

CHAPTER 1

Database resources for wet-bench scientists

Neil Hall and Lynn M. Schriml

1. INTRODUCTION

With the increasing amount of data being generated by genomic-scale studies, it has become much more important for biologists to store data in a structured way that makes it easily accessible and allows the integration of different data types from different sources. Hence, there are now hundreds of specialized databases available to biological researchers that cover a vast array of different data types from transcription factor binding sites to metabolic pathways and from protein domains to scientific journals. For these data resources to be properly exploited, one has to understand how the data is generated and structured, otherwise there is a danger that important data may be missed or, worse still, incorrect data blindly trusted.

Because of the number and diversity of data resources available, we cannot describe them all in a single chapter, so here we will describe the publicly available databases at the National Center for Biotechnology Information (NCBI) (1) (<http://www.ncbi.nlm.nih.gov/>^{1,1}) and provide examples of how you can query different information using their online tools. This chapter complements Chapter 2, which explores resources for navigating sequenced genomes with the emphasis on a second major group of tools – Ensembl. It should be noted also that there are many other tools available at different web sites and we will mention some of these later in this chapter.

1.1 Types of databases

The publicly available web resources described in this chapter can be divided into two types: primary and secondary databases. Whilst both types of database serve a useful purpose, one has to understand the distinction before making assertions based on a database query.

1.1.1 Primary databases

Primary databases include all of the repositories of the primary output from experimental work, such as GenBank (2) (<http://www.ncbi.nlm.nih.gov/Genbank/index.html>^{1.2}), which contains nucleotide sequences, and ArrayExpress (3) (<http://www.ebi.ac.uk/arrayexpress/>^{1.3}) at the European Bioinformatics Institute (EBI), which contains microarray expression data. These repositories generally house information that is submitted by the scientist who generated it and little is done to process, curate, or provide quality control over what is entered. Therefore, they are usually very comprehensive, but one must always treat the data with caution.

1.1.2 Secondary databases

Secondary databases are less all-inclusive than the primary databases, but instead concentrate on data quality and include additional information and cross-referencing. Secondary databases usually draw on (and may be linked to) a number of primary databases, collecting together information centered on a particular topic. For example, Pfam (3) (<http://www.sanger.ac.uk/Software/Pfam/>^{1.4}) is a secondary database that curates protein domains and allows users to search proteins for known domains, whereas the Mouse Genome Informatics (MGI) (<http://www.informatics.jax.org/>^{1.5}) collects and curates genomic, genetic, and functional data associated with the laboratory mouse. There is a clear distinction between these two examples of secondary databases: Pfam covers one theme (protein domains) and does so for all organisms, whereas MGI (4) covers many types of data but for only one species.

1.2 Database resources at NCBI

At the time of writing, NCBI has over 20 databases, which can be searched either individually or en masse. Entrez (5) is a system that provides a common interface to all of the major NCBI databases, including PubMed (6), nucleotide and protein sequences, protein structures, complete genomes, taxonomy, and many others. It provides a consistent user interface and format, and allows queries to be made across multiple NCBI databases at once. The starting point for Entrez is <http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>^{1.6}, and an overview, showing all of the Entrez-accessible databases and the connections between them, is available at <http://www.ncbi.nlm.nih.gov/Database/datamodel/>^{1.7}.

Each of these databases (which are often called nodes) will contain entries with unique identifiers (UIDs). These identifiers are stable over time, whilst the data associated with them can change. For example, a gene will always have the same UID, but over time we may discover a new function for it or new splice sites so its annotation and sequence could change. These entries can be linked within and between nodes, which allows, for example, publication entries to be linked to protein entries. Each node has a specific entry format, although many features will recur among the different nodes. For example, there is a fully annotated GenBank record at <http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>^{1.8}. Of particular interest are the three identifiers you will find in this single record: the *locus* name, the *accession number*, and the *GI*. The locus name (in this example

SCU49845) is unique to each entry and in many cases it may be identical or similar to the accession number (in this case, U49845). The accession number will always remain the same even if the entry is changed. There is also a *version number*, which indicates how many times the entry has updated: U49845.1 indicates that this is the first version of this entry. The GI is the 'GenInfo Identifier': if a sequence or protein translation changes in any way, a new GI number will be assigned, so GI sequence identifiers run parallel to the version numbers.

1.2.1 Nucleotide databases at NCBI

The main nucleotide database at NCBI (and, like the rest, accessible through Entrez) is GenBank (2), which contains all of the publicly available DNA sequences. However, there are subdivisions of GenBank that can be searched independently of the complete database. Depending on what you are looking for, it may be better to search a subdivision rather than the whole dataset. For example dbEST (7, 8) contains only that subset of sequence data and other information that relates to 'single-pass' cDNA sequences or expressed sequence tags (ESTs); similarly, dbGSS contains only single-pass genome survey sequences.

Further nucleotide databases (again, Entrez-accessible) exist outside GenBank. For example, dbSNP (9) (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp>^{1.9}) is a database of single-nucleotide polymorphisms, small insertions, and deletions. There is also a sequencing trace archive (<http://www.ncbi.nlm.nih.gov/Traces/trace.cgi>^{1.10}), which contains sequences that have been submitted along with all of their underlying experimental data, so that you can view the original trace file that generated them. This data could be particularly useful if you wish to check the validity of a frameshift or insertion in a gene of interest.

1.2.2 Protein databases at NCBI

Protein databases can contain different levels of information from primary sequence to secondary and tertiary structures. As well as databases that contain entire peptide sequences, there are a number of resources dedicated to collecting and curating protein domains and motifs. The NCBI protein database is a concatenation of a number of subdatabases: it includes sequences from Swiss-Prot (10, 11), the Protein Information Resource (PIR) (12), the Protein Research Foundation (PRF), and the Protein Data Bank (PDB) (13), along with protein sequences translated from nucleotide sequences of RefSeq (10, 11) and GenBank. Therefore, when you search using the Entrez server, you will be doing a comprehensive search of all publicly available sequences. Like the nucleotide databases, these proteins will be based on variable data quality depending on their source. For example, Swiss-Prot has highly curated annotations and should be nonredundant, whereas the translations from GenBank will contain sequence errors and misannotations in a number of cases.

Additional protein-related information is available in NCBI's Structure database (<http://www.ncbi.nlm.nih.gov/Structure/>^{1.11}) and NCBI's Conserved Domains Database (CDD) (<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>^{1.12}). CDD contains domains from Pfam (14), Simple Modular Architecture Research Tool (SMART) (15), and Clusters of Orthologous Groups (COG) (16), as well as other

domains curated at NCBI. The major utility of the domain database is for identifying domains in a protein sequence, which will allow the user to infer a function (also see Chapter 8).

1.2.3 Other databases at NCBI

As well as the major nucleotide and protein databases, NCBI houses a number of other related databases (nodes) that are linked to the sequence databases, as well as being themselves browsable and searchable through Entrez. One of the most commonly used databases is PubMed, which is a repository of biomedical journal articles. As well as being searchable using text queries, all of the articles in PubMed are linked to other NCBI entries related to them, such as nucleotide sequences. Similarly, there are databases of chemical structures (PubChem), microarray experiments (Gene Expression Omnibus, GEO) (17), taxonomy (Entrez Taxonomy), genes (Entrez Gene), maps (Map Viewer), and inherited diseases (Online Mendelian Inheritance in Man, OMIM) (18, 19) among others. Whilst some of these datasets may not seem obviously useful to your particular area of research, much of their functionality is derived from the fact that related records in each database are all linked so the user is able to traverse between datasets by following the links provided. For example, search the PubChem database for 'ethanol' and it will return not only a structure and description of the compound but also links to protein databases of enzymes that bind ethanol, as well as toxicology reports in the National Library of Medicine and relevant publications in PubMed.

2. METHODS AND APPROACHES

Here we discuss, in a little more detail, some of the tools available at NCBI through Entrez, tailored for searching their nucleotide, protein, and other databases. Additionally, we then provide a set of protocols, illustrating how to answer specific typical bioinformatics questions.

It is not possible to cover more than a tiny fraction of the resources, tools, and methods of query that are available through Entrez. However, we suggest that you start with the examples in the protocols and then take these as starting points to explore on your own.

2.1 Searching databases at NCBI

2.1.1 Text searches

The simplest and broadest search of NCBI databases is offered via the Entrez entry page: <http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>^{1,6}. Here, you can perform a simple text-based search across all databases or you can choose a specific NCBI database to search such as PubMed or Nucleotide. If you are searching across all databases, the simplest search you can perform is a text search of all fields. *Protocol 1* gives a simple example.

Protocol 1

A simple text search for *Plasmodium* across all NCBI databases using Entrez

1. Start at the NCBI home page: <http://www.ncbi.nlm.nih.gov>^{1,1} and select **Entrez home** on the right to go to the Entrez cross-database search page: <http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>^{1,6}.
2. In the text box towards the top of the screen ('Search across all databases'), type '*Plasmodium*'. Click the adjacent **Go** button, or press 'Enter' on your keyboard.
3. The page will be updated. Next to each of the database names and icons in the lower part of the screen will be a number (or, in a few cases, 'none'). This indicates the number of entries in each of these databases that contained, anywhere within it, the word '*Plasmodium*'.
4. Many of these records will not relate to *Plasmodium* itself. For example, some will describe proteins from other species that are noted as interacting with, or being similar to, *Plasmodium* proteins.
5. We therefore want to limit the search to entries in which the organism is *Plasmodium*. To do this, repeat the query, this time using the text '*Plasmodium*[ORGANISM]^a'
6. The page will be updated, this time giving the number of entries (in each of the databases) in which *Plasmodium* is in the 'Organism' field.
7. Clicking on any of the results will take you to the respective results page, listing all of the *Plasmodium* entries found. For example, clicking on **Nucleotide** or the adjacent icon will take you to the start of a list of over 200 000 *Plasmodium* nucleotide sequence entries.

Note

^aA list of fields that can be searched is given here: http://www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary_Matrices.html#Search_Fields_and_Qualifiers^{1,13}. Many fields can be abbreviated; for example, 'ORGN' can be used in place of 'ORGANISM'.

Search terms can also be combined; for example, searching for '*malaria* AND *mosquito*' will find all entries that contain (anywhere within the entry) both '*malaria*' and '*mosquito*'. Similarly, '*Plasmodium* NOT *Plasmodium*[ORGANISM]' will find all entries that refer to *Plasmodium*, but that do not originate from the organism *Plasmodium* itself. More sophisticated searches can be made by querying each database individually, rather than globally. The advantage of this is that it will allow you to define your search using fields specific to that database. It is also possible to view a 'history' of previous searches and to combine these together to refine the search further. *Protocol 2* gives a simple example of searching a single database, using 'limits' and 'history' to build up a progressively more refined query.

Protocol 2

A search in PubMed using Limits and History

1. Navigate to the Entrez entry page (<http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>^{1,6}) and click on the **PubMed** link towards the top of the page.
2. In the text field at the top of the page, enter 'malaria' and click the adjacent **Go** button.
3. The result is a list of over 40 000 (at the time of writing) entries that contain the word 'malaria' in any part of the entry.
4. We will repeat this search, restricting it to articles with 'malaria' in their title. Click on the **Limits** tab (just below the text entry box). Scroll to the bottom of the new page to find the pull-down menu **Default tag**. Select **Title** from the pull-down menu. Click the adjacent **Go** button (or scroll back up the page and click the one at the top).
5. The result is a list of about 20 000 articles, all with 'malaria' in their title.
6. We will now look at our previous searches and combine them to refine the search. Click on the **History** tab.
7. Into the text entry field, type 'mosquito'. Also, click on the ticked box on the **Limits** tab to 'untick' it. This removes our previous limits settings.
8. Below, you will see a list of your most recent searches. The top-most one in the list will be:

#xx Search **malaria** Field: **Title**

where 'xx' is a number. Click on the number.
9. A pop-up menu of options will appear, asking how you want to combine your previous search (for articles with 'malaria' in the title) with the current search (for 'mosquito', not limited to the title). Click on **AND**.
10. The text box should now show:

(mosquito) AND (#xx)

meaning that we are about to search for records that contain mosquito in any part of the entry and that also contain 'malaria' in the title (from our previous search). Click **Go**.
11. The result is around 3500 entries (at the time of writing), each with 'malaria' in the title and with 'mosquito' somewhere in the entry.

NCBI provides a web page giving further details of how to search their databases at <http://www.ncbi.nlm.nih.gov/entrez/query/static/help/helpdoc.html#Searching>^{1,14}. The following protocols give some examples of other ways to search the NCBI databases. The examples are in no way exhaustive, but they will introduce you to a range of search types that can form the basis of your own explorations.

Protocol 3

Determining the set of web resources available for a genome

1. Start at NCBI's home page (<http://www.ncbi.nlm.nih.gov/>^{1.1}) and click on the **Genomic Biology** link in the left blue bar. This link takes you to the genomic biology page: <http://www.ncbi.nlm.nih.gov/Genomes/>^{1.15}.
2. Under **Genome resources** on the right, select **Eukaryotic** to go to an alphabetic list of genome projects, listed by species: <http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi>^{1.16}.
3. Scroll down to find **Plasmodium falciparum 3D7** (at the time of writing, only one genome project is listed for this strain). On the same line, you will see a number of links, including a taxonomic identifier ('36329') and a link to the sequencing consortium's home page.
4. You will also see, at the right of the line, a series of colored abbreviations for different NCBI databases (**PM, R, G**, etc.). Clicking on any one of these will bring up data on *Plasmodium falciparum* 3D7 from the appropriate database. For example, clicking on **G** will show you all of the entries in the Genes database for this organism. If necessary, use your browser's 'back' button to return to <http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi>^{1.16}.
5. Clicking on the organism name at the left will take you to the Genome Project database and display entries for *P. falciparum*, offering further links to data and resources for this organism.

NCBI Map Viewer provides one way to access positional genome information and to integrate it into searches. *Protocol 4* takes you through a typical use of Map Viewer.

Protocol 4

Finding sequence-tagged site (STS) markers on chromosome 3 of *P. falciparum* using the NCBI Map Viewer

1. Navigate to the Map Viewer home page (<http://www.ncbi.nlm.nih.gov/mapview/>^{1.17}) by the **Map Viewer** link from the NCBI home page (<http://www.ncbi.nlm.nih.gov/>^{1.1}).
2. On the Map Viewer home page, select **Plasmodium falciparum** (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=36329^{1.18}) from the pull-down **Search** menu (leave the text field empty) and click **Go**. You will be taken to the Map Viewer page for *P. falciparum*, including an ideogram of the karyotype.
3. Enter 'STS' and '3' in the text fields **Search for** and **on chromosome(s)**, respectively, and click **Find**.
4. The results of your query are presented as hits on chromosome ideograms and in a tabular format. View the results in the Map Viewer graphical display by clicking on the **3** underneath the chromosome in the ideogram (to show all STSs) or by clicking on the blue links in the table below. Click on the first Map Element in the table (at the time of writing, this was **Pf2541**).
5. The resulting page will show STSs in a part of the chromosome, with the chosen STS (Pf2541) indicated. Clicking on its name will call up further information on that STS, including the polymerase chain reaction primer sequences.

Protocol 5

Searching for α -tubulin genes in *P. falciparum*

This search could be started from the Entrez Home page (searching all NCBI databases and then selecting those hits from the Gene database). Alternatively, as here, we can navigate to the Entrez Gene page to search only that database.

1. Navigate from the NCBI home page (<http://www.ncbi.nlm.nih.gov/>^{1.1}) to the Entrez home page (<http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>^{1.6}) using the link on the right of the screen.
2. Click on the **Gene** link or adjacent icon (left side of screen) to go to the Entrez Gene page (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene>^{1.19}).
3. Into the text box at the top of the screen, type:


```
Plasmodium falciparum[ORGANISM] alpha tubulin
```

(see *Protocol 1* for an explanation of using '[ORGANISM]' to limit a search to entries originating from a species). Click **Go**.
4. The query should return 11 genes (correct at the time of writing) with α -tubulin in the annotation. Click on the link for **PFI0180w** to get a detailed summary of its annotation.
5. In this case, the PubMed reference for the gene (under **Links** on the right of the screen) links back to the genome project paper (Hall *et al.*), rather than to an original paper about α -tubulin. From this, we might assume that this gene's annotation has been predicted by homology to other tubulin genes, rather than having been verified experimentally.
6. Use your browser's 'back' button to return to the Entrez Gene page for PFI0180w. Scrolling down the page to the section headed 'General gene information' shows that all of the Gene Ontology terms relating to this protein (see Chapter 9 for an introduction to Gene Ontology) were assigned on the basis of evidence-code 'IEA' ('Inferred from Electronic Annotation'), confirming our assumption that the annotation was not verified experimentally.
7. Additional examples of gene searches are given on the Entrez Gene home page.

2.1.2 Sequence similarity searches and alignment of transcripts to genomic sequences

A common method of querying sequence databases is by similarity searching. The most well-known tool for similarity searching is BLAST (Basic Local Alignment Search Tool) (17), which allows you to search your query sequence against a database of your choosing. Chapter 3 gives detailed information about BLAST and related tools; here we introduce the use of such tools in the context of the NCBI databases.

Similarity searches of NCBI's nucleotide and protein sequence databases can be restricted to sequences from one or more species either by specifying the organisms in the Options section of the BLAST page or by submitting searches against databases on the organism-specific BLAST pages (<http://www.ncbi.nlm.nih.gov/BLAST/>^{1.20}). Multiple query sequences can also be submitted in the same search using the organism-specific BLAST pages.

Protocol 6

Simple BLAST searches at NCBI

1. Navigate to the BLAST home page (<http://www.ncbi.nlm.nih.gov/BLAST/>^{1,20}) from the BLAST link included in the query bar found at the top of most NCBI pages.
2. The BLAST home page provides links to the suite of BLAST tools for comparisons between nucleotide or protein sequences. Searches may be conducted against highly divergent organisms (discontinuous megablast), the trace archive, the CDD, gene expression data in GEO, single-nucleotide polymorphisms, immunoglobulins, etc. In this case, we will search for proteins related to a *Plasmodium* α -tubulin, so select **protein blast** (under the 'Basic BLAST' heading).
3. On the 'Protein BLAST' page, leave all settings at their default values (note that we will be searching against the 'nr' or non-redundant database).
4. Open a new browser window and find the *P. falciparum* α -tubulin gene PFI0180w (see Protocol 5). When you have found the Entrez Gene page for this gene, scroll down to find the heading 'NCBI Reference Sequence (RefSeq)' and click on the link to the gene product (XP_001351911.1). This should bring up the corresponding Entrez Protein page and, scrolling down, you will find the complete amino acid sequence for this protein. Copy this sequence (along with the numbers and spaces) and paste it into the **Search** box on the BLAST page.
5. Click on **BLAST**. You will be taken to a page saying that your request has been successfully submitted. The page will be automatically updated when your results are ready (BLAST searches can take some time to complete).
6. When the results are ready, you will see a diagram representing the best matches. The colored bars indicate the score of the match and the portion of your query sequence that it matches. In this case, there will be many full-length red bars indicating many close matches to the complete α -tubulin sequence.
7. Below this are listed the hits in order of BLAST score (best first). If you click on the accession number of the hit, you will go to the GenBank entry for that protein. If you click on the BLAST score, you will go to the alignment (remember that there may be more than one alignment per hit).
8. Now try repeating the search, but looking only for matches in *Arabidopsis thaliana*. To do this, navigate to the 'Protein BLAST' page (step 3 above), but this time, type *Arabidopsis thaliana* into the 'Organism' text field (under 'Choose Search Set'), before continuing as before. (Note: as you type the organism name, you will be prompted with a list of likely organisms—simply click on *Arabidopsis thaliana* to save typing the complete name.
9. The result this time is a smaller number of matches, some of them shorter or of lower score, to *Arabidopsis* sequences.

BLAST is not always the best tool for sequence alignment and NCBI provides other tools that may be more appropriate for your needs. Alignment of a mRNA or cDNA sequence to a genomic sequence can be computed using NCBI's SPIDEY (17) (<http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/>^{1,21}) or SPLIGN (<http://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi>^{1,22}) alignment tools. Chapters 4 and 7 cover the alignment of transcripts with genomic sequence in more detail. Protocol 7 gives a simple example of using SPLIGN to compare a cDNA sequence with genomic sequence.

Protocol 7

Aligning a cDNA sequence to genomic sequence

1. From the NCBI home page, select **Tools**. From the tools page (<http://www.ncbi.nlm.nih.gov/Tools/>^{1.23}) click on the **Splign** link (scroll down to find it) to go to the SPLIGN page (<http://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi>^{1.22}).
2. SPLIGN can be downloaded to run locally or you can submit a cDNA and genomic sequence to SPLIGN at NCBI, which we will do here. Click on the **click here** link (<http://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi?textpage=online&level=form>^{1.24}) to submit an online job.
3. You will see text boxes to accommodate the cDNA sequence and the genomic sequence with which to align it. In the cDNA box, you can either paste a sequence or specify a sequence by its accession number. In this case, we will specify the cDNA for a chicken cDNA sequence: type the accession number 'AJ744697' into the cDNA box.
4. In the 'Genomic' box, you can again specify the sequence either by pasting it in or by giving an accession number. However, you can also select from a list of whole genome sequences using the pull-down menu underneath. Use this to select **Gallus gallus** (chicken).
5. Click on the **Align** button and wait until the results are ready.
6. The cDNA aligns to two places in the same genomic sequence contig (listed under 'Subject' at the top of the results page). Each alignment is a 'Model'.
7. For each model, the alignment of the cDNA to the genomic sequence is shown. This alignment will return six segments (six putative exons). The yellow boxes at the top represent the cDNA divided into the aligned segments; the genomic sequence is shown below. The vertical blue lines in the cDNA represent indels and the red lines mismatches in the alignments. To view the alignment for each segment, click on the graphical display or on the segment number.

Results of whole genome-to-genome pre-computed sequence comparisons are available in NCBI's Homologene resource, with highly similar sequences being represented as distinct Homologene groups. Text queries of the Homologene database yield results pages of matching Homologene groups containing highly related sequences from multiple organisms.

Pre-computed orthologs can be searched and browsed using Clusters of Orthologous Groups of proteins (COGS): <http://www.ncbi.nlm.nih.gov/COG/new/>^{1.25}. These are genes that are predicted to be functional equivalents due to the fact that they are derived by vertical descent from a single ancestral gene in the last common ancestor of the compared species. You can also compare your gene to known COGS using the kognitor tool: <http://www.ncbi.nlm.nih.gov/COG/grace/kognitor.html>^{1.26}.

Results of pre-computed protein comparisons can be viewed at NCBI's BLink (BLAST Link) resource (<http://www.ncbi.nlm.nih.gov/sutils/static/blinkhelp.html>^{1.27}). BLink results are provided by links on other Entrez database pages (e.g. **Entrez Protein**, **Entrez Gene**). BLink provides a tabular display of pre-computed highly related proteins for all organisms in the Entrez Protein database including hits to the CDD (<http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>^{1.28}).

2.2 Downloading NCBI datasets

For most scientists, web sites such as NCBI will provide all of the functionality they will ever need for their research. However, for people who want to do a lot of data searches, web sites become impractical and it is sometimes necessary to download entire datasets and software tools for searching them, so that the analysis can be done on a local computer.

NCBI provides this at their FTP site: <ftp://ftp.ncbi.nih.gov/>^{1.29}. In the `blast` directory, you will find all of the protein and nucleotide databases available through the NCBI web site, as well as executable files that you can install on Windows, Mac OS X, or Linux operating systems. In the `genomes` directory, you can download individual genomes. For people who want to create their own scripts that incorporate NCBI datasets or NCBI software tools, there is a set of file standards and software tools called the NCBI toolbox that will allow you to process files, run searches, and format output on your own machine: <http://www.ncbi.nlm.nih.gov/IEB/ToolBox/MainPage/index.html>^{1.30}.

One important caveat with installing your own datasets is that you will need to update them constantly as they go out of date, whereas if you are running searches over the internet this is not a problem. You should also test the output of any executable that you install to make sure that it is running properly and that your databases are indexed correctly.

3. TROUBLESHOOTING

- **I can't find my genome at NCBI. Where else can I search?**

This problem is becoming less widespread but it is not unusual. If a genome is published, then the rule is that it must be submitted to NCBI. However, if the genome project is still ongoing or it is unpublished, then it may not be submitted to GenBank but it may still be available. Many genome centers will submit ongoing projects to the trace archive at NCBI (<http://www.ncbi.nlm.nih.gov/Traces/trace.cgi>^{1.31}), so check there first. If that does not work, many ongoing projects have links from the genome project page: <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj>^{1.32}. If you still cannot find it, then you will have to search yourself; the major academic genome centers are listed below as additional web resources.

- **How do I reconcile different versions of a genome at the various sites?**

Unfortunately, you will find more than one version of a genome depending on where you look. This will happen if a genome project is ongoing but puts an intermediate version in GenBank in order to make the sequence widely available. The genome center may then continue to update their own version whilst leaving the old version in GenBank. GenBank will give a submission date in the top line of the entry. One must keep in mind that the GenBank version will be the 'official' version of the genome; this means that you will be able to give an accession number for this record in publications so that others can see the data. The genome center version is a transient file on a web site that

may disappear at any moment. So, whilst you may want to use the more recent version of the data, remember that it may not be there tomorrow.

4. ADDITIONAL WEB RESOURCES

Listed here is a selection of other web sites that are likely to be useful.

Major primary sequence generators

- The Wellcome Trust Sanger Institute: <http://www.sanger.ac.uk/>^{1.33}
- JGI Genomes: Eukaryota, Archae, Bacteria: http://genome.jgi-psf.org/tre_home.html^{1.34}
- Human Genome Sequencing Center, Baylor College of Medicine: <http://www.hgsc.bcm.tmc.edu/>^{1.35}
- The Broad Institute: <http://www.broad.mit.edu>^{1.36}
- The Institute for Genomic Research: <http://www.tigr.org>^{1.37}. Now renamed the J. Craig Venter Institute (JCVI; <http://www.jcvi.org>)
- Washington University Genome Sequencing Center: <http://genome.wustl.edu/>^{1.38}
- Genoscope: <http://www.genoscope.cns.fr/>^{1.39}

Bioinformatics institutes

- The European Bioinformatics Institute: <http://www.ebi.ac.uk>^{1.40}
- National Center for Biotechnology Information: <http://www.ncbi.nlm.nih.gov/>^{1.1}
- Center for Information Biology and DNA Data Bank of Japan: <http://www.cib.nig.ac.jp>^{1.41}

Genome annotation databases

- KEGG (Kyoto Encyclopedia of Genes and Genomes): <http://www.genome.jp/kegg/>^{1.42}
- GeneDB (Sanger Institute Pathogen Sequencing Unit annotation database): <http://www.genedb.org/>^{1.43}
- Ensembl Genomes: <http://www.ensembl.org/index.html>^{1.44}
- CMR (Comprehensive Microbial Resource) Annotated Microbial Genomes: <http://pathema.tigr.org/tigr-scripts/CMR/CMrHomePage.cgi>^{1.45}
- BRC Central (central web site of NIAID Bioinformatics Resource Centers, which houses databases of biodefense-related organisms): <http://www.brc-central.org>^{1.46}
- Genome properties (a database of curated and calculated properties of microbial genomes): <http://cmr.tigr.org/tigr-scripts/CMR/shared/GenomePropertiesHomePage.cgi>^{1.47}

Protein families, domains, and structures

- Pfam (curated protein domains): <http://www.sanger.ac.uk/Software/Pfam/>^{1.4}

- TIGRFAM (curated protein domains of microbes): <http://www.tigr.org/TIGRFAMs/index.shtml>^{1.48}
- SMART: <http://smart.embl-heidelberg.de/>^{1.49}
- InterPro (protein families, domains, and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences): <http://www.ebi.ac.uk/interpro/>^{1.50}
- Protein data bank (provides a variety of tools and resources for studying the structures of biological macromolecules): <http://www.rcsb.org/pdb>^{1.51}

Miscellaneous

- OBO (Open Biomedical Ontologies: an umbrella web address for well-structured controlled vocabularies for shared use across different biological and medical domains): <http://obo.sourceforge.net/>^{1.52}
- Amigo (a web interface for browsing gene ontologies, which will allow you to search for genes with specific functions or cellular locations, or that are involved in specific processes): <http://www.godatabase.org/>^{1.53}

5. REFERENCES

- ★ 1. **Wheeler DL, Barrett T, Benson DA, et al.** (2006) *Nucleic Acids Res.* **34**, D173–D180. – *This publication gives an overview of the NCBI databases and their associated tools. This reference is the most up to date at the time of writing, but NCBI publishes an overview of changes and updates in the Nucleic Acids Research database issue, which is published every year.*
2. **Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J & Wheeler DL** (2006) *Nucleic Acids Res.* **34**, D16–D20.
3. **Parkinson H, Sarkans U, Shojatalab M, et al.** (2005) *Nucleic Acids Res.* **33**, D553–D555.
4. **Eppig JT, Bult CJ, Kadin JA, et al.** (2005) *Nucleic Acids Res.* **33**, 5.
- ★ 5. **Geer RC & Sayers EW** (2003) *Brief. Bioinform.* **4**, 5. – *A tutorial paper that covers some of the ground of this chapter but in more detail. It gives a useful overview of the concepts behind the Entrez tool using example tasks.*
6. **McEntyre J & Lipman D** (2001) *CMAJ*, **164**, 1317–1319.
7. **Boguski MS, Lowe TM & Tolstoshev CM** (1993) *Nat. Genet.* **4**, 332–333.
8. **Banfi S, Guffanti A & Borsani G** (1998) *Trends Genet.* **14**, 80–81.
9. **Sherry ST, Ward MH, Kholodov M, et al.** (2001) *Nucleic Acids Res.* **29**, 308–311.
10. **Boeckmann B, Bairoch A, Apweiler R, et al.** (2003) *Nucleic Acids Res.* **31**, 365–370.
11. **Boeckmann B, Blatter MC, Famiglietti L, et al.** (2005) *C. R. Biol.* **328**, 882–899.
12. **Barker WC, Garavelli JS, McGarvey PB, et al.** (1999) *Nucleic Acids Res.* **27**, 39–43.
13. **Sussman JL, Lin D, Jiang J, et al.** (1998) *Acta Crystallogr. D Biol. Crystallogr.* **54**, 1078–1084.
- ★ 14. **Bateman A, Coin L, Durbin R, et al.** (2004) *Nucleic Acids Res.* **32**, D138–D141. – *Pfam is possibly the most useful web resource available for gene function analysis. It is useful to understand how it is built and annotated before diving in and using it. This publication should be updated each year in the database issue of Nucleic Acids Research.*
15. **Letunic I, Goodstadt L, Dickens NJ, et al.** (2002) *Nucleic Acids Res.* **30**, 242–244.
- ★ 16. **Tatusov RL, Fedorova ND, Jackson JD, et al.** (2003) *BMC Bioinform.* **4**, 41. – *An in-depth description of how orthologous groups have been calculated for the COG database; it also describes the eukaryotic clusters (or KOGS). This database is an excellent tool for studying the phylogenetic coverage of genes.*

17. **Barrett T, Suzek TO, Troup DB, et al.** (2005) *Nucleic Acids Res.* **33**, D562–D566.
18. **Hamosh A, Scott AF, Amberger JS, Bocchini CA & McKusick VA** (2005) *Nucleic Acids Res.* **33**, D514–D517.
19. **Cantor MN & Lussier YA** (2004) *Medinfo*, **11**, 753–757.