomic libraries we stock. Interpretation sheets are supplied with the membranes to assist with the identification of positive clones from the high density grid.

PAC and BAC Libraries

P1-derived Artificial Chromosome (PAC) and Bacterial Artificial (BAC) libraries are an important complement to YACs. The insert size is smaller, usually in the range of 100 - 150 kb, although inserts of up to 300 kb have been reported. The possible disadvantage of the smaller insert size is balanced by the fact that there appears to be no problems with chimaerism or instability.

In addition to a human PAC library, PAC and BAC libraries are available from mouse, chicken, pig, Drosophila, rat and Cryptosporidium for comparative mapping purposes.

PAC and BAC libraries are screened like YAC libraries, with the important difference that there is no need to isolate the DNA for PCR screening. An initial denaturation step before the main reaction cycle is sufficient to break down the cells and expose the DNA. In difficult cases, isolating the DNA can help.

Cosmid Libraries

Cosmid libraries have the smallest insert size of the vectors routinely used for genomic cloning. A typical insert size would be in the range of 40 kb. Total genomic libraries have been made for some species, and cosmids have also been used as the cloning vector for creating single chromosome specific libraries, using chromosomes isolated by flow-sorting.

The UK HGMP Resource Centre is an authorised distributor of the human single chromosome cosmid libraries produced at the Lawrence Livermore National Laboratory (LLNL) [Chromosomes: 1, 2, 3, 7, 9,18, 21, 22, X and Y] and at the Los Alamos National Laboratory (LANL) [Chromosomes: 4, 5, 6, 8, 10, 11, 13, 14, 15, 17 and 20].

All of the LLNL libraries are in the cosmid vector Lawrist 16, with the exception of chromosome 2 which consists of three libraries: a cosmid (Lawrist 16) library, a fosmid (pFOS1) library and a PAC (pCYPAC2N) library. Details on the methods used to produce these libraries have been published: Gingrich *et al* 1996 Construction and Characterization of Human Chromosome-2-Specific Cosmid, Fosmid and PAC Clone Libraries. Genomics **32** 65-74. The LANL libraries are in the vector sCOS1 and were constructed according to the method described in Longmire, JL *et al* 1993 Genetic Analysis, Techniques and Applications **10** 69-76. All filters are 4 x 4 array, double spotting, but the number of filters per library varies with the size of the library.

In order to use these libraries users must sign a "Materials Transfer Agreement" form, available from our WWW pages. This contract need only be signed once, at the initial stage. The signed contract must be posted to The Resource Centre - **we cannot accept faxed copies**. There are different MTAs for the LLNL and LANL libraries.

Further information on the libraries is available from:

http://www.hgmp.mrc.ac.uk/Biology/descriptions/human_single_chrom_lib_main.html

The pufferfish (*Fugu rubripes*) has a genome size of only about 400 Mb, although it has essentially the same number of genes as the human genome. This genome compression coupled with the homology make it an ideal model organism for gene identification studies (Baxendale *et al*. (1995) Nature Genetics *10*, 67-76.). A Fugu cosmid library is available from the HGMP-RC as high density gridded membranes for hybridisation.

A landmark map of the Fugu genome is being generated at the Resource Centre, using the cosmid library described above. The aim of the project is to sequence 1000 Fugu cosmids using a scanning approach. Between 50 and 100 random sequences (400-500 bp per scan) are being obtained per

cosmid. These sequences are analysed by BLAST search of the EMBL and SwissProt databases, minimally edited and placed in the EMBL sequence database. This database is accessible through our WWW page at http://fugu.hgmp.mrc.ac.uk/

CpG Island Libraries

CpG islands are short stretches of DNA containing a high density of non-methylated CpG dinucleotides. They usually occur associated with coding regions, and it has been estimated that 60% of human genes have CpG islands attached to them. A CpG island is thus effectively a genomic library which has been enriched for coding sequences.

A human CpG island library (Cross *et al*. (1994) Nature Genetics *6*, 236-244), as well as libraries for mouse, chicken and pig, are available from the HGMP-RC as *E.coli* XL1-BlueMRF culture (~ 10^7 cells/0.1 ml). Primers flanking the cloning site are supplied with the library.

The Sanger Centre has systematically sequenced clones picked from the human library, and submitted the sequences to the EMBL and Genbank databases. The sequenced clones are available from the Resource Centre.

4.2.4 Hybrid Panels

Hybrid panels offer an alternative method for mapping and localisation of markers. Monochromosomal somatic cell hybrids can be used for the initial allocation of an unknown marker to the correct chromosome. Whole genome radiation hybrid mapping panels (Gyapay *et al*., (1996) Human Molecular Genetics *5*, 339-346) can be used for the construction of high-resolution, contiguous maps. A major advantage is that non-polymorphic markers such as expressed sequence tags (ESTs), which are uninformative in linkage mapping, can be used. The distance estimated by this method is directly proportional to physical distance.

4.2.5 Support

If you need help with any aspect of our service, please contact the biology helpdesk by emailing biohelp@hgmp.mrc.ac.uk or by telephoning us on the contact number listed at the beginning of this guide.

---

5. Frequently Asked Questions

To be conscious that you are ignorant is a great step toward knowledge

Anon.

This section attempts to answer some of the questions that we are most frequently asked.

5.1 General Questions

5.1.1 I've forgotten my password, what do I do?

Phone the Computing Helpdesk (01223 494520) or use our simple web form at

http://www.hgmp.mrc.ac.uk/cgi-bin/password_gone.pl

You will be issued with a new password that will be sent to you by post.

5.1.2 How do I register to use the HGMP-RC?

To use the services of the HGMP-RC, you first have to become a registered user. Registrationis free of charge to all academic users and is simply a matter of filling in the application form and returning it to us. The annual registration fee for commercial users is at the moment [sterling]5,000, and this allows up to 10 individual registrations from the same company.

The form is on our publicly-accessible WWW pages at

http://www.hgmp.mrc.ac.uk/About/Registration, or it can be obtained by contacting HGMP-RC Administration. Commercial companies wishing to become registered users should contact our Administration for a registration form, or use the web form. The form should be returned to us by post as we cannot accept faxed or emailed copies.

5.1.3 I have changed my address/name/title/other details, do I need to re-register?

No, you do not need to re-register - but please do keep us informed. Phone, fax, write to, or email Administration and tell us the new information. If you do not inform us of a change of address, your requests may be delayed, you will not receive your copy of *Genome News*, and your user status may be cancelled. You can reach us on:

Tel:	01223 494500
Fax:	01223 494512
Email	admin@hgmp.mrc.ac.uk

Or you can use our simple web form at

http://menu.hgmp.mrc.ac.uk/menu-bin/change_address.pl

5.1.4 How do I register for an HGMP-RC training course?

Phone, fax or email Administration to reserve a place on your preferred course giving details of your name, department, and full address. You will receive confirmation of your place, or an offer of a place on the next available course if your first choice is fully booked. Course notes and details will be dispatched two weeks before the date of the course.

5.1.5 How much do the goods and services cost?

Computing and biological resources and services are supplied free to registered UK academic users, except for the small fee for using the GCG package, and fees for attending training courses. Current prices for non-UK or commercial registered users, and for unregistered users, are available on our WWW pages or by contacting Administration.

5.2 Computing Questions

See our Frequently Asked Questions (FAQ) page at

http://www.hgmp.mrc.ac.uk/MANUAL/faq/.If you have problems doing this, see the earlier sections of this manual on accessing the HGMP-RC via the World Wide Web, or contact the HGMP-RC Computing Helpdesk.

The Computing FAQ has sections on:

* Who can I get help from?

* Terminals and Networking

* WWW

* News and Mail

* Files and Disk Space

* Programs

* Sequence Databases

* OMIM/GDB

* GCG

* Phylogeny

* Linkage

5.3 Biological Questions

5.3.1 How long should I wait between sending in an order and receiving the goods?

We aim for goods to arrive within two weeks of order. You will be informed directly if there is an exceptional delay in processing an order. You can now track the progress of your order using our order tracking facility at

http://menu.hgmp.mrc.ac.uk/menu-bin/CheckMyOrders

5.3.2 How do I obtain order/request forms?

For the fastest service, submit your order online using the request forms on our WWW pages. Requests can also be submitted by fax or post if desired by printing out the request form and completing it by hand.

5.3.3 How do I find out if a YAC has already been identified for the gene (the chromosome, the region) I'm interested in?

There are a number of databases that you can screen to avoid duplicating what somebody else has already done. GDB, which is accessible through the HGMP-RC menu, is a good place to start. If you are not familiar with GDB, our computing helpdesk can advise you; alternatively, you can attend one of our computing courses to learn more about how to access genome databases.

Other databases with this type of information include RLDB (the ICRF reference library database) and the CEPH-Généthon database. The latter only contains data on their own markers (the AFM microsatellites).

5.3.4 What is meant by pools for PCR screening?

The sensitivity of PCR makes it possible to find a specific clone in a library containing tens of thousands of clones. The clones in a library are stored as glycerol stocks in microtitre plates. Clones from a number of plates are pooled together, making up what is referred to as primary pools. The first step is to determine in which of these pools is the clone you're looking for. Sub-pools of the same primary pool (consisting, for instance, of clones from one plate each) are then screened, narrowing down the number of possibilities. The sub-pools are usually arranged in a three-dimensional way so that screening a limited number gives you the 'address' of the positive clone.

5.3.5 I have screened a YAC library and found positive pools, but I can’t find a positive clone.

This unfortunately happens occasionally. As described in the overview, the YACs can be unstable and delete the part of the insert which you are screening for. When you receive the clone, you should streak out for single colonies and test a large number of these, as it is possible that the colony is mixed and the non-deleted version will be found among them. If this fails, we can go back to our archive copy in case this still has the original insert.

5.3.6 How do I find out what cDNA libraries you have?

A comprehensive list can be found on our WWW pages; please see
http://www.hgmp.mrc.ac.uk/Biology/resources_index.html

5.3.7 What do I do if the clone you have sent us doesn’t grow?

If your clones do not grow, inform us as soon as possible and we’ll send you a new slope/plate, if necessary using a different copy of the library.

5.3.8 I want to request an I.M.A.G.E. clone but I do not know the I.M.A.G.E. ID.

You can find out the I.M.A.G.E ID by querying dbEST for the sequence you are interested in:

http://www2.ncbi.nlm.nih.gov/dbST/dbest_query.html

In the database entry, look in the **Comment** field for a reference to the I.M.A.G.E. consortium. The I.M.A.G.E. ID can be found under **Definition** near the top of the screen. It is a five or six digit figure, with *no* preceding letter, often followed by 5’ or 3’. Any ID codes given under **Comment** should be ignored *; the proper I.M.A.G.E. ID is the one given at the top of the screen under **Definition.*** Also see our help pages at http://www.hgmp.mrc.ac.uk/Biology/descriptions/image_IDs.html

5.3.9 I need technical information on your cDNA/genomic libraries, including vector, average insert size and restriction sites for cutting.

Much of this information can be found under "Biological Services" on our WWW pages. If you want to discuss a specific problem, please contact biohelp by email or by telephoning the Biology Helpdesk [01223 494510], and you will be put in touch with the relevant member of staff.

5.3.10 Why can’t I order the I.M.A.G.E clone I want?

We receive regular deliveries of new plates from the I.M.A.G.E Consortium. As soon as these plates are replicated and the clones are then made available this information is added to our WWW pages dealing with I.M.A.G.E clone availability:

http://www.hgmp.mrc.ac.uk/Biology/descriptions/image_availability.html

This WWW page is regularly updated so if you consult it at frequent intervals you will be able to see when new clones become available. Unfortunately we cannot predict dates for release as we do not know in advance when certain clones will reach us.

In addition, there are certain clones we cannot supply as they have been withdrawn due to phage contamination. For full details please see:

http://www.hgmp.mrc.ac.uk/Biology/descriptions/image_phage_contamination.html

5.3.11 I’ve forgotten which I.M.A.G.E ID I ordered and only have the clone name

On your dispatch note you will find the clone name against the I.M.A.G.E ID. However, if you no longer have this you can translate clone name into I.M.A.G.E ID at http://www.hgmp.mrc.ac.uk/Biology/PlateConversion.html

---

Any Comments, Questions? Support@hgmp.mrc.ac.uk

6. Application Worked Examples

Everything should be as simple as it is, but not simpler.

Albert Einstein (1879-1955)

6.1 Sequence based database searches: BLAST

The easiest way to run ***BLAST*** at the HGMP-RC is to used our web-based interface at http://www.hgmp.mrc.ac.uk/Registered/Webapp/blast/

The interface to ***BLAST*** is a form with various fields for you to fill in with your data, and options to choose the databases to search. The small ' **i**' next to some of the options is a link to additional help as to the type of data that is expected for that field. Here is an example for searching databases using ***BLAST***:

From the main WWW-menu: (http://www.hgmp.mrc.ac.uk/Registered/Menu/)	
	Choose ' **BLAST - Fast Database Similarity Search'** from the ' **Integrated Analysis Services'** section.
	Fill in your sequence file name, or paste the sequence in. For example, you could use Entrez or SRS to find and display a sequence and then copy and paste this in. You need to tell the program whether your sequence is nucleic acid or protein so that it can select the appropriate ***BLAST*** algorithm. You can also enter a description to help you recognise the search later; this is particularly useful if you will be running more than one search.
	Choose the database(s) you want to search in - for example, click on the **Fugu Project sequences** and **Primates** boxes.
	Have a look through the other options; the filtering can be useful. Choose other options as desired.
	If you would like to have your results emailed to you, make sure that the address in the **Enter your e-mail address** box is correct. The system should have filled in the correct address but you can change it if you need to.
	Finally, click on **Do the search**.

Your search is now set up and will be automatically queued and run. You will receive email with the resultant files when the search is complete. The results will be back within minutes, hours or possibly a day, depending upon how busy the database searching queues are.

From the telnet menu, you can run this search by typing **blast** at the menu prompt.

6.2 Sequence based database searches: FASTA

The HGMP-RC ***FASTA*** web interface is very similar to the ***BLAST*** interface:

From the main WWW-menu: (http://www.hgmp.mrc.ac.uk/Registered/Menu/)	
	Choose ‘ **FASTA - Database Similarity Search’** from the ‘ **Integrated Analysis Services’** section.
	Fill in your sequence file name, or paste the sequence in. Tell the program whether your sequence is nucleic acid or protein, plus a description if desired.
	Choose the database(s) you want to search in.
	Choose other options as desired.
	Check that the address in the **Enter your e-mail address** box is correct.
	Finally, click on **Do the search**.

From the telnet menu, you can run this search by typing **fasta** at the menu prompt.

6.3 Changing file formats with ReadSeq

readseq will convert sequences between the different file formats used by various molecular biology programs. We have written a forms based interface to ***readseq***:

From the main WWW-menu: (http://www.hgmp.mrc.ac.uk/Registered/Menu/)	
	Click on **Nucleic** Then **Sequence Editors and formats** Then **ReadSeq** Then **Run READSEQ Now!**
	Fill in your sequence file name, or paste the sequence in.
	Choose the output format you require.
	Provide a filename for the file that will be produced.
	Click on **REFORMAT**
	You can view the converted sequence by using *more* on the output file in a UNIX window.

From the telnet menu, you can run this program by typing **readseq** at the menu prompt.

6.4 Sequence Analysis: EMBOSS

EMBOSS already contains many applications and the list available is always growing. We cannot show them all to you here; we will give worked examples of a few applications to get you started, and refer you to the EMBOSS web pages for details of the others:

http://www.sanger.ac.uk/Software/EMBOSS/

6.4.1 Starting EMBOSS

If you are using the WWW menu: (http://www.hgmp.mrc.ac.uk/Registered/Menu/)	
	Click on **Nucleic**
	Then **General Sequence Analysis**
	Then **EMBOSS**
	Then **Run EMBOSS Now!**
	If you are using X-windows a new window will pop up.
	If you choose to use Java a new screen will pop up in your browser and the UNIX session will run in an applet.
If you are using the Telnet menu:	
	Type *emboss* at the telnet prompt.
	If you are using X-Windows, a new window will pop up
	If you do not have X installed on your machine, the EMBOSS session will run in your telnet window.

Some help messages will appear on your screen and you will see a UNIX prompt (something like **unix %)**displayed. From now on we will refer to this as 'the prompt'. All the following examples assume you have started EMBOSS before attempting to run the programs.

In these examples, what you see on the screen is represented in **bold** type, and what you should type in is represented in *italics*.

6.4.2 wossname

wossnamesearches the EMBOSS documentation for a keyword that you enter; you can use it to produce lists of programs that perform different tasks, or a list of all the programs currently in EMBOSS.

	At the prompt, start ***wossname***by typing *wossname*		
		unix % *wossname*	
		Finds programs by keywords in their one-line documentation	
Keyword to search for: *alignment*			
SEARCH FOR 'ALIGNMENT'			
emma	**Multiple alignment program - interface to ClustalW**		
matcher	**Finds the best local alignments between two sequences**		
needle	**Needleman-Wunsch global alignment**		
prophecy	**Creates matrices/profiles from multiple alignments**		
prophet	**Gapped alignment for profiles**		
simplesw	**Simple Smith-Waterman alignment**		
stretcher	**Finds the best global alignment between two sequences**		
water	**Smith-Waterman local alignment**		

Many programs have additional parameters that can be seen by appending the flag

-opt to the program name. A default for each option is given in square brackets: either press *return* to accept it, or enter the value you require.

Start ***wossname*** again, this time using the ***-opt*** flag:

	unix % *wossname -opt*
Output program details to a file [stdout]: *myfile*	
Format the output for HTML [N]:	
Output only the group names [N]:	
Output an alphabetic list of programs [N]:	

This time the output has been written to *myfile*.

wossname can also produce a list of all EMBOSS programs. Run it again, but this time press *return* instead of specifying a keyword. Scroll up and down the list of programs that appears on your screen to see them all. How could you get this data into a file? (Hint: use -opt)

You can see all the command flags available for any EMBOSS program by using the flag ***-help***. For example:

```
unix % wossname -help
```

6.4.3 showdb

EMBOSS reads from sequence databases provided the sequence is referred to in the form **database:entry**. Use ***showdb*** to see the databases currently available at the HGMP:

```
unix % showdb
Displays information on the currently available databases
#Name        Type  ID  Qry  All  Comment
swissprot    P     OK  OK   OK   -
pir          P     OK  OK   OK   PIR/NBRF
nbrf         P     OK  OK   OK   PIR/NBRF
sw           P     OK  OK   OK   -
embl         N     OK  -    -    EMBL HGMP IDS
.
.
.
```

showdb writes a table displaying database names, types (**P**rotein or **N**ucleic acid) and access methods:

- **ID** (programs can extract a single named database entry eg *embl:x13776*)
- **Query** (programs can extract a set of matching wildcard entry names eg *swissprot:pax*_human*)
- **All** (programs to analyse all the entries in the database eg *embl:**)

6.4.4 seqret

seqret reads in a sequence, and writes it out. You can specify sequences by accession number or by sequence identifier:

Using sequence identifiers

	unix % *seqret*
	Reads and writes (returns) a sequence
	Input sequence: *embl:xlrhodop*
	Output sequence [xlrhodop.fasta]:
	unix % *more xlrhodop.fasta*
	>XLRHODOP L07770 Xenopus laevis rhodopsin mRNA, complete cds.
	agtagaacagcttcagttgggatcacaggcttctagggatcctttgggcaaaaaaga aacacagaaggcattctttctatacaagaaaggactttatagagctgctaccatgaa cggaac
	.
	.
	.

Using accession numbers

You can also retrieve sequences using accession numbers; at the HGMP, that means you may need to use ***embla*** - the names of databases will change, so you should always use ***showdb*** to check them.

```
unix % seqret
Reads and writes (returns) a sequence
Input sequence: embla:L07770
Output sequence [xlrhodop.fasta]: xlrhodop2.fasta
unix % more xlrhodop2.fasta
>XLRHODOP L07770 Xenopus laevis rhodopsin mRNA, complete cds.
gtagaacagcttcagttgggatcacaggcttctagggatcctttgggcaaaaaaga
acacagaaggcattctttctatacaagaaaggactttatagagctgctaccatgaa
ggaac
.
.
.
```

You could also run this example entirely from the command line.

```
unix % seqret embl:xlrhodop -outseq xlrhodop.fasta
```

By default, ***seqret*** writes the sequence in fasta format. You can select different output formats:

```
unix % seqret embl:xlrhodop -outseq xlrhodop.gcg -osf gcg
```

EMBOSS can also read sequences from files. For example, we can reformat our fasta sequence into gcg format:

```
unix % seqret xlrhodop.fasta -outseq xlrhodop.gcg -osf gcg
```

If we wanted to be really careful we could make sure ***seqret*** knew this sequence was fasta format:

```
unix % seqret fasta::xlrhodop.fasta -outseq xlrhodop.gcg -osf gcg
```

You have seen some of the command line options available for use with ***seqret***. To see the full range, type `seqret -help` at the **`unix %`** prompt.

6.4.5 Pairwise alignments: water

We will align a cDNA to a genomic sequence to clearly show the introduction of gaps (due to introns) in two pairwise alignment methods. The gene we are using is *Xenopus laevis* rhodopsin and has five exons. ***water*** uses the Smith-Waterman algorithm for finding local alignments.

unix % *water*

Smith-Waterman local alignment.

Input sequence: *embl:xlrhodop*

Second sequence: *embl:xl23808*

Gap opening penalty [10.0]:

Gap extension penalty [0.5]:

Output file [xlrhodop.water]:

unix % *more xlrhodop.water*

. . .

XLRHOD	**272**	**aacttcatgaccttgtttgttaccatccagcacaagaaactcaga**	**316**
XL23808	**1452**	**||||||||||||||||||||||||||||||||||||||**	**1496**
		aacttcatgaccttgtttgttaccatccagcacaagaaactcaga	
XLRHOD	**317**	**acacccctaaactacatcctgctgaacctggtatttgccaatcac**	**361**
XL23808	**1497**	**||||||||||||||||||||||||||||||||||||||**	**1541**
		acacccctaaactacatcctgctgaacctggtatttgccaatcac	
XLRHOD	**362**	**ttcatggtcctgtgtgggttcacggtgacaatgtacacctcaatg**	**406**
XL23808	**1542**	**||||||||||||||||||||||||||||||||||||||**	**1586**
		ttcatggtcctgtgtgggttcacggtgacaatgtacacctcaatg	
XLRHOD	**407**	**cacggctacttcatctttggccaaactggttgctacattgaaggc**	**451**
XL23808	**1587**	**||||||||||||||||||||||||||||||||||||||**	**1631**
		cacggctacttcatctttggccaaactggttgctacattgaaggc	
XLRHOD	**452**	**ttctttgctacacttggt.....**	**469**
XL23808	**1632**	**||||||||||||||||**	**1676**
		ttctttgctacacttggtggtaagttccaatgggctttcgtcact	
XLRHOD	**1677**	**.....**	**1721**
XL23808		**gatattgttgtagcaataaattcttggaaagctcgtaagggaaca**	

. . .

The vertical bars (**|**) represent bases that are conserved between the two sequences, and the dots (**.**) represent gaps. We've only shown part of the output as it is very long. You should look at the whole output and note that there are five aligned regions that represent the five exons as predicted from the dotplot.

6.4.6 Pairwise alignment:needle

needle uses the global alignment algorithm of Needleman and Wunsch to align two complete sequences, maximizing matches and minimizing gaps.

	unix %*needle*
	Needleman-Wunsch global alignment.
	Input sequence: *embl:xlrhodop*
	Second sequence: *embl:xl23808*
	Gap opening penalty [10.0]:
	Gap extension penalty [0.5]:
	Output file [xlrhodop.needle]:

Look at the output file *xlrhodop.needle*: the alignment doesn't look nearly as convincing as the output from ***water***. This illustrates a valuable point: when running ***needle*** we accepted the default gap penalties - but remember, the program doesn't know anything about your sequence and you shouldn't trust it to provide sensible parameter values. We were lucky with ***water*** that the default values gave us a ''good'' answer. Try rerunning ***needle***:

	unix % *needle*
	Needleman-Wunsch global alignment.
	Input sequence: *embl:xlrhodop*
	Second sequence: *embl:xl23808*
	Gap opening penalty [10.0]: *3*
	Gap extension penalty [0.5]: *0.3*
	Output file [xlrhodop.needle]: *xlrhodop2.needle*

Look at this output file (type *more xlrhodop2.needle* at the **unix %** prompt); you should see a more convincing result.

Why did we choose these values? The simple answer is we experimented until we found values that gave us the five exons we were expecting.

6.4.7 Motif searching: patmatmotifs

Many protein families can be recognised by a specific 'fingerprint' or 'motif'. ***patmatmotifs*** looks for sequence motifs by searching your protein sequence for PROSITE patterns. PROSITE (http://www.expasy.ch/prosite/) is a database of protein families and domains. Proteins or protein domains belonging to a particular family generally share functional attributes and are derived from a common ancestor.

	unix % *patmatmotifs -full*			
	Search a motif database with a protein sequence			
	Input sequence: *sw:opsd_xenla*			
	Output file [opsd_xenla.patmatmotifs]: *xlrhodop.pat*			
	unix % *more xlrhodop.pat*			
	Number of matches found in this Sequence = 1			
	Length of the sequence = 354 basepairs			
	Start of match = position 123 of sequence			
	End of match = position 139 of sequence			
	Length of motif = 17			
	patmatmotifs of G_PROTEIN_RECEPTOR with OPSD_XENLA from 123 to 139			
	TLGGEVALWSLVVLAVERYMVVCKPMA			
		|		**|**
		123		**139**
	** *** G-protein coupled receptors signature *** **			
	G-protein coupled receptors [1 to 4,E1,E2] (also called R7G) are an extensive group of hormones, neurotransmitters, odorants and light receptors which transduce extracellular signals by interaction with guanine nucleotide-binding (G) proteins. The receptors that are currently known to belong to this family are listed below.			
	Number of matches found in this Sequence = 1			
	. . .			

We already know that our sequence is a rhodopsin. However, we hope you can see that identifying motifs in an unknown sequence can provide information to help you plan further experiments.

6.4.8 Protein fingerprints: pscan

A fingerprint is a group of conserved motifs used to characterise a protein family. Usually the motifs are separated along a sequence, though they may be contiguous in 3D-space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs. ***pscan*** compares your sequence against the PRINTS protein fingerprints database and is a useful complement to ***patmatmotifs***.

	unix % *pscan*
	Scans proteins using PRINTS
	Input sequence: *sw:opsd_xenla*
	Minimum number of elements per fingerprint [2]:
	Maximum number of elements per fingerprint [20]:
	Output file [opsd_xenla.pscan]: *xlrhodop.pscan*
	Scanning OPSD_XENLA...
	unix % *more xlrhodop.pscan*
	CLASS 1
	Fingerprints with all elements in order
	Fingerprint GPCRRHODOPSN Elements 7
	Accession number PR00237
	Rhodopsin-like GPCR superfamily signature
	Element 1 Threshold 54% Score 61%
	Start position 39 Length 25
	Element 2 Threshold 49% Score 49%
	Start position 72 Length 22
	Element 3 Threshold 48% Score 55%
	Start position 117 Length 23

6.4.9 transeq

transeq will translate nucleotide sequences into peptide sequences. You can select the frame to be translated and can specify ranges to be included in the translation. For example, if we look at the original EMBL entry for out rhodopsin gene (for example, using SRS), we will see that the start and end points of the exons have been included in the sequence annotation. We can pass this information into ***transeq*** and translate only the exons:

```
unix % transeq embl:xl23808 -regions''1290..1650,
1899..2067, 2669..2834,3085..3324,4030..4158''
-outseq xl23808.pep
Translate nucleic acid sequences
unix % more xl23808.pep
>XL23808_1 Xenopus laevis rhodopsin gene, complete cds.
MNGTEGPNFYVPMSNKTGVVRSPFDYPQYYLAEPWQYSALAAYMFLLILLGLPINF
MTLFVTIQHKKLRTPLNYILLNLVFANHFMVLCGFTVTMYTSMHGYFIFGQTGCYI
EGFFATLGGEVALWSLVVLAVERYMVVCKPMANFRFGENHAIMGVAFTWIMALSCA
APPLFGWSRYIPEGMQCSCGVDYYTLKPEVNNESFVIYMFIVHFTIPLIVIFFCYG
RLLCTVKEAAAQQQESATTQKAEKEVTRMVVIMVVFFLICWVPYAYVAFYIFTHQG
SNFGPVFMTVPAFFAKSSAIYNPVIYIVLNKQFRNCLITTLCCGKNPFGDEDGSSA
ATSKTEASSVSSSQVSPA*
```

6.4.10 restrict

restrict can be used to identify cut sites for restriction enzymes within a nucleic acid sequence. By default, the cut sites are output in the order they appear along the sequence. In this example, we will use the flag ***-alpha*** to cause the enzymes that cut the sequences to be listed alphabetically. We will search for enzymes that cut a minimum of once and a maximum of twice, and have a recognition site length of at least six bases:

```
unix % restrict -min 1 -max 2 -alpha
Finds restriction enzyme cleavage sites
Input sequence: embl:xl23808
Minimum recognition site length [4]: 6
Comma separated enzyme list [all]:
Output file [xl23808.restrict]:
Scanning XL23808...
unix % more xl23808.restrict
# Restrict of XL23808 from 1 to 4734
#
# Minimum cuts per enzyme: 1
```

	# Maximum cuts per enzyme: 2					
	# Minimum length of recognition site: 6					
	# Blunt ends allowed					
	# Sticky ends allowed					
	# DNA is linear					
	# Ambiguities allowed					
	# Base Number	Enzyme	Site	5′	3′	[5′ 3′]
	829	AatI	AGGCCT	831	831	
	1762	Acc113I	AGTACT	1764	1764	
	3851	AclI	AACGTT	3852	3854	
	2878	AclNI	ACTAGT	2878	2882	
	4210	AflII	CTTAAG	4210	4214	
	408	AgeI	ACCGGT	408	412	
	418	Ama87I	CYCGRG	418	422	
	1217	BamHI	GGATCC	1217	1221	
	3249	BbsI	GAAGAC	3237	3241	
	794	BbuI	GCATGC	798	794	
	492	BcgI	CGANNNNNNTGC	481	479	515 513
	3438	BciVI	GTATCC	3427	3426	
	161	BglII	AGATCT	161	165	

6.4.11 fuzznuc and fuzzpro

fuzzpro searches for patterns in protein sequences, allowing mismatches. You can use PROSITE syntax to define your pattern - so, for example, **[DE] - x(2) - {ED}** defines a four residue pattern consisting of an acidic residue, followed by any two residues, followed by a non-acidic residue. For example, here we are searching for the G-protein-coupled-receptor motif as defined in PROSITE:

```
unix % fuzzpro
Protein pattern search
Input sequence: sw:opsd_xenla
Search pattern: [GSTALIVMFYWC]-[GSTANCPDE]-{EDPKRH}-x(2)-[LIVMNQGA]-x(2)-[LIVMFT]-[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSH]-x(2)-[LIVM]
Number of mismatches [0]:
Output file [opsd_xenla.fuzzpro]:
unix % more opsd_xenla.fuzzpro
OPSD_XENLA 123 VALWSLVVLAVERYIVV
```

If we allow three mismatches, we see additional hits to our sequence:

```
unix % more opsd_xenla.fuzzpro
OPSD_XENLA 52 LPINFMTLFVTIQHKKL
OPSD_XENLA 72 LNYILLNLVFANHFMVL
OPSD_XENLA 89 CGFTVTMYTSMHGYFIF
OPSD_XENLA 123 VALWSLVVLAVERYIVV
OPSD_XENLA 149 GENHAIMGVAFTWIMAL
OPSD_XENLA 165 LSCAAPPLFGWSRYIPE
```

fuzznuc *searches for patterns in nucleic acid sequences.*

6.5 Sequence Analysis: GCG/EGCG

We cannot give an example of every GCG program here. We have chosen a few that you will probably use frequently to give you an idea of how the command line interface works.

6.5.1 Starting GCG

If you are using the WWW menu: *(http://www.hgmp.mrc.ac.uk/Registered/Menu/)*	
	*Select **gcg10** from the subsection titled **Common options***
	If you are using X-windows a new window will pop up.
	If you choose to use Java a new screen will pop up in your browser and the UNIX session will run in an applet.
If you are using the Telnet menu	
	Type gcg at the telnet prompt .
	If you are using X-Windows, a new window will pop up
	If you do not have X installed on your machine, the GCG session will run in your telnet window

*Some help messages will appear on your screen. If this is the first time this year you have used GCG at the HGMP-RC, and you are not listed in our database as an MRC employee, you will be asked whether you are liable for the GCG charge. Answer these questions as appropriate and eventually you will see a UNIX prompt (something like **unix %)** displayed. From now on we will refer to this as 'the prompt'. All the following examples assume you have started GCG before attempting to run the programs.*

6.5.2 fetch

We are going to extract a sequence from a database by using its EMBL accession number (x07732) and its database name (hshepsh). Either could be found from the literature or by doing a keyword search.

	unix % fetch em:x07732
	Fetch copies GCG sequences or data files from the GCG database into your directory or displays them on your terminal screen.

The program then tells you the filename of the sequence you have extracted:

	x07732.em_hum1

*Look at the output from **fetch**:*

	unix % more x07732.em_hum1

The position of the hepsin coding sequence is indicated by 'CDS' in the file.

6.5.3 translate

We can now translate the region corresponding to the hepsin product listed in the feature table of the file produced by ***fetch.****.*

	unix % *translate*
	Translate translates nucleotide sequences into peptide sequences.
	TRANSLATE from what sequence(s) ? *x07732.em_hum1*
	Begin (* 1 *) ? *826*
	End (* 2363 *) ? *2079*
	Reverse (* No *) ?
	Range begins ATGGC and ends TCTGA. Is this correct ***(* Yes *) ?***
	That is done, now would you like to:
	A) Add another exon from this sequence
	B) Add another exon from a new sequence
	C) Translate and then add more genes from this sequence
	D) Translate and then add more genes from a new sequence
	W) Translate assembly and write everything into a file
	Please choose one (* W *):
	What should I call the output file (* x07732.pep *) ?

Now view the output file:

	unix % *more x07732.pep*

6.5.4 seqed

seqed *is GCG's program for editing or creating new sequences. It has two main modes:*

- ***Editing*** *where you actually edit the sequence.*
- ***Command*** *where you carry out various functions on the sequence.*

	At the prompt, type seqed new.seq, where new.seq is your existing sequence or a sequence file you wish to create. When you are creating a new sequence ***seqed*** *first puts you in the header section. Type in any annotation required and then type <CTRL-D> to enter sequence editing mode.*
	Type in your sequence from the keyboard.
	To go between editing and command mode, use <CTRL-D>. The cursor will jump down to the command prompt (a colon). To go back to editing the sequence, simply press <RETURN>.
	To see what commands are available and how to use them, in the command mode type: help
	When you have finished editing and wish to save the sequence and exit type the command: exit

6.5.5 map

Here we are going to create a restriction map for our sequence with enzymes that cut only once or twice. To do this we need to use command line switches.

	***unix** % map -mincut=1 -maxcut=2*
	MAP maps a DNA sequence and displays both strands of the mapped sequence with restriction enzyme cut points above the sequence and protein translations below. Map can also create a peptide map of an amino acid sequence.
	***(Linear) MAP of what sequence ?** hshepsh.em_hum2*
	Begin (* 1 *) ?
	End (* 2363 *) ?
	Enzyme(* * *):
	What protein translations do you want:
	a) frame 1 b) frame 2 c) frame 3
	d) frame 4 e) frame 5 f) frame 6
	t)hree forward frames s)ix frames o)pen frames only
	n)o protein translation q)uit
	Please select (capitalize for 3-letter) (* t *):
	What should I call the output file (* hshepsh.map *) ?
	Mapping .....
	Writing ....
	MAP complete with:
	Sequence Length: 2,363
	Enzymes Chosen: 229
	Cutsites found: 75
	CPU time: 00.23
	Output file: hshepsh.map

You can now view the text output using the UNIX more command. You will see the restriction enzymes that would cut the sequence only once or twice and their cutting sites. Lists of enzymes that cut or fail to cut, given our criteria, are given at the end of the file.

6.5.6 lookup

***lookup** is an old implementation of SRS. You may find the 'real' SRS service at http://srs.hgmp.mrc.ac.uk/ more useful. **lookup** presents you first with a menu of databases, and then with a set of fields in which you can search. Use the arrow keys to move to the field and enter the required search terms, and then press <CTRL-D> to start the search.*

	unix *% lookup*
	LookUp identifies sequences by name, accession number, author, organism, keyword, title, reference, feature, definition, length, or date. The output is a list of sequences.
	LOOKUP in what sequence libraries:
	a) embl
	b) est
	c) em_new
	d) swissprot
	e) swissprot_new
	f) trembl
	g) trembl_new
	h) sptr
	i) pir
	j) nrl3d
	k) owl
	l) All libraries
	q) quit
	Please choose one or more (* l *): *(enter your choice here)*

You will now be presented with a list of options to refine your search. Use the <TAB> key to move between fields and enter the terms you require. When you are finished, press <CTRL-D> to perform the search. We will enter " mRNA" in the Definition field and " Carassius auratus" in the organism field.

	Complete the query form below:			
	All text:			
	Definition:	*mRNA*		
	Author:			
	Keyword:			
	Sequence name:			
	Accession number:			
	Organism:	*Carassiusauratus*		
	Reference:			
	Title:			
	Feature:			
	On or after (dd-mmm-yyyy):		***On or before (dd-mmm-yyyy):***	
	Shortest sequence length:		***Longest sequence length:***	
	Inter-field operator:	***AND***	***Form of output list:***	***Whole Entries***
	Press <Ctrl>D to continue. *^D*			
	Searching embl			
	131 entries were found.			
	Do you wish to:			
	1) write out this list to a file			
	2) preview the results			
	3) refine the query			
	4) choose different libraries			
	q) quit			
	Please choose one (* 1 *):			
	What should I call the output file (* lookup.list *) ?			
	...			
	131 entries were written to "lookup.list"			

The list file 'lookup.list' could then be used to provide a list of sequences for any program that can search databases.

6.5.7 findpatterns

findpatterns is useful for searching for ambiguous patterns in sequences or for searching for short sequences in databases. BLAST and FASTA have problems searching for sequences of under about 30 bases long and findpatterns should be used instead in this case. If you want to use a list file with this program you must put an '@' symbol in front of the filename, otherwise it would assume that you are giving it a file with a sequence in it and it would get confused.

	unix % findpatterns
	FINDPATTERNS identifies sequences with short pattern queries like GAATTC or YRYRYRYR. You can define the patterns ambiguously and allow mismatches. You can provide the patterns in a file or simply type them in from the terminal.
	FINDPATTERNS in what sequence(s) ? @lookup.list
	Enter patterns individually, one per line.
	End the list with a blank line.
	Pattern 1: GAATTC
	Pattern 2:
	What should I call the output file **(* findpatterns.find*)**

You can view the results using the UNIX more command.

Any Comments, Questions? Support@hgmp.mrc.ac.uk

A , B , C , D , E , F , G , H , I , J , L , M , N , O , P , R , S , T , U , V , W , X , Y

A

B

C

D

E

F

G

H

I

J

L

M

N

O

P

R

S

T

U

V

W

X

Y