Molecular biology and the Internet

J Rashbass

Introduction
There are few subjects that have produced such a sustained response in the press over the past few years as the Internet and the Information Super Highway. Every week we are bombarded with details of new facilities that can be found on this vast computer network. In the United States an initiative led by Vice-President Al Gore three or four years ago established the infrastructure necessary for super-computer communications countrywide. As a result, the American public and the medical profession are fully aware of the strengths of this new technology. In Britain, however, we have been slow to exploit the potential of this worldwide communications network. Even though the Internet has now entered the political debate, lack of a suitable national computer network and a centrally co-ordinated policy has meant that we are rapidly falling behind in the range of resources available to us at our desks. Despite the amount written about the Web in the medical literature in Britain, the number of medical users remains small. In these few columns I should like to provide a glimpse of the variety of resources that are available to the molecular and diagnostic histopathologist, the limitations of the Internet, and what this invaluable facility promises in the future.

What is the Internet?
The Internet is a worldwide collection of computers that share a common and fairly simple communication protocol called TCP/IP (Transmission Control Protocol/Internet Protocol). There is no restriction on the types of wires or links between these computers and they range from the radio waves that link satellites to the simple telephone wires that connect our homes. Each machine on the Internet, and now there are about 31 million connected users, has its own unique identity called an IP (Internet Protocol) number. Information leaving that machine, or going to it, is identified electronically by the presence of this number. Once assigned an IP number, which might be from a commercial vendor (an Internet provider) or through the computer services administered by larger institutions, the user has access to all the facilities on the Internet. This may simply be electronic mail, where messages and documents can be sent from one computer user to another, or bulletin boards where users with a common interest can post and read articles. Users also have access to the vast array of more sophisticated facilities on the World Wide Web.

The World Wide Web has come to dominate the use of the Internet. It is its simplicity and the fact that it incorporates text, images, video, and sound that makes it so attractive. Fundamental to the Web is the concept of "hyper-text". This allows the text within one document to be linked with another document, picture or video which can be on any computer anywhere on the Web. The key to this is a hidden link or Uniform Resource Locator (URL) buried within the text. This means of "publishing documents" involves converting them into a modified format called HyperText Markup Language (HTML); there are many tools now available to allow a word processor document to be converted into this format within a matter of seconds. Information on the Web is provided from computers called Web Servers or Web Sites and accessed by a program called a browser, run on the user's local computer. The majority of resources I will discuss are on the Web and the URLs are given in the appendix.

Molecular biology databases
The Internet has been an essential tool for molecular biologists for many years. It provides the communication network to search for DNA and protein sequence data at sites such as the Sanger Centre in Britain\(^1\), GenBank\(^2\) in the USA or EMBL\(^3\) in Heidelberg. Sequences can be compared with those already known in the database and homologies rapidly identified (fig 1). At present, the GenBank database contains nearly 500 000 DNA sequences, totalling more than 350 million nucleotide bases from over 11 000 biological species. New human DNA sequences are being added to the database at a rate of 1500 a day.\(^4\) At this rate of increase, sending the information by post on digital tapes or CD-ROMs is clearly an inefficient means of distribution. Using the Web gives access to the information from any site in the world within minutes. Programs such as Entrez\(^5\) allow the user to search for sequences in GenBank and, at the same time, find the Medline records from the National Library of Medicine with references and abstracts that refer to that particular sequence. Entrez will also allow the user to visualise the three dimensional structure of the proteins that interest him/her. Searches for sequence similarity can be made with programs such as BLAST\(^6\) (Basic Local Alignment Search Tool), in which the test sequence is sent to a computer at the National Centre for Biotechnology Information in Bethesda (Maryland, USA).\(^7\) The alignment is performed within a matter of minutes and the answer is e-mailed back to your computer.

Features such as these highlight the power of the Internet. The ability to perform complex computational calculations is no longer limited by power of the user's computer. The request and the necessary data are transferred to a centrally maintained computer anywhere on the Web, where the power is adequate to perform
Molecular biology and the Internet

allows form Science Fisheries Pathogenesis Utilization Research Strom.

Figure 1 A schematic diagram of the molecular biology databases available on the Web from the Sanger Centre in the UK (http://www.sanger.ac.uk/short). This image is an “active map”—clicking on an icon in the picture takes the user straight to that database, where electronic forms then allow him/her to enter questions into the computer to perform searches and retrieve data. Reproduced with kind permission of Thure Etzold.

the tasks. The answer, be it a picture, text or a video, is simply e-mailed back to your computer.

Biomedical databases

The Internet is the largest “single” collection of documents ever accumulated. All the information is accessible from the user’s computer and constitutes a vast electronic library which is relatively easily searched, rapidly updated and allows a single resource to be accessed by many people at the same time. One of the most elegant medical facilities available, which has evolved with the development of the Internet and the World Wide Web, is Victor McKusick’s online version of Mendelian Inheritance in Man (OMIM)⁰, Mendelian Inheritance in Man was first published as a two volume work in the early 1960s and has been constantly updated since then. The first electronic version appeared in 1985 and could be searched using tools such as the Internet Gopher, a simple but very powerful text retrieval tool. With the introduction of the World Wide Web and the graphical browsers which support pictures and sound, OMIM has been updated to include video and audio clips. It is now possible to hear the cry of a baby with “Cri du chat” syndrome or see the choreoform movements of a patient with Huntington’s disease. The OMIM database gives extensive information on the classification of genetic diseases, clinical presentations and genomic mapping of the disease locus. The database is updated continuously and provides probably the best “first port of call” when looking for information on hereditary conditions.

There are many other biomedical resources on the Web and many have a molecular or pathological bias. There are on-line handbooks¹ of laboratory protocols with clear instructions on areas from cloning to running

Figure 2 There are several Web sites which provide electronic access to laboratory protocols for molecular biology. This one is run by Mark Strom at the Microbial Pathogenesis/Utilization Research Division, Northwest Fisheries Science Center (NMFS/NOAA) in the USA (http://research.nwfs.noaa.gov/protocols.html). There is a range of methods from full techniques to helpful tips, all instantly available on the user’s Web browser. An electronic form allows users to submit additional comments or techniques. Reproduced with kind permission of Mark Strom.
electrophoretic gels (fig 2). There are histopa-
thology atlases with macroscopic and micro-
scopic pathology, some with accompanying
clinical information and x ray films. Many
medical schools now make their course mate-
rnal available on the Web, providing informa-
tion not only for their own students but also for
those at other medical schools. In the USA
physicians can gain CME points by completing
electronic cases over the network. 4

Electronic journals
Our main source of new information comes
from reading current journals, but it can be
frustrating if the journal required is not in the
library or if someone else is reading it. An obvi-
ous solution is to provide the journal in an
electronic form over the Internet. Several
dozen science journals, including Nature, Cell
and Science, make their contents pages and
abstracts available on the Web on the day of
publication. This allows multiple users from
anywhere in the world access to the same issue
at the same time. The text can easily be
searched for key words and hypertext links can
be used to allow the reader to jump between
other articles published in other issues of the
same journal. The field is evolving rapidly and
standards are still being developed, but it is
only a matter of time before clicking on a text
citation will bring up the full Medline refer-
ce of the article.

BioMOO: an electronic meeting place for
biologists
The majority of Web resources provide infor-
mation for individual users. There is little
scope for immediate interaction with other
Web users. One exception is a MOO (Multi-
user Dimension Object Orientated), a virtual
reality meeting place at which multiple users
can be present at once. BioMOO is one such
facility for biological scientists, in which users
from around the world can log on to a computer
at the Weizmann Institute in Israel and
converse with each other, hold symposia and
read notice boards, posters and job adver-
tisements. 5 The structure of the MOO illus-
trates the beauty of the Web because the images
and slide collections found in BioMOO are
stored on a separate computer at CalTech in
California and the two sites are linked
seamlessly by hypertext.

Too much of a good thing?
Initial enthusiasm about the Web must be tem-
pered by reality. There are times when the network
is so slow that it is unusable and while the number
of resources is remarkable, there is in general no
peer review system for determining the quality or
the reliability of the information. There are
reputable institutions which make information
available, but there is also much more information
produced from dubious sources on the "fringe".
Sorting between this information can be difficult,
time consuming and costly, especially if you pay by
the minute for your connection. At present, most
of the resources on the Web are free and without
copyright, provided by non-commercial organisa-
tions and individuals. Many who use the Web are
funded by research grants and use the Internet as
a means of disseminating their knowledge. The
service is unlikely to remain free. After all, if a pub-
lisher makes their journal available free on the Web
why should subscribers pay for the paper version
to arrive (several days later) on their desk?

One of the biggest fears expressed by the medi-
cal profession is that of patient confidentiality.
How can you be certain that if the computers in
your hospital allow you access to information from
the outside world that others from the outside
world cannot access hospital information? This is
a very real concern and one that is being
addressed. There are several security measures
available—for example, a 'firewall', in which a
combination of software and hardware is used to
restrict traffic coming into a site while allowing
free traffic out of the site. These systems work
well—it must be remembered that organisations
whose job it is to be paranoid, such as the
intelligence gathering services, make use of them.
The security of medical records even within a hos-
pital can be difficult to achieve; one strength of
computer security is that you can at least maintain
a check on who has been looking at what and
when.

The future
The power of computers, the speed at which
they run and their storage capacity doubles
every 18 months, while in the same period
their price halves. With this rapid advance in
the technology every few weeks, the variety
of resources available on the Internet is changing.
Encryption technology is being developed
which, in the next few months, will allow the
secure use of credit card numbers on the Web.
Suddenly, a whole world of commerce will
appear on the Internet. The same technology
could permit medical information to be sent
securely over the Web and improvements in
communication technology in the next few
years will allow real-time video information to
be transmitted. We will all be attending virtual
conferences and seminars within the next five
years.

Conclusions and suggestions
In Britain we are being bombarded by
information about the resources available on
the Web; however, the infrastructure necessary
is unavailable to many in the medical profes-
sion. Facilities which would be of considerable
benefit to all pathologists are restricted to a few
isolated enthusiasts. Those of us who are
fortunate enough to be able to use the Internet
easily, support the idea of an official body to
co-ordinate and represent the interests of all
pathologists. Communication and information
are fundamental to all branches of medical
practice and it is clear that pathologists
throughout Britain should not be denied
access to these facilities.

## Appendix

<table>
<thead>
<tr>
<th></th>
<th>Sanger Centre</th>
<th><a href="http://www.sanger.ac.uk/searching.html">http://www.sanger.ac.uk/searching.html</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>(b)</td>
<td>EMBL</td>
<td><a href="http://www.embl-heidelberg.de/srs/srsc">http://www.embl-heidelberg.de/srs/srsc</a></td>
</tr>
<tr>
<td>(d)</td>
<td>BLAST</td>
<td><a href="http://www.ncbi.nlm.nih.gov/BLAST/">http://www.ncbi.nlm.nih.gov/BLAST/</a></td>
</tr>
<tr>
<td>(f)</td>
<td>Protocols</td>
<td><a href="http://research.nwpsc.noaa.gov/protocols.html">http://research.nwpsc.noaa.gov/protocols.html</a></td>
</tr>
<tr>
<td>(g)</td>
<td>Teaching</td>
<td><a href="http://www.primenet.com/~gwa/med.ed/courses.html">http://www.primenet.com/~gwa/med.ed/courses.html</a></td>
</tr>
<tr>
<td>(h)</td>
<td>CME units</td>
<td><a href="http://ns.gamewood.net/pathcase/nephpath.html">http://ns.gamewood.net/pathcase/nephpath.html</a></td>
</tr>
<tr>
<td>(i)</td>
<td>Nature</td>
<td><a href="http://www.nature.com">http://www.nature.com</a></td>
</tr>
<tr>
<td>(j)</td>
<td>Cell</td>
<td><a href="http://www.cell.com">http://www.cell.com</a></td>
</tr>
<tr>
<td>(l)</td>
<td>BioMOO</td>
<td><a href="http://bioinfo.weizmann.ac.il:8888">http://bioinfo.weizmann.ac.il:8888</a></td>
</tr>
</tbody>
</table>