
CHAPTER 2

Internet Resources for the Geneticist

MICHAEL R. BARNES¹ and CHRISTOPHER SOUTHAN²

¹*GlaxoSmithKline Pharmaceuticals*
Harlow, Essex, UK

²*Oxford GlycoSciences UK Ltd*
The Forum, 86 Milton Science Park
Abingdon OX14 4RY, UK

- 2.1 Introduction
 - 2.1.1 Hypothesis construction and data mining — essentials for genetics
- 2.2 Sub-division of biological data on the internet
- 2.3 Searching the internet for genetic information
- 2.4 Which web search engine?
 - 2.4.1 Google
 - 2.4.2 Scirus
- 2.5 Search syntax: the mathematics of search engine use
 - 2.5.1 Using the ‘+ and –’ symbols to filter results
 - 2.5.2 Using quotation marks to find specific phrases
 - 2.5.3 Restricting the searching domain of a query
- 2.6 Boolean searching
- 2.7 Searching scientific literature — getting to ‘state of the art’
 - 2.7.1 PubMed
- 2.8 Searching full-text journals
 - 2.8.1 HighWire
 - 2.8.2 Literature digests and locus-specific databases
- 2.9 Searching the heart of the biological internet — sequences and genomic data
- 2.10 Nucleotide and protein sequence databases
 - 2.10.1 Entrez
 - 2.10.2 Sequence Retrieval Server (SRS)
- 2.11 Biological sequence databases — primary and secondary
 - 2.11.1 Primary databases
 - 2.11.2 Secondary databases — nucleic acids and proteins
 - 2.11.3 Nucleic acid secondary databases
 - 2.11.4 STSs and SNPs
 - 2.11.5 Protein databases and websites

2.1 INTRODUCTION

The World Wide Web ('the web') and our knowledge of human genetics and genomics are both expanding rapidly. By allowing swift, universal and largely free access to data, particularly the human genome sequence, the web has already played an important role in the study of human genetics and genomics. Increased data accessibility is dramatically changing the way the scientific community is communicating and carrying out research. The internet biology community is expanding daily with an organic development of websites, tools and databases, which could eventually replace the conventional scientific paper as the predominant form of communication. Already we are starting to see successful website/journal hybrids such as *Genome Biology* (<http://genomebiology.com/>) and biomednet (biomednet.com) which offer high quality peer-reviewed scientific articles and reviews alongside bioinformatics databases and tools. Many more established journals like *Nature* and *Science* are rapidly following suit with user-friendly websites, which offer much more than the full text of the journal.

The web is offering more than just information. Virtual research communities have been organized around databases and specialist research groups. These communities are even influencing the way bioinformatics tools are being developed, a good example being Ensembl the human genome browser developed at the EBI and Sanger Institute in Hinxton, Cambridgeshire (Hubbard *et al.*, 2002). In the spirit of open source community projects such as the free UNIX operating system Linux, the Ensembl development team has developed Ensembl on an 'open source' basis. This means all code is freely available to anyone who wishes to download it. But further still, Ensembl is developed by a 'virtual community' of developers from institutes, industry and academia around the world who are free to modify and add to the central software code (subject to a peer review). So the tools and interfaces, though primarily developed in Hinxton, may include contributions from developers in Singapore, North Carolina and New York or elsewhere.

2.1.1 Hypothesis Construction and Data Mining – essentials for Genetics

Genetics is a science which calls for analysis and interpretation across a wide range of biological research. Many chapters in this book deal with focused tools. Beyond these specialist applications however, geneticists need access to a wide range of databases and literature, both to update particular research areas and formulate new hypotheses. This requires expertise across the gamut of biological data on the internet. This ranges from the review literature to highly specific databases. This can illuminate biology from gene function to biological pathways. Effective data mining needs an understanding of the general principles by which it is organized, particularly the sequence-based data resources. This needs to be backed up by good scientific judgment concerning quality and significance.

An exhaustive description of biological data and databases on the internet would be beyond the scope of this book. Confucius might not have been thinking of internet searching when he said 'give a man a fish and he will live for a day, teach a man to fish and he will live forever', but the principle still applies. So, instead of reviewing the data

resources themselves the most useful thing we can do here is to review search methods to help identify both current and future resources.

2.2 SUB-DIVISION OF BIOLOGICAL DATA ON THE INTERNET

Biological information on the internet can be roughly subdivided into two broad categories, which we will term 'the biological internet' and 'biological information on the internet'. This distinction may not be immediately apparent—we define 'the biological internet' as purpose-built biological tools and databases which index and contain detailed biological information, such as the human genome sequence, nucleotide and protein sequences, genetic markers, polymorphisms and the full range of biological literature. The majority of these tools and databases are maintained in a highly integrated form by major biological organizations such as NCBI and EMBL. We define 'biological information on the internet' as biological data which is less formally maintained on the web, this could include information on research laboratory homepages, conference abstracts, tools, boutique databases and any other data that scientists have seen fit to present on the web.

These distinctions are more clearly defined by the tools that are available to search the data. Firstly there are general purpose web search engines, such as Google, Lycos and Excite (see Table 2.1 for a full list), these tools index and search the full range of the internet and have the capability to identify webpages, tools and databases by simple keyword searching. A second category of tools are the specialist biological search tools, such as Entrez-PubMed and BLAST (see Chapter 4). The former uses keyword searching or accession number queries, the latter uses similarity searching to find related sequences.

The choice of search tool depends on the kind of information that needs to be retrieved. The scope of biological and genetic information on the internet is so broad that no single tool is available to index all data. The key point to understand is which tool is most suitable to identify a specific form of data. For example literature is most comprehensively indexed by PubMed or Scirus (see below), whereas nucleotide records can only be identified with any specificity by Entrez or BLAST. This is in contrast to a laboratory homepage or a boutique web resource. Unless a description is published in PubMed these resources may only be identified by a web search tool. If it is not clear what information needs to be retrieved then clearly both specific and general search tools should be used.

TABLE 2.1 Key Internet Search Engines with Reported Index Size (Equivalent to the Number of Documents Indexed)

Search Engine	URL	Reported Index Size
Google	http://www.google.com/	560 M
AltaVista	http://www.altavista.com/	350 M
FAST	http://www.alltheweb.com/	340 M
Northern Light	http://www.northernlight.com/	265 M
Excite	http://www.excite.com/	250 M
HotBot	http://www.hotbot.com/	110 M
Lycos	http://www.lycos.com/	110 M
MetaCrawler	http://www.metacrawler.com/	ND
Scirus	http://www.scirus.com/	69 M (science only)

2.3 SEARCHING THE INTERNET FOR GENETIC INFORMATION

The World Wide Web began as an information-sharing and retrieval project at the European particle physics laboratory CERN (Berners-Lee *et al.*, 1999). It has only recently evolved into the mass media beast that we all know. But just as the internet began, so it continues as an information-sharing resource for scientists in all fields. One cannot deny that commercial proliferation has not been an unmitigated success for the growth of the web but this has led many scientists to perceive the internet as a rising tide of irrelevant noise that has largely washed away any intrinsic value. This is a misconception. We will demonstrate that some web resources for biological sciences are both outstanding and indispensable. Internet biology suffers as much as any other field of scholarship from: data of dubious provenance, broken links, outdated sites and newsgroup spam. But it also contains valuable and novel data which can be crucial for scientific discovery. The skill is to recognize chaff and know how to sift the wheat from it. To do this we need tools that are capable of highlighting relevant information in an organized manner.

In the process of linking genotypes to phenotypes it is important to know about the function of a gene or gene family, for example to prioritize candidate disease-association genes. In such cases biological search tools and internet search tools may provide complementary results. To give an hypothetical example let us assume that a genetic locus associated with a familial form of basal cell carcinoma includes a novel gene with homology to WNT genes. With no knowledge of WNT genes it would be difficult to include or exclude this gene as a candidate. A search of PubMed would reveal a daunting range of over 1000 publications mentioning members of the WNT gene family. Some might contain specific information to link WNT genes to carcinoma but it would take a long time to read and digest all the available information. Using Google to search for 'WNT gene' would identify a range of conference abstracts and laboratory homepages. Towards the top of the hit-list this would include the 'World Wide WNT Window' (www.stanford.edu/~rnusse/wntwindow.html). This is an excellent summary of the whole WNT signalling pathway maintained by prominent researchers in the WNT signalling field. The page includes a detailed and regularly maintained summary of all genes in this highly complex pathway, which is currently unpublished. Examination of this pathway would identify the Patched receptor upstream, which has been shown to cause 80% of sporadic basal cell carcinomas. This is just one of many examples of how a thriving unpublished and unpublicized on-line research community can be identified by opportunistic internet searching.

2.4 WHICH WEB SEARCH ENGINE?

In a nutshell the availability of full-text search engines allows the web to be used as a searchable 15-billion-word encyclopedia. However, because the web is a distributed, dynamic, and rapidly growing information resource, it presents many difficulties for traditional information retrieval technologies. This why the choice of the search methodology used for searching can lead to very different results.

An important point to make is that all search engines are not the same. A common misconception is that most internet search engines index the same documents for a large proportion of the web. In fact the coverage of search engines may vary by an order of magnitude. An estimated lower boundary on the size of the indexable web is 0.8 billion pages

(<http://www.neci.nec.com/~lawrence/websize.html>). Many engines index only a fraction of the total number of documents on the web and so the coverage of any one engine may be significantly limited. Combining the results of multiple engines has been shown to significantly increase coverage. This is done automatically with metasearch engines such as MetaCrawler (www.metacrawler.com), which search and combine the results of several search engines. Table 2.1 presents a selection of web search engines with direct applicability to biological searching. We also recommend the website, SearchEngineWatch.com, for reviews and reports on new search engines.

2.4.1 Google

It is apparent from Table 2.1 that Google offers the widest indexing capacity. This is an innovative search engine based on scientific literature citation indexes (Butler, 2000). Conventional search engines use algorithms and simple rules to rank pages based on the frequency of the keywords specified in a query. Google exploits the links between webpages to rank hits. Thus the highly cited pages of the web world with many links pointing to them are ranked highest in the results. This is an efficient searching mechanism which effectively captures the internet community 'word of mouth' on the best and most frequently used webpages.

2.4.2 Scirus

The greatest limitation for web search engines is unindexed databases. These include many of the databases that make up the biological internet, such as sequence databases and some subscription-based resources such as full-text journals, and commercial databases. Although limited material from these sites, such as front pages, documentation and abstracts are indexed by search engines, the underlying data is not available because of database firewalls and/or blocks on external indexing.

In an attempt to solve this problem, the publisher Elsevier has developed Scirus (<http://www.scirus.com/>). This is a joint venture with FAST, a Norwegian search engine company who have produced an excellent specialist scientific search engine. Scirus enhances its specificity and scope by only indexing resources with scientific content. These include webpages, full-text journals and Medline abstracts. This makes Scirus an effective tool for both web and literature searching tool. Both full text and PDF format journal content is indexed by performing a MetaSearch of the other major providers of full text—Elsevier's ScienceDirect and Academic Press's IDEAL collection. Scirus also searches the web for the same key words, including Medline, patents from the databases of the US Patent Office, science-related conferences and abstracts. The Medline database is provided on the BioMedNet platform, which requires a free BioMedNet login and password for access. Scirus offers the user several options to customize their searches to search only free sites, only membership sites or only specific sites. The advanced interface also allows boolean queries (see below). By March 2002 Scirus had indexed 69 million science-related pages, including PDF files and peer-reviewed articles, thereby covering the majority of the biologically relevant internet.

Although Scirus expands the scope of biological data searching beyond other search engines it falls short in some areas. For example the full-text journals are restricted to Elsevier and Academic Press. Coverage is also restricted by index pre-filtering that might miss some websites. Another disadvantage is that search results tend to be redundant. Although for literature searching there are alternative full text searching tools such as

HighWire (see below) Scirus is tantalizingly close to what a universal biological search engine should be.

2.5 SEARCH SYNTAX: THE MATHEMATICS OF SEARCH ENGINE USE

The best search engine in the world will not retrieve relevant results unless the query is correctly defined. This is easy to master and a few basic commands can turn a poor specificity keyword search into a highly targeted query. The key to successful sifting of the web is to select for the minimum number of irrelevant hits (maximize specificity) but avoiding the exclusion of relevant hits (minimize false negatives).

2.5.1 Using the ‘+ and –’ Symbols to Filter Results

Sometimes it is necessary to ensure that a search engine finds pages that have all the words you enter, not just some of them. This can be achieved by using the ‘+’ symbol. Similarly you may wish to exclude a specific word from your search by using the ‘–’ symbol. These commands work with nearly all the major web search engines and are similar in function to the boolean operators ‘AND’ and ‘NOT’ respectively.

As an example let’s say you wish to find information about human promoter prediction tools. You could search using [+ promoter + prediction + tool]. This will only retrieve pages that contain all three words. If the search returns excessive information by including tools for plant and bacterial promoter prediction, one could further refine the search by using the following search query [+ promoter + prediction + tool – plant – prokaryote]. This will subtract pages which mention plants and prokaryotes. Be aware though that this might filter out valid hits to tools which analyse *both* prokaryote and eukaryote sequences.

2.5.2 Using Quotation Marks to Find Specific Phrases

The most complex filtering syntax on our promoter prediction query still manages to retrieve over 1000 results, so we need to consider other methods of reducing the number of hits. One approach is to use a phrase search that will find only those pages where the terms appear in exactly the order specified. This is achieved by putting quotation marks around the phrase, so we might search with [‘promoter prediction tool’]. This retrieves six relevant hits but clearly many sources have been filtered out, so it is important to beware of over-specifying search terms.

2.5.3 Restricting the Searching Domain of a Query

A final measure that can be taken to fine tune your query is to restrict the internet domain. For example you can restrict your search to only identify hits in the .edu (educational) domain or to ignore hits from the .com (company) domain. This is achieved in Google and most other sites by using the [+ site:.edu] to include a domain or [– site:.com] to exclude a domain. This command can be extended further to search only a specific site, e.g. to search the NCBI website for SNP information try [+ SNP + site:ncbi.nlm.nih.gov].

Table 2.2 includes the search results obtained from the different variations on the search for promoter prediction tools, using both Google and Scirus. This shows the improvements

TABLE 2.2 Different Results Obtained from Different Query Targeting Methods. Results Compare the Number of Hits Returned by the General Search Engine Google and Specialist Science Search Engine Scirus

Query	Google Hits	Scirus Hits*
+ promoter + prediction + tool	4050	2379
'promoter prediction tool'	6	2
'promoter prediction tools'	14	8
+ promoter + prediction + tool – plant	2630	1312
+ promoter + prediction + tool – plant – bacterial	2080	936
+ promoter + prediction + tool – plant – bacterial – site:.com	1750	NA

*Queries to Scirus were designed using the equivalent boolean syntax in the advanced search form.

from filtering on the query. The final word on fine tuning web search queries is to be as flexible as possible. Try to use keywords which are likely to be specific to the kind of website or tool you are looking for. Sometimes it is useful to go to a page or tool similar to the one you are looking at to check for very specific words that might be shared by similar sites. For example, in the case of promoter prediction tools, a commonly occurring word was 'server'; exchanging this for 'tool' significantly improves the relevance of the hits.

2.6 BOOLEAN SEARCHING

Although the familiar boolean search commands (AND, OR, NOT) are widely used for many forms of database searching, including PubMed, they are not universally supported by all web search engines. Table 2.3 lists those supported by the most popular search engines. The functionality offered by AND and NOT mirrors the functionality of [+ and –]. Other commands have a distinct function, for example [SNP OR Analysis] will retrieve all webpages that contain the words SNP or analysis. The NEAR command is not

TABLE 2.3 Boolean Commands Supported by Popular Web Search Engines

Command	How	Supported by
Or	OR	AltaVista, Excite, Google, Lycos, Northern Light
	None	FAST, LookSmart,
And	AND	AltaVista, Excite, Lycos, Northern Light
	None	FAST, Google, LookSmart
	NOT	Excite, Lycos, Northern Light
Not	AND NOT	AltaVista
	None	FAST, Google, LookSmart,
Near	NEAR	AltaVista (10 words), Lycos (25 words)
	None	FAST, Google, LookSmart

widely supported but can be useful to help to identify two keywords in close proximity to each other.

2.7 SEARCHING SCIENTIFIC LITERATURE – GETTING TO ‘STATE OF THE ART’

Effective mining of the literature is important at the stages of conception, design and construction of genetic studies. At the most basic level it is important to be aware of the ‘state of the art’ in a research area before embarking on new efforts. At the very least this avoids duplication of effort, but it can also provide previously unrecognized clues which need to be followed up. Unfortunately this important informatic process is still lacking truly innovative tools and databases. We are still struggling with tools that cover the fundamentals of literature searching, such as making the full text of *all* journals available for searching. Even with unlimited access to full text, the problems with effective literature mining are profound. Some of these problems stem from the limitation of language as a precise query tool—there is simply too much vocabulary to describe or specify the same target information. Some databases attempt to minimize the impact of this problem by the use of controlled vocabulary and gene nomenclature. But in the absence of such measures, flexible composition of queries becomes quite critical to obtain comprehensive coverage of a research area.

There are many commercial and publicly available tools and databases for mining scientific literature which vary in their data content. Some offer access to proprietary curated databases but they all employ essentially similar keyword-based interfaces with a facility for boolean operators to combine and subtract keywords.

2.7.1 PubMed

PubMed is the most widely used free literature searching tool for biologists. It forms part of the Entrez-integrated database retrieval system at the NCBI and is essentially a web interface to the Medline database which indexes >11 million journal abstracts. It also provides links to the full text of more than 1100 journals available on the web, although search queries are restricted to the text in abstracts. The interface allows the user to specify a search term (any alpha numeric string) and a search field (e.g. title, text word, journal or author). Queries retrieve abstracts from most of the major journals, although not all journals are indexed, particularly newer journals or journals with lower impact factors. There is a surprisingly stringent threshold applied before a journal will be considered for Medline indexing.

Many of the same guiding principles applied to searching the web also apply to PubMed, but there are some differences between this tool and other more general web search engines. Firstly the boolean operators are limited to the three main operators AND, OR and NOT. One major improvement over most web search engines is the availability of a wildcard function (*) to designate any character or combination of characters. The creative use of wildcards and boolean terms is important to widen the search without retrieving excessive and irrelevant results. For example, to find publications which present evidence of schizophrenia association on chromosome 8q21, an appropriate PubMed query might be [schizo* AND 8q*] searching the *text word* field. Using a wildcard search with ‘schizo*’ instead of ‘schizophrenia’ retrieves articles which mention schizoaffective, schizophrenia or schizophrenic, all of which may be relevant. By using a wildcard with

'8q' the search will retrieve nearby loci or larger loci which may encompass 8q21, e.g. 8q13–8q22. These are simple points but they are integral to a successful search strategy. Those using these facilities extensively will find additional searching guidelines on the NCBI website.

2.8 SEARCHING FULL-TEXT JOURNALS

Prospects for literature searching have improved recently with the greater availability of full-text articles. We have already described the advances offered by Scirus in searching full-text journals and the web simultaneously. Other highly recommended websites are HighWire which is approaching comprehensive coverage of available full-text journals and Medline (see below). However, searching scientific publications is still somewhat decentralized and there is still no completely comprehensive central tool to search all full-text journals, although it is possible to search the full text of most of the major genetics journals by visiting the top three or four major publishers. Table 2.4 lists the major sites which index the full text of a large range of science journals. As a benchmarking test we queried each tool, with a standard full-text query for the keyword [WNT], where searching Medline was also an option we identified the combined number of full text and Medline hits in parentheses. The highest number of results was retrieved with Scirus, however these results were very redundant. The HighWire tool seemed most effective in the benchmarking test, identifying a high number of hits with no redundancy.

2.8.1 HighWire

HighWire was set up as a non-profit making organization in 1995 by Stanford University to help universities and societies to publish on the web at low cost (Butler, 2000). Since its launch HighWire has expanded to become the world's second-largest scientific repository, after the US space agency NASA's Astrophysics Data System (which contains no biological information). Many journals available on the HighWire site make their content free immediately, or 1 or 2 years after print publication often coupled with an early view service for papers in press. In March 2002, HighWire had indexed 410,821 free full-text articles, derived from a list of 324 full-text journals. These are listed on the website along with Medline records from January 1948 through to April 2002. In our benchmark test

TABLE 2.4 Major Websites Providing Full-text Journal Access and Searching

Site/Publisher	Test Query Hits (with Medline)	URL
PubMed	(1615)	http://www.ncbi.nlm.nih.gov/entrez
Scirus	5061 (7015)*	http://www.scirus.com/
HighWire	2651 (3738)	http://highwire.stanford.edu/
Biomednet	1192 (2749)	http://www.bmn.com
ScienceDirect (Elsevier)	1264	http://www.sciencedirect.com
IDEAL	565	http://www.idealibrary.com/
Nature Publishing Group	255	http://www.nature.com/nature/
Wiley InterScience	196	http://www.interscience.wiley.com/

*Results from Scirus were redundant.

against other full-text search tools a comparative search of PubMed and HighWire with the keywords [Wnt16 OR Wnt-16] identified two papers with PubMed and eight papers with HighWire.

2.8.2 Literature Digests and Locus-specific Databases

The literature searching process can be simplified by searching locus-specific databases. The most widely used is On-line Mendelian Inheritance in Man (OMIM). As the name suggests, this focuses on Mendelian monogenic disorders, although it also offers some coverage of complex diseases. As a manually curated digest of the literature extracted from the full text of publications it can contain more information than PubMed. Although this has the disadvantage that not all entries are fully comprehensive or current, the database usually captures the most salient information and is therefore a good place to start. In addition OMIM is fully integrated with the NCBI database family. This facilitates rapid and direct linking between disease, gene sequence and chromosomal locus.

Other databases are available which provide curated information about genes and diseases which can also help to speed up the literature searching process. One of these is GeneReviews (www.geneclinics.org). This complements the molecular genetics emphasis of OMIM by offering a distinctly different focus. GeneReviews is a medical genetics information resource aimed at physicians and other healthcare providers. The site provides current, expert-authored, peer-reviewed, full-text articles describing the application of genetic testing to the diagnosis, management and genetic counselling of patients with specific inherited conditions. It also contains an international genetic testing Laboratory Directory and an international genetic and prenatal diagnosis Clinic Directory.

2.9 SEARCHING THE HEART OF THE BIOLOGICAL INTERNET – SEQUENCES AND GENOMIC DATA

So far we have reviewed a range of tools and approaches for searching the wider internet and the specialist scientific literature for biological information which may be useful for genetics. All of the tools reviewed so far may provide links, but will stop short of direct retrieval of actual biological database records, such as DNA or protein sequence records. This biological information is the heart of the biological internet. However, the flood of sequence data from genome sequencing has rapidly pushed biological sequence data beyond the reach of general internet searching tools. Instead sequence data can be searched and retrieved by using specialist bioinformatics tools based on sequence homology, map location, keyword, accession number and other features in the records. At a basic level this can be done by keyword searching using search tools such as, Entrez at the NCBI (Schuler *et al.*, 1996) or SRS at the EBI (Zdobnov *et al.*, 2002). Moving beyond simple searching methods the biological databases are constantly being updated and re-engineered to allow more powerful data query methods. These methods are covered in many other chapters throughout this book.

2.10 NUCLEOTIDE AND PROTEIN SEQUENCE DATABASES

There are three major organizations that collaborate to collect publicly available nucleotide and protein sequences. These organizations share data on a daily basis but they are distinguished by different international catchment areas for submissions, different formats and

sometimes differences in the nature of their submitter annotations. Genbank is maintained by the NCBI in the United States (<http://www.ncbi.nlm.nih.gov/Genbank/index.html>). EMBL is maintained by the European Bioinformatics Institute in the United Kingdom (<http://www.ebi.ac.uk/>). The third member is the DNA Database of Japan (DDBJ) in Mishima, Japan (<http://www.ddbj.nig.ac.jp/>). All three organizations offer a wide range of tools for sequence searching and analysis but two integrated database query tools have become pre-eminent. These are Entrez from the NCBI and SRS from the EBI.

2.10.1 Entrez

Entrez (<http://www.ncbi.nlm.nih.gov/Entrez/>) is the backbone of the NCBI database infrastructure. It is an integrated database retrieval system that allows the user to search and browse all the NCBI databases through a single gateway. Entrez provides access to DNA and protein sequences derived from many sources, including genome maps, population sets and, as already described, the biomedical literature via PubMed and On-line Mendelian Inheritance in Man (OMIM). New search features are being added to Entrez on a regular basis. Most recently facilities have been added to allow searches for DNA by 'ProbeSet' data from gene-expression experiments and for proteins by molecular weight range, by protein domain or by structure in the Molecular Modelling Database of 3D structures (MMDB).

2.10.2 Sequence Retrieval Server (SRS)

The sequence retrieval server (SRS) serves a similar role to Entrez, for the major European sequence databases. SRS is a flexible sequence query tool which allows the user to search a defined set of sequence databases and knowledge-bases by accession number, keyword or sequence similarity. SRS encompasses a very wide range of data, including all the major EMBL sequence divisions (Table 2.5). SRS goes one step further than Entrez by enabling the user to create analysis pipelines by selecting retrieved data for processing by a range of analysis tools, including ClustalW, BLAST and InterProScan.

2.11 BIOLOGICAL SEQUENCE DATABASES – PRIMARY AND SECONDARY

Anyone entering the heart of the biological internet encounters a bewildering number of accession numbers, identifiers and gene names. To get to grips with this flood of terminology it is important to understand the difference between primary and secondary databases and their associated accession numbers. This is not proposed as a rigorous definition but it does have a utility for understanding the information flow between sequence databases.

2.11.1 Primary Databases

Primary accession numbers have a number of key attributes; they refer to nucleic acid sequences derived directly from a sequencing experiment, the results are submitted by authors in a standardized format to GenBank, EMBL or DDBJ, the accession numbers are both unique and stable (if they are updated or amended by the submitting authors the accession number will signify a version change as .1, .2 etc.), the data records from every accession number can be retrieved, a contactable submitter is included in every record,

TABLE 2.5 Databases Indexed by the Sequence Retrieval Server at the EBI

Data Type	Database
Scientific literature	Medline, GO, GOA
Protein sequence libraries	European, Japanese and US protein patents, SWISS-PROT, SpTrEMBL
DNA sequence libraries	EMBL, Ensembl HUMAN, global DNA patents
Protein motifs	INTERPRO, PROSITE, PRINTS, PFAM, PRODOM, NICEDOM
DNA sequence related	UTR, UTRSITE, BLOCKS, TAXONOMY, GENETICCODE, REBASE, EPD, CPGISLAND, ENSEMBLCPG, UNIGENE
Transfac (Transcription factor analysis)	TFSITE, TFFACTOR, TFCELL, TFCLASS, TFMATRIX, TFGENE
Protein3DStruct	PDB, DSSP, HSSP, FSSP, RESID
Mutations	SWISSCHANGE, EMBLCHANGE, OMIM, HUMUT, HUMAN_MITBASE, P53LINK, Locus Specific Mutations (see Chapter 3)
SNPs	HGBASE, HGBASE_SUBMITTER
RH mapping	RHDB, RHEXP, RHMAP, RHPANEL
Metabolic pathways	LENZYME, LCOMPOUND, PATHWAY, ENZYME, EMP, MPW, UPATHWAY, UREACTION, UENZYME, UCOMPOUND
SRS pipelineapplications	FASTA, FASTX, FASTY, NFASTA, BLASTP, BLASTN, CLUSTALW, NCLUSTALW, PPSEARCH, RESTRICTIONMAP, HMMPfam, InterProScan, FingerPRINTScan, PFScan, BlastPRODOM, ScanRegExp

they are explicitly redundant in that all submissions are accepted regardless of partial or complete overlap with existing entries and lastly the growth rate remains close to exponential and now exceeds 16 million sequence records. The concept of authors' needs stretches to encompass consortia that run high-throughput sequencing projects. One of the most valuable and perhaps overlooked principals of these unique public repositories is that there is always (with the exception of patent data, see below) an identified individual or laboratory representative listed with the sequence record who can be contacted for any queries regarding experimental details, data quality and availability of source material. There is a large amount of information associated with primary sequence records. These include primary accession numbers, version numbers, protein ID numbers, gene identifier (GI) numbers, header records and feature identifiers. These cannot be covered in detail here but full descriptions are given in database guides (<http://www.ebi.ac.uk/embl/index.html>) and release notes (<ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>).

Geneticists should be encouraged to contact submitting authors in cases where anything seems non-obvious about primary data records for an mRNA or a finished genomic clone. They may have extra information that has a crucial bearing on the interpretation of genetic experiments. Authors may be difficult to track down if they have moved institutions but they are usually pleased to assist in the utilization of their data, because as with scientific publishing, this is the principle behind public sequence databases. Technical errors,

anomalies, miss-annotation in submissions or artefacts are entirely the responsibility of submitting authors not the database administrators. Although we should be sanguine concerning anomalies in the high-throughput data divisions (EST, GSS, STS, HTG, HTC and SNP) if problems are pointed out authors can certainly amend or update their entries or in some cases may withdraw them. The primary data is deposited in good faith so authors should certainly not be harshly judged if an error has occurred in the rough and tumble of cloning, sequencing and submitter annotation. The exception to author responsibility for GenBank records is the patent division (gbPAT) where inventors are not equivalent to academic authors. These sequence records are processed by the US, European and Japanese patent offices and forwarded on to the databases. Although author contact may not be practical database users should be aware that patent applications are public documents and for an increasing number of gbPAT records the documentation can be accessed via the patent number on-line and free of charge (<http://ec.espacenet.com/espacenet/> and <http://www.uspto.gov/patft/>). It is also possible to get to these patent full-text links directly from sequence entries via SRS.

2.11.2 Secondary Databases — Nucleic Acids and Proteins

By definition secondary databases are derived from the primary data. The word secondary should not be taken to imply lower value; indeed they include sources of the highest utility for genetic research. However they are defined, it is important to understand how they are linked back to the experimental data. The good news for geneticists is that there is now a comprehensive selection of high quality secondary databases that extract and collate subsets of mRNA, genomic or protein sequences from primary GenBank entries. The bad news is that the proliferation of features that make secondary databases so powerful also presents a bewildering range of options to the user. Testimony to both the good and bad news is given by the 2002 update of the Molecular Biology Database Collection (<http://nar.oupjournals.org/cgi/content/full/30/1/1/DC1>). This covers no less than 355 databases, up from 281 in 2001, of which the primary databases, GenBank, EMBL and DDJB, constitute only three entries. Although this compendium includes many non-human data sources almost all of these secondary databases contain information that could be pertinent to mammalian genetics. These review issues appear every January in *Nucleic Acids Research* and are definitely worth browsing. Are the genome portals secondary databases? This is where the definitions become blurred. Because NCBI generate their own genomic contig accessions (NT numbers) and Ensembl generate their own exon and gene identifiers they could be considered secondary databases. In so far as the UCSC genome portal marks up only external sequence record identifiers (primary and secondary) they are not strictly a secondary database. However, because they usefully give every type of gene prediction in the display a retrievable identity number, they could be considered as a secondary database.

The value of secondary databases includes the following:

- Distilling down a massive number of overlapping and/or redundant primary GenBank entries to a manageable range of genomic sections, unique transcripts and translated protein sequences
- Maintaining a running total of gene products, they partition human gene products and other vertebrates with extensive genomic data such as mouse, rat and zebra fish
- The inclusion of informative graphic displays for sequence features
- Providing access to a vast amount of pre-processed bioinformatic data

- Extensive interconnectivity through web hot-links
- Many of them are backed up by extensive institutional resources and expertise

However, users of these secondary databases also need to be aware of their shortcomings:

- They all suffer from the snapshot problem i.e. the time to re-build or update massive data sets means they are always out of date with respect to the new data cascading into the primary databases (given the complexity of the processes this is entirely expected but they often do not display the dates when the primary records were extracted)
- They all have different look-and-feel interfaces thereby necessitating regular practice to get the best out of them
- The web-based interoperativity can leave a lot to be desired; e.g. broken links, link-outs to databases that are not maintained to the same standards and overkill by linking out to too many similar sources
- Their automated annotation schema can be confounded by sequence artefacts (Southan *et al.*, 2002)
- The overlap between utility and content between major databases is extensive but is never enough for any of them to be the mythical 'one-stop-shop'
- Non-redundant transcript and protein collections may seem conceptually similar but because they diverge in schema details and update frequency they all give different statistics
- Some secondary databases such as SwissProt keep sequence identifiers both unique and stable but for technical reasons others, such as UniGene EST clusters or Ensembl genes, may change identifiers between builds
- Many specialized 'boutique' databases are never updated when their originators move on or run out of resources
- Last but not least some secondary databases that initially had free access can become commercial and require a subscription fee

2.11.3 Nucleic Acid Secondary Databases

For the analysis of their results the geneticist must become acquainted with these feature-rich sources of gene product information. A key example, based around nucleic acid sequence but including protein of secondary databases is LocusLink/RefSeq (LLRS) for mRNAs. The LLRS system is built round a reference sequence (RefSeq) which is usually the longest available mRNA of those coding for the same protein. RefSeq includes splice variants and if only genomic sequence is available, such as for many of the 7TM receptors, the system defaults to the predicted coding sequence annotated as a 'CDS' in the database entry. For example there is no experimentally determined human rhodopsin mRNA in GenBank, only a model mRNA predicted from the genomic sequence U49742. This presents an immediate problem for the geneticist, as the untranslated region (UTR) of the rhodopsin locus, which defines the boundaries and functional regions of the gene may be extensive. Chapter 4 takes a detailed look at approaches to help define the true extent of gene loci.

The end-product of the RefSeq pipeline is a unique mRNA, coding sequence (CDS), or set of splice variants for those gene products where data or predictions are available.

The LocusLink side of things, as suggested by the title, is directed towards mapping the RefSeq gene products onto the genomic sequence and checking the consistency between the two. LocusLink has linked sections of key importance to the geneticist. These are: variation which assigns SNP data, OMIM which includes verified monogenic disease links, homologene which indicates close homologues in other species, UniGene which specifies ESTs clusters associated with the gene product, and PubMed that links to all publications that can be specifically linked to the primary GenBank accession numbers. There are also links to all three genome portals, NCBI, UCSC and Ensembl. There has been some confusion in the past where the portals could not synchronize their builds and track displays with GP version updates but this problem has been addressed and they should all be on version 28 (from December 2001) at the time of writing.

The RefSeq identifier is secondary in the sense that it is a supplementary identifier assigned to one particular mRNA chosen as the reference sequence. These accession numbers have the prefix NM_ for mRNA entries and NP_ for protein entries. The LocusLink/RefSeq system goes one step further in assigning a third identifier, XM_ for nucleic acid and XP_ for proteins, which are the genomic counterparts of the NM and NP numbers. A BLAST search against the NCBI protein database will show all three entries, the primary accession number, the NM_ and the XM_ entries. There is the added complication that the XP_ sequences have a variable evidence support level and include *ab-initio* genomic predictions both with and without EST support. Secondary accession numbers are also important for ESTs. ESTs can be considered as mRNA fragments that, with sufficient sampling (now just exceeding 4 million human entries in dbEST) can be clustered or assembled to form a contiguous extended transcription product and in some cases, the splice variants from the tissue types sampled for EST preparation. The main post-genomic utility of EST collections is as exon detectors. In addition to splice variants these can reveal possible gene transcription activity where no extended mRNA has been experimentally verified. The primary data source for ESTs is the dbEST division of GenBank.

The geneticist should be aware of two major secondary EST databases, UniGene (Wheeler *et al.*, 2002) and the TIGR human gene index (Liang *et al.*, 2000). The principles by which these different databases are constructed, are explained in the appropriate source references but in fact they both converge to a similar set of 'virtual' surrogate transcripts. In the TIGR case, the virtual transcripts assembled from overlapping ESTs can be retrieved; in the Unigene case, the individual EST reads can be batch downloaded. As with most secondary databases, built from the same source data, the two databases have both overlap and complementarity. The TIGR assemblies are particularly useful for extending the 3' UTR of known mRNAs but the assemblies are re-compiled at long time intervals. UniGene is updated more frequently and is fully interlinked to the LocusLink/RefSeq system but the clusters are built on mRNAs from the preceding version of GeneBank.

2.11.4 STSs and SNPs

These are two of the most important data sources for the geneticist involved in disease mapping. The dbSTS database contains sequence and mapping data on short genomic landmark sequences. Although they have a primary sequence record and GB accession number they also have a number of alternative marker names. These have been cross-referenced into a secondary database called UniSTS that integrates all available marker and mapping data (<http://www.ncbi.nlm.nih.gov/genome/sts/>). The dbSNP database is an interesting exception in that it is not a division of GenBank so it is not strictly a primary database. The

submissions (SS numbers) are equivalent to a primary record but overlapping sequences with the same polymorphism are collapsed into the Reference SNP Cluster Report with an RS number. This can be considered a secondary database where the RS numbers are non-redundant and stable. These RS numbers, currently at 2,640,509 for human, are integrated with other NCBI genomic data and primary GenBank records containing overlapping sequences deduced or stated to be from the same location. The HGVbase has a smaller set of 984,093 highly curated records (<http://hgvbase.cgb.ki.se/>). They have their own secondary accession/ID number and these can be queried and retrieved from the Ensembl genome annotation. Chapter 3 presents detailed examination of the major databases of genetic variation.

2.11.5 Protein Databases and Websites

A website of central importance in protein analysis is the Expert Protein Analysis System (ExpPAS; <http://www.expasy.ch/>). In addition to protein analysis tools, such as PROSITE (<http://www.expasy.ch/prosite/>) and Swiss-3Image (<http://www.expasy.ch/sw3d/>) Swiss-Prot protein database contains high-quality annotation and web-linked cross-references to 60 other databases. It is accompanied by TrEMBL, a computer-annotated supplement that contains the translations of all coding sequences present in primary nucleotide sequence databases not yet in SwissProt. Sequence records are merged where possible to minimize the redundancy. Sequence conflicts and splice variants are indicated in the feature table of the corresponding entry. The combined database is referred to as SwissProt/TrEMBL (SPTR). Amongst the links in SPTR it is worth mentioning the InterPro system which is of very high utility for finding protein family-specific domain matches (Apweiler *et al.*, 2000). Acquiring this information is one of the main goals of the bioinformatic analysis of proteins so it is useful to find that this piece of the work is already done and updated with new releases of InterPro. Other major sites provide PFAM, PROSITE, and other tools for protein sequence analysis. The Sanger Institute (<http://www.sanger.ac.uk/>) provides access and maintains PFAM and multiple other useful links and genomic tools, including three-dimensional protein structure prediction (<http://genomic.sanger.ac.uk/123D/123D.shtml>).

Any division between the universe of DNA and protein sequences is clearly artificial. Protein information can be accessed from within the LLRS system, just as it is also possible to link out to primary nucleic acid sequence record accession numbers from SPTR. However, the complementarity between LocusLink/RefSeq and SPTR is clear. The focus is on nucleic acid sequences in the former and protein sequences in the latter. The message for the user is that both sources will be essential for interpreting the results of genetic experiments.

2.12 CONCLUSIONS

In this chapter we have introduced the major data sources available on the internet that geneticists increasingly need to access for their research. The choice was based on our direct working experience of their utility. Rather than restrict ourselves to just cataloguing these, we have also included some discussion of the principles behind the organization of biological data, such as the concept of primary and secondary sequence databases. We have also demonstrated the power of web search engines, both of the specialist and common variety. Mastering these is essential for interrogating biological resources on the

internet. They also allow the user to search for new developments, tools and databases. This is something we strongly recommend to future-proof your own research, even if we cannot future-proof this book!

REFERENCES

- Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, *et al.* (2000). InterPro — an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16**: 1145–1150.
- Berners-Lee T, Fischetti M, Dertouzos M. (1999). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. Harper: San Francisco.
- Butler D. (2000). Biology back issues free as publishers walk HighWire. *Nature* **404**: 117.
- Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, *et al.* (2002). The Ensembl genome database project. *Nucleic Acids Res* **30**: 38–41.
- Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet* **25**: 239–240.
- Southan C, Cutler P, Birrell H, Connell J, Fantom KG, Sims M, *et al.* (2002). The characterization of novel secreted Ly-6 proteins from rat urine by the combined use of two-dimensional gel electrophoresis, microbore high performance liquid chromatography and expressed sequence tag data. *Proteomics* **2**: 187–196.
- Schuler GD, Epstein JA, Ohkawa H, Kans JA. (1996). Entrez: molecular biology database and retrieval system. *Methods Enzymol* **266**: 141–162.
- Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, *et al.* (2002). Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res* **30**: 13–16.
- Zdobnov EM, Lopez R, Apweiler R, Eitzold T. (2002). The EBI SRS server — recent developments. *Bioinformatics* **18**: 368–373.