The largest repository of current information and tools for biological analysis, is the World Wide Web. This can be in the form of a large database of sequence information, or databases of clinical information. More recently, integrated databases are being built, and sites such as Ensembl (http://www.ensembl.org) and LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/) offer sequence information from *Homo sapiens* together with preliminary analysis and location within the genome. Much of this data is linked to other sources of information about a particular biological molecule.

Common analysis tools are also found on the web, and many of them can be used through a browser interface for free. A particularly wide range of tools for protein analysis is offered by the Expert Protein Analysis System housed at the Swiss Institute of Bioinformatics in Geneva (http://www.expasy.ch). Reliability is particularly important when choosing a site from which to use bioinformatics tools. Large national nodes are generally funded by the local government, or an EU grant and should not only remain under the same URL, but should also retain a useful connection. Many of these sites have a helpdesk to pass on comments or ask for assistance. A good site will also have on-line help for each parameter within a tool.

Not everything is free to industry and many sites will ask for registration of an individual with commercial connections. This often corresponds to a fee, as institutions need extra sources of funding.
Contents

Internet Data Resources ............... 1

Non-sequence Databases ................... 1
   GeneCards ............................................. 1
   GDB .................................................. 2
   OMIM ................................................. 3

Bibliographic Databases .................... 4
   PubMed ............................................... 4
   World of Science .................................... 5
   PubCrawler ......................................... 5

Clinical Databases ........................ 5
   Human Gene Mutation Database ............... 5
   dbSNP ............................................... 7

Sequence Databases ......................... 7
   SRS – Sequence Retrieval System ............... 7
   ExPASy .............................................. 10
   Ensembl ............................................. 11

More databases ............................ 14
   KEGG ............................................... 14
   WIT ............................................... 14
Internet Resources

TIGR .................................................................................. 15

ArkDB .............................................................................. 15
Internet Data Resources

Much commercial interest is invested in medical genetics, which often entails information on Human genetics, or the bacteria that affect them, and other animals. Many of the databases will contain information on all species, but others are species specific. This selection will introduce you to what is available for Human genome analysis, and subsequent practicals concentrate on sites for other animals, plants and bacteria.

This course will use aniridia and its underlying protein as an example throughout.

Non-sequence Databases

GeneCards

For a subject (such as aniridia) you know little or nothing about, the best plan is to find some information on it! GeneCards is an integrated database, which assimilates biochemical information from a wide range of resources and is a good place to start. It is automatically curated at the Weizmann Institute in Israel and offers concise information on human genes and their functions. The data it provides is free to academics, although use by industrial employees requires a licence.

Go to http://bioinformatics.weizmann.ac.il/cards/ 1 and search the site using the keyword² aniridia. You are rewarded with a screen offering you the summary of the contents of a single GeneCard³. To display the complete Card, follow the link to the left of this summary.

There is a wealth of information available from this GeneCard. All names are used according to the HUGO Gene Nomenclature Committee⁴ and the correct name of the gene involved in aniridia is given as the name of the GeneCard, with

---

1 Although this site is held in Israel, there are several mirror sites around the world. Because internet traffic can be slow, it is generally a good idea to use sites as close to home as possible. Unfortunately, some mirror sites are not always up-to-date. The Israeli site is also linked from the HGMP bioinformatics pages.

2 Ensure that the “Search/Display GeneCards™ by” is set to keyword.

3 If your search had produced many cards, the “open a small search window” button would allow you to conduct a more stringent search in a separate display.

4 URL: http://www.gene.ucl.ac.uk/nomenclature/
Internet Resources

synonyms used by various databases also listed. The Card is divided up into two vertical columns. The left hand column displays a definition of the information held in the right hand column. The links in this left hand column will take you to the homepage of any of the databases listed. In contrast, the links in the right hand column take you to the database entry for the specified gene.

Words that are identical to those specified in your original search are marked in red.

It is worth noting down certain information from this and subsequent databases as it will help to constrain other searches you may do.

Gene responsible for aniridia .......................................................*

Cytogenetic location .................................................................

GDB ID number ........................................................................

Each sequence is given a unique accession number when it is deposited into a database. This number should remain with it permanently and enable it to be found rapidly as the result of a keyword search.

Scroll down until you reach the sequences section. Check out the sequences listed under "Additional Gene/cDNA sequence". Why are there several listed?

GeneCards is not the only integrated database. There are many others with links to entries relating to your keyword search. For example, GDB and OMIM. You can also access both these databases from the GeneCard.

GDB

The Genome Database (GDB) started life as the central repository for human mapping and genomic data. It is hosted at the bioinformatics centre of the Hospital for Sick Kids in Toronto, Canada. The term "Genomic Segment" is used to define a region of the genome, which may be anything from cytogenetic marker, to a complete clone. The entries are owned by the individuals who deposited them, and these owners maintain sole editing rights. The GDB does not store sequence or raw mapping data.

---

5 You should have found out that the gene implicated in aniridia is pax6 - a paired box homeotic gene.

5 Both GDB and OMIM are directly accessible from the HGMP WWW menu.
Internet Resources

Open another browser window and go to http://www.gdb.org/gdb/. Select the Search for "Genomic segments" and search by Name/GeneID. Enter in the GDB identification number (prefixed by GDB:) for the aniridia gene that you wrote down from your GeneCards search and click on "submit".

GDB also offers plenty of information of the names and cytogenic localisation of the pax6 gene. It also displays information of the experimental determination of genome localisation. Further down there are links to mutations, phenotypes and homology.

Follow the link “PAX” under the section header “Families” and note down the consensus pattern of the paired box in the “definition” section.

Paired box pattern .................................................................
Pattern span (residue to residue) ..........
........................................................................
The information to be found includes the identity of other members of the family and what their function is.

Number of PAX genes .................................................................
Possible gene function .................................................................

OMIM

OMIM is the Online version of Victor McKusick's Mendelian Inheritance in Man. It is a database of phenotypes of human diseases - with a substantial genetic component.

Follow link to this database from the Pax6 GeneCard. Read some of the OMIM entry. In particular have a look at some of the database links. Are there any that might be useful to you in the future?

6 It is always best to use identification numbers to search databases if at all possible. To illustrate this point, you may like to do a keyword search of the GDB using the word “aniridia”, or a search such as pax* which uses a wildcard.
Bibliographic Databases

PubMed

If you follow a citation link from the GDB, and then the Medline Unique ID link, you will arrive in PubMed, a freely available implementation of the Medline database offered by the National Library of Medicine (NLM) as part of the service of the American public bioinformatics effort, the NCBI. Most widely used molecular biology databases link to literature references via PubMed.


This takes you to the site of the NCBI in the U.S.A and their own search engine called Entrez. It is linked to all information at the NCBI and can be searched using different categories.

Ensure that the Search has contains the word PubMed, in order to search for a literature reference. To alter it, pull down the menu by clicking on the arrow and select the appropriate category. Type in aniridia and press hit "Go". See how many references you get. Try that again with the keyword pax6. Do you get the same number of references? Now try pax6 and aniridia. What about using wildcard searches (e.g. pax*)?

You should get fewer entries on using "pax6" as the keyword and much less if you combine both "pax6 and aniridia" together. As with other keyword searches, it is worthwhile thinking about what you really want before entering words in the text field.

As a component of Entrez, PubMed links to all other types of information held at the NCBI. PubMed is a comprehensive database containing over 11 million Medline citations dating back to 1966. The coverage of this database is a very broad interpretation of the medical literature, but if your sequence of interest is not even remotely connected to a human disease, it may be that Medline does not contain all papers of interest, and other sources contributing to the PubMed resource should display them.

---


---

Human Genome Mapping Project Resource Centre http://www.hgmp.mrc.ac.uk
Internet Resources

**World of Science**

Alternatively, you may want to investigate the **World of Science (WoS)** at [http://wos.mimas.ac.uk](http://wos.mimas.ac.uk). This database has taken over where the Bath Information Data Service (BIDS) left off, and can be searched for any scientific category, including social sciences, arts and humanities. To access this database, you need to register your institution and gain an Athens username and password.

**PubCrawler**

If you were really going to enter this field, you may wish to subscribe to one or more "alerting services". These will automatically send you an email when, for example, relevant papers are published or nucleotide or peptide sequences are released. **PubCrawler** is an alerting service that scans daily updates to the PubMed and Genbank\(^8\) databases and can be found at: [http://www.pubcrawler.ie/](http://www.pubcrawler.ie/). You need to register to use this service. You might want to check out their sample results page.

**Clinical Databases**

**Human Gene Mutation Database**

If you are studying a disease, the chances are that the cause of this is an underlying genetic mutation. The **Human Gene Mutation Database (HGMD)** is held at the Institute of Medical Genetics in Cardiff, Wales and contains sequences and phenotypes of human disease-causing mutations. The database can be searched in a variety of ways including disease name, gene name or symbol and also by OMIM or GDB accession number.

The collaboration of HGMD in October 2000 with Celera Genomics\(^9\) highlights the funding crisis that many public databases face. Initial grants to set up databases fade away leaving nothing left for updating and maintaining these systems. Celera has now bought a period of exclusivity at the HGMD. The data

---

\(^8\) Genbank is the American repository for nucleotide sequences.

\(^9\) Celera Genomics is an American organisation, which is aiming towards becoming “the definitive source of genomic and related biological information”. The slight flaw in this plan is that you must subscribe to the service, and whilst biotechnology firms may be able to muster the required finances, not a lot of government funded research posts have the money available.
Internet Resources

for newly discovered mutations will become available to Celera subscribers prior to it becoming available to the public. The revenue this has produced will ensure the survival of HGMD.

Go to http://www.hgmd.org and choose "HGMD Search". Type aniridia into the keyword field and submit the search. The results of your search are at the bottom of the page where you will see a link to "PAX6"10. Follow this link to the index of mutations where the entries are grouped in tables according to mutation type or phenotype.

The data below the links from the phenotype groupings link directly to several integrated databases.

The HGMD only records the first phenotype caused by a mutation, so it is not impossible that reported mutations may also cause other diseases.

Select "nucleotide substitutions (splicing)" and follow the link. Check there is the same number of mutations as specified in the original table. Do all mutations cause aniridia? The number on the right hand side indicates the number of the reference attributed to that mutation. Follow one of these reference links - they display the abstract of that particular reference from the PubMed entry.

Each splice mutation is given a unique accession number; IVS11 and whether it is a donor or an acceptor splice site12. The location of the substitution is then relative to the relevant splice site of the defined intron. The substitution is then displayed together with the phenotype and the number of references.

Substitution (accession number CS982309) ............................................................

Location .................................................................................................................

10 Your results appear below the search options. If your search retrieved no results, then nothing will appear in this space and you should conduct a new search using a different keyword.
11 IVS – Intervening sequence – i.e. introns.
12 Donor splice sites are located on the boundary between the right end of an exon and the left end of an intron. Acceptor splice sites are located on the opposite boundary. The mutation is given relative to the ds or as of a specified intervening sequence.
Internet Resources

**dbSNP**

One of the features to be observed from the sequencing of the Human Genome, was the presence of individual base-pair mutations, or Single Nucleotide Polymorphisms (SNP). There are several SNP databases held around the world, possibly the best known of these is dbSNP held at the NCBI. Searching these databases, however, is not the easiest of tasks if you do not have the SNP identification number. There are several ways of obtaining this, and we will look at a couple in the next section using sequence database searches.

**Sequence Databases**

There is one main nucleic acid sequence database and one main protein sequence database in widespread general use among the biological community. The nucleotide database is EMBL\(^1\) (or GenBank or DDBJ) and the protein database is SWISS-PROT.

Shortcuts to all these databases can be found on the Genome Web (accessed from the HGMP main page at [http://www.hgmp.mrc.ac.uk](http://www.hgmp.mrc.ac.uk)).

**SRS – Sequence Retrieval System**

As these databases contain hundreds of thousands of sequences, searching through them requires the processing power of a computer search engine. The Sequence Retrieval System (SRS) has been designed to do just that. SRS is available at many sites over the world. However, every site allows access to a different set of databases and, sometimes, search and analysis tools.

⭐⭐⭐ Go to the HGMP homepage at [http://www.hgmp.mrc.ac.uk](http://www.hgmp.mrc.ac.uk) and select the SRS link from the right hand side. Click on the “start” paw pad\(^1\) and you should be faced with a page offering you several databases\(^1\) under a “sequence libraries- complete” heading.

---

1. Neucleic acid sequences may be submitted to any to the three databases, depending on whether you are resident in Europe (EMBL), America (Genbank) or Asia (DDBJ – DNA DataBase of Japan). Newly submission sequences are transferred between the three databases on a daily basis.

2. This is Version 6 of SRS and the paw pad is that of a lion. This represents the company, which produce SRS, Lion Biosciences Ltd. Although Lion is a commercial company, SRS software is made freely available to academic institutions.

3. If you do not know what these databases are, click on the name and it will link to a description of the data held.

4. EMBL and SwissProt contain hundreds of thousands of sequences, and new sequences are added every day. There is a new release of these databases approximately every three months, and any new submissions before that date are released as updates. The separate libraries can be seen by exploding the “Sequence libraries – subsections” heading.
Internet Resources

Select the EMBL database by clicking in the little box to the left of it and then select the **standard** query form on the left hand side of the page.

In the first editor field, enter **pax6** and leave the query as "All Text". In the second field, enter **human** and alter the query to "Organism" using the scroll menu. At the bottom of your page, under the "Create your own view" section, highlight "AccNumber". Now press the Control key on your keyboard and also highlight "Description".

Ensure that the "append wildcards to query" option on the left hand side of the page is turned off (there should be no tick in the box) and click on the yellow "Submit Query" button.

Scroll down the list of hits looking for the genomic sequence. Why would the rest of these be unsuitable?

You should have received a list of approximately 19 sequences. The left hand column contains the sequence identifiers. An identifier consists of the name of the databases housing the sequence, followed by a colon and some alphanumeric characters defining the sequence. These identifiers are unique to the sequence and database. The subsequent columns are the results of your choice on the standard query form - in this case accession number and description. Accession numbers are unique to the sequence, and should not change between databases. The final column contains the requested description of the sequence and is perhaps a better initial indication of what you are looking at.

You should also have noticed that one clone mentioned in GDB, with accession number M77844, was not pulled out from this SRS search. If you go back to SRS and search for M77844 in the field AccNumber, you will find that the word "pax6" is not mentioned. This gene is described as ocularhombin, which you may remember has been mentioned before as an as alternative name for PAX6. SRS does text-based searches; you need to think hard about the best terms to use in your search.

From the description you should have identified EMBL:HSA1280 (Acc: Z83307) as the genomic sequence. The other sequences are either incomplete or contain an mRNA sequence.

Select the genomic sequence link and go directly to the EMBL entry. Select "text entry" and examine the EMBL entry for the full length PAX6 clone, as an illustration of the format used in EMBL data files.
The EMBL Data Format

<table>
<thead>
<tr>
<th>ID</th>
<th>Sequence identity, type, organism and length in short form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>Accession number</td>
</tr>
<tr>
<td>DT</td>
<td>Dates of entry creation and modification</td>
</tr>
<tr>
<td>DE</td>
<td>Sequence description</td>
</tr>
<tr>
<td>KW</td>
<td>Keywords</td>
</tr>
<tr>
<td>OS, OC</td>
<td>Organism and full taxonomy</td>
</tr>
<tr>
<td>RN</td>
<td>References linked to Medline (PubMed entry numbers)</td>
</tr>
<tr>
<td>DR</td>
<td>Link to protein sequence in SwissProt</td>
</tr>
<tr>
<td>CC</td>
<td>Comments</td>
</tr>
<tr>
<td>FT</td>
<td>Feature Table: table of features within the sequence; each can be accessed as a separate “pseudo-entry” including: source coding sequence (CDS) “miscellaneous features”, here a CpG island repeat regions</td>
</tr>
<tr>
<td>SQ</td>
<td>Sequence (beginning with sequence composition)</td>
</tr>
</tbody>
</table>

The feature table is one of the most important aspects of the EMBL data format. Any sequence feature can be expressed in this format. It is easy to express complex ideas, as features can be combined and arranged hierarchically. As one example, a number of exons can be arranged into a single coding sequence.

The sequence format used in GenBank, which is similar to EMBL format, also includes feature tables.

You can return to the "pretty" view of the sequence by pressing the link at the top of the page to "Embl entry". The default for EMBL and SwissProt is to show you the pretty view of the database entry - but both text and pretty view hold the same information.

It would be possible to select the SwissProt entry for the PAX6 protein sequence just by clicking on the single link in this entry. This may not be the case in a more complex query. We will therefore go back to SRS to retrieve the protein sequence by linking to the SwissProt database.

Use the "Back" button on your browser's tool bar to go back to the query page of SRS defining your hits. Ensure that the "all but selected" box is ticked and then select the yellow "Link" button. Select the "SwissProt" database and click on "Submit Link". Follow this link to the Swissprot entry of the Pax6 protein. Verify the
Internet Resources

name of the genomic clone in the EMBL cross-references. Why are there several references?

SwissProt Accession Number .................................................................
SwissProt Identifier .............................................................................

If you are looking for SNPs, the dbSNP can also be queried using the “Link” facility.

Go back again to the SRS query page defining your protein hit. Once more ensure that the “all but selected” box is ticked and click on the “Link” button. Select the “SNP” database header in black towards the bottom of the page and explode it by clicking the plus sign in the box. Click the box beside “HGBASE” so that it is ticked, and “Submit Link”. SNP identifier .................................
Submitter identifier .................................................................

ExPASy

If you are doing more work with proteins, you might want to investigate the Expert Protein Analysis System (ExPASy) held at the Swiss Institute of Bioinformatics. This site not only holds the SwissProt and TrEMBL databases, but also offers many tools for the user to analyse their protein sequences. We will return to ExPASy later on in the course, but for the moment, we will query the protein database for our Pax6 protein.

Go to http://www.expasy.ch and select “SWISS-PROT and TrEMBL from the left hand side of the page. Type EITHER the accession number OR the identifier you retrieved using SwissProt into the “Quick Search” field and click the button. Examine the

---

17 This approach is only possible where the SRS search engine is connected to the required databases.
18 The Human Genic Bi-allelic Sequences Database attempts to summarise all sequence variation data in the Human Genome. This also includes repeats and indels.
19 These are important if you wish to query dbSNP at the NCBI
20 TrEMBL is a database which holds peptide sequences resulting from automatic translation and annotation of the EMBL nucleotide sequences - hence Translated EMBL. It is generally seen as a precursor to SwissProt which contains manually annotated peptide sequences.
entry. Although the layout of the data in the NiceProt entry may be slightly different, the information is exactly the same as you saw using NiceView via SRS at the HGMP.

You will also notice that this view offers the possibility of some analysis – in particular the inclusion of this protein in a BLAST search. Following this option will display a textual output of a sequence homology search of the databases SwissProt and TrEMBL. We will be looking at BLAST searching later in the course.

There have always been various sites that combine the information gleaned from the sequencing of the Human Genome and if you insist on using the NCBI in America, you would find the information at http://www.ncbi.nlm.nih.gov/genome/guide/human/ very useful. In particular the list of FTP sites. This site incorporates cytogenetic data with sequence information, to offer you a detailed insight into the Human Genome.

Ensembl

This information can also be found at the European site http://www.ensembl.org. Ensembl is a collaborative project between the Sanger Centre and the EBI and is designed to allow you free access to all the genetic information currently known about the Human Genome. It automatically annotates current genomic data, and as this project advances, the site will change, and information will be added or removed as more evidence for or against a particular genetic feature becomes available. Genes are discovered using the Genscan and GeneWise programs, and corroborated with supporting evidence found in the protein databases.

Go to http://www.ensembl.org and select “Ensembl tour”, as it describes to you the various features of the site. Now go back to the home page, select the species “Human” and then browse the chromosome you know the Pax6 gene to be located on. Find the location on the chromosome where you know the gene to be. Is it a gene dense region? What about SNP density?

---

21 File Transfer Protocol is a method of transferring large files to your computer. It is preferable to email for this purpose and from public sites, the only way of accessing the data and software information.
22 Centre at Hinxton, Cambridgeshire, responsible for sequencing one third of the Human Genome as well as other genomes. Visit the centre at http://www.sanger.ac.uk
23 European Bioinformatics Institute – specialist EMBnet node responsible for the EMBL database and specialised bioinformatics tools. Located at Hinxton, visit the EBI at http://www.ebi.ac.uk

----------------------------------
Human Genome Mapping Project Resource Centre  http://www.hgmp.mrc.ac.uk
Internet Resources

Scroll to the bottom of the page and follow the "Browse OMIM diseases" link. The diseases associated with this chromosome are in alphabetical order. Work through until you find aniridia and follow the link to the gene (then long number beginning ENSG24)

You are now in the Ensembl GeneView, which contains a summary of all the information on the gene, plus links to that gene entry in many other databases. Do you recognise any of them? You will see from the GeneView, just as you saw in the overview, that there are two possible transcripts. This is because Ensembl uses the protein sequences currently stored in Refseq25 as its reference database. Refseq currently has two isoforms of this protein, to which two cDNA transcripts can be mapped.

Ensembl Gene ID ........................................................................................................................................

Gene Location ...............................................................................................................................................

Only one of these transcripts is homologous to a protein found in SwissProt - the PAX6_HUMAN. When writing the subsequent information, use only the information from this transcript (to find this, go back and look at the information from the pax6 database entries).

Number of Exons ........................................................................................................................................

Start and finish of each ................................................................................................................................

Number of splice sites ...................................................................................................................................

Splice site 7 sequence .....................................................................................................................................

Follow the link in the GeneView back to the ContigView by selecting the genome location link Z83307. Select the region of the chromosome containing the pax6 gene and then on the new page click on the actual pax6 gene (marked in red fill) to reach the overview.

The box displays one megabase of sequence and the contig26 region, which contains it. The region you choose for the overview may be altered using the...

24 Everything in Ensembl is numbered using the letters ENS followed by one which denotes the identity of the specification, i.e. G for gene, P for protein, T for transcript, etc.
25 Refseq is a standard, non-redundant database automatically annotated at the NCBI. The Reference Sequence project aims to provide reference sequence standards for all molecules involved in the Human Genome Project, from DNA through mRNA to proteins. Visit it at http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html
26 Contig is short for contiguous region and represents the sections of sequence that were initially...
Internet Resources

Zoom arrow. The long arrow at the top of the detail view screen represents the length of sequence you are viewing, in this case 1Mb. Below this there are several purple filled boxes indicating ESTs. By holding the mouse over these regions, a drop down menu will appear with the ID of the EST in question.

The blue line cutting across the detailed view shows the contigs, which were originally sequenced to provide the data. Each separate contig is shown in an alternating shade of blue. Below are the various features found on the other strand of that particular section of the genome. This strand looks much more promising. Just below the contig line, there are transcripts of the gene and the information from the Genscan program. You will see homologies found in the SPTR and EST databases, any markers known in that area, and any CpG islands that have been found. The bottom line displays the megabase scale of the sequence in that region of the chromosome.

Contig (containing pax6)

You should also see that three exons belonging to the pax6 gene are located on a different contig.

The transcript in this case is marked in both red and black - indicating this gene has two transcripts associated with it. The thick vertical blocks represent the exons, connected by the thinner horizontal lines. If you look closely, you can see there is a discrepancy in the number of exons in each transcript.

In addition to the information on the overview, you may now view SNPs and other information, by selecting further sources of data:

Select the "Features" tab and select coding SNP by clicking once on the text. Click once more on the "Features" tab to redraw the contig view. The SNP information is towards the bottom of the display. Do all the SNPs found here have an entry in HGBASE? Did you find this number of entries when you did the SRS link from SwissProt to HGBASE? If not, why do you think this is?

You will notice that to view the trace correctly, your browser must be "Java enabled". Java is a programming language that is becoming more and more popular. Netscape or Internet Explorer browsers using Version 4 and upwards should have this capability.

---

cloned, sequenced and aligned to created continuous areas of the genome.

27 Swissprot and Trembl (SPT) databases hold the majority of protein sequences available.

28 If you have a discrepancy in the number of SNPs found with various methods, the most likely reason is that the database you are using hasn’t been updated yet. Try using SRS at a different site, or go straight to the database homepage and compare the number of entries in each.
Internet Resources

More databases

We have only been able to include a few of the many databases available on the Internet in this case study. You might like to try exploring a few others if you have the time, using either PAX6 or a gene or protein you are interested in.

KEGG

The primary objective of KEGG is to computerise the current knowledge of molecular interactions; namely, metabolic pathways, regulatory pathways and molecular assemblies. At the same time, KEGG maintains gene catalogs for all the organisms that have been sequenced, and links each gene product to a component on the pathway. It also contains a database of all chemical compounds in living cells and links each component to a pathway component http://www.genome.ad.jp/kegg/kegg.html

WIT

WIT is short for "What is There". It is similar to the KEGG database in that it tries to integrate all the knowledge there is about an organism. WIT includes a set of over 2900 diagrams depicting metabolic pathways. This collection is from the Metabolic Pathways Database constructed by Evgeni Selkov and his team. Each pathway diagram includes a set of functional roles. The goal of producing a metabolic reconstruction is to identify which pathways are present in an organism and which genes implement the functional roles. WIT is based in the USA, at http://wit.mcs.anl.gov/WIT2/

Most of these databases are available via the WWW. Most of the important ones can be accessed from the HGMP-RC - try searching for them in the HGMP’s WWW menu if you are stuck. Some, especially certain ACeDB-style databases, require an X-Windows environment.

Most of the WWW-based databases will have links to related data in other databases and these links should be followed, as it will save you much time. Be aware that sometimes the links between databases are incorrect and you may

29 ACeDB (which stands for “A C. elegans DataBase”) started life as the repository for mapping and genomic data for the nematode worm C. elegans. Data on several human chromosomes and other organisms’ genomes are also held in ACeDB-style databases. You still need X-Windows to access some databases in this format, although they are increasingly being made available on the WWW.
Internet Resources

arrive at an inappropriate entry. If this is the case, and there is a helpdesk address available, it is very worthwhile letting them know. This includes the HGMP-RC - if we have got something wrong, we'd like to know about it.

If you are not working on Human genes, you may have to search further for information. The NCBI is a good place to start for completed genomes. These include many bacterial genomes at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db= Genome

TIGR

The Institute of Genome Research (TIGR) is also a good place to look. TIGR also offer Gene Indices of organisms that have not yet been completely sequenced. Try TIGR at http://www.tigr.org.

ArkDB

The ArkDB is a repository for sequence information from various animals generally seen in Britain as Poultry and Livestock. The database also offers information on animals such as the cat and salmon. http://www.thearkdb.org/ also has rather nice little cartoons of the animals in question! It is curated at the Roslin Institute just outside Edinburgh in Scotland using funds provided by the BBSRC\(^\text{30}\).

Also remember WWW search engines. http://www.google.com is probably one of the best, but use your favourite for a text search.

\(^\text{30}\) The Government funded Biotechnology and Biological Sciences Research Council.