BSC4933/ISC5224: Introduction to Bioinformatics

Laboratory Section: Wednesdays from 2:30 to 5:00 PM in Dirac 152.

Biological Molecular Databases

Lab Two, Tuesday, January 14, 2009

Author and Instructor: Steven M. Thompson

The GCG Package, biomolecular databases, and how they are organized and accessed:

This lab introduces various methods for biological sequence and structural database access including those available on the World Wide Web, in particular at the National Center for Biotechnology Information (NCBI). Emphasis is on using the Accelrys GCG¥ Wisconsin Package® and its graphical user interface (GUI) SeqLab® with the local, on-site ¥GCG sequence databases. Data entry and format conversion are also briefly covered, as they are an important component of most bioinformatics chores.

Steve Thompson
BioInfo 4U
2538 Winnwood Circle
Valdosta, GA, USA 31601-7953
stevet@bio.fsu.edu
229-249-9751

Introduction

Standard disclaimer: I write these tutorials from a 'lowest-common-denominator' biologist's perspective. That is, I only assume that you have fundamental molecular biology knowledge, but are relatively inexperienced regarding computers. As a consequence they are written quite explicitly. Therefore, if you do exactly what is written, it will work. However, this requires two things: 1) you must read very carefully and not skim over vital steps, and 2) you mustn't take offense if you already know what I'm discussing. I'm not insulting your intelligence. This also makes the tutorials longer than otherwise necessary. Sorry.

I use three writing conventions in the tutorials, besides my casual style. I use **bold** type for those commands and keystrokes that you are to type in at your keyboard or for buttons or menus that you are to click in a GUI. I also use bold type for **section headings**. Screen traces are shown in a '`typewriter`' style `Courier` font and "`//////////`" indicates abridged data. The dollar sign (`$`) indicates the system prompt and should not be typed as a part of commands. Really important statements may be underlined.

The public biological molecular databases

This tutorial will restrict itself to openly available public biological sequence and structural databases and modified versions thereof. Most proprietary databases, especially as marketed to big drug discovery and development firms, are more-often-than-not beyond the financial means of many universities and research centers. The commercial exceptions are the on-site Wisconsin Package databases. These databases are modified versions of the public databases, modified by the GCG database routines included with the package, either by your GCG system administrator (me), or purchased pre-built on a license plan from Accelrys.

What are primary sequences?

Remember biology's Central Dogma: DNA → RNA → protein. Primary refers to one dimensional — all of the 'symbol' information written in sequential order necessary to specify a particular biological molecular entity, be it polypeptide or polynucleotide. The symbols are the one letter alphabetic codes for all of the biological nitrogenous bases and amino acid residues and their ambiguity codes (see http://virology.wisc.edu/acp/CommonRes/SingleLetterCode.html). Biological carbohydrates, lipids and structural information are not included within this sequence; however, much of this type of information is available in the documentation and annotation associated with each primary sequence in the databases.

What are sequence databases?

Sequence databases are an organized way to store an exponentially accumulating amount of sequence data. Each database has its own specific format with access most easily handled through various software packages and interfaces, either on the World Wide Web or otherwise. Three major database organizations worldwide are responsible for maintaining most of this data.

In the United States the National Center for Biotechnology Information (NCBI http://www.ncbi.nlm.nih.gov/), a division of the National Library of Medicine (NLM), at the National Institute of Health (NIH), supports and distributes the GenBank and WGS (Whole Genome Shotgun) nucleic acid sequence database and the unannotated GenPept CDS (CoDing Sequence) translations database, as well as a host of specialized sequence databases (in particular see the non-redundant RefSeq genomic, cDNA, and protein databases). The National Biomedical Research Foundation (NBRF http://www-nbrf.georgetown.edu/), an affiliate of Georgetown University Medical Center, maintains the Protein Identification Resource (PIR) database of polypeptide sequences, which is now a part of UniProt (see below).

The European Molecular Biology Laboratory (EMBL http://www.embl-heidelberg.de/) and the European Bioinformatics Institute (EBI http://www.ebi.ac.uk/) maintain the EMBL nucleic acid sequence database and the excellently annotated Swiss-Prot protein sequence database (also supported by the Swiss Institute of Bioinformatics, SIB, at ExPASy http://www.expasy.org/), as well as the minimally annotated TrEMBL (Translations from EMBL — those EMBL translations not yet in Swiss-Prot) protein sequence databases, in Heidelberg, Germany; Cambridge, UK; and Geneva, Switzerland. EBI, SIB, and PIR together constitute the UniProt Consortium (http://www.uniprot.org/)— a single, nearly non-redundant, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces freely accessible to the scientific community. Additional, less well known, sequence databases include sites with the military, with private industry, and in Japan (the DNA Data Bank of Japan, DDBJ http://www.ddbj.nig.ac.jp/). In most cases data is openly exchanged between the databases so that many sites 'mirror' one another. This is particularly true with GenBank, EMBL, and DDBJ; there is never a need to look in more than one of these places.

What information do they contain, how is it organized, and how is it accessed?

Sequence databases are often mixtures of ASCII and binary data. Even though expensive proprietary sequence databases are usually true relational or object oriented data structures, locally installed ones usually aren't, although some public domain ones are MySQL. It's all a very complicated mess with little standardization. Typical local sequence database installations contain several very long ASCII text files that contain information of all the same type, such as all of the sequences themselves, versus all of the title lines, or all of the reference sections. Binary files usually help 'tie together' all of the files by providing indexing functions. Software specific routines, as exemplified by genome browsers and text search tools, are by far the most convenient method to successfully interact with these types of databases.

Nucleic acid databases (and TrEMBL) are split into subdivisions based on taxonomy (historical) and data type. Protein databases are often split into subdivisions based on the level of annotation that the sequences have. This annotation includes much extremely valuable information — author and journal citations, organism and organ of origin, and the features table. The features table lists all sorts of important regulatory, transcriptional and translational (CDS coding sequence), catalytic, and structural sites, depending on the database. Actual sequence data usually follows the annotation.

Becoming familiar with the general format of sequence files for the type of software you want to use can save a lot of grief. Unfortunately most databases and many different software packages have conflicting format requirements. Fortunately there are many excellent format converters available. However, most sequence analysis software requires that you specify a proper sequence name and/or database identifier. These are usually discovered with some sort of reference text searching program, either on the World Wide Web or not. This brings up a point, locus names versus accession numbers. The LOCUS, ID, and ENTRY names category in the various databases are different than the Accession number category. Each sequence is given a unique accession number upon submission to the database. This number allows tracking of the data when entries are merged or split; it will always be associated with its particular data. Entry names may change; accession numbers are forever; they just pile up, primary becomes secondary, on *ad infinitum*.

What changes have occurred in the databases — history and development?

The first well recognized sequence database was Margaret Dayhoff's *Atlas of Protein Sequence and Structure* begun in the mid sixties (1965-1978). GenBank began in 1982 (1986), EMBL in 1980 (1986). They have all been attempts at establishing an organized, reliable, comprehensive and openly available library of genetic sequence. Databases have long-since outgrown a hardbound atlas. They have become huge and have evolved through many changes. Changes in format over the years are a major source of grief for software designers and program users both. Each program needs to be able to recognize particular aspects of the sequence files; whenever they change, it's liable to throw a wrench in the works. People have argued for particular standards such as XML, but it's almost impossible to enforce. NCBI's ASN.1 format and its Entrez interface attempt to circumvent these frustrations, but nobody else will adopt it. EMBL's SRS (Sequence Retrieval System) found on the World Wide Web at all EMBL OutStations and the Wisconsin Package's LookUp derivative of SRS also search for text in, interact with, and allow users to browse in the sequence databases. Both SRS and Entrez provide 'links' to associated databases so that you can jump from, for instance, a chromosomal map location, to a DNA sequence, to its translated protein sequence, to a corresponding structure, and then to a MedLine reference, and so on. They are incredibly helpful!

What other types of bioinformatics databases are used?

Specialized versions of sequence databases include sequence pattern databases such as restriction enzyme (e.g. http://rebase.neb.com/rebase/rebase.html) and protease (e.g. http://merops.sanger.ac.uk/) cleavage sites, promoter sequences and their binding regions (e.g. http://www.gene-regulation.com/pub/databases.html and http://www.epd.isb-sib.ch/), and protein motifs (e.g. http://us.expasy.org/prosite/) and profiles (e.g. http://pfam.janelia.org/); and organism or system specific databases such as the sequence portions of ACeDb (A *C. elegans* Database http://www.acedb.org/), FlyBase (the *Drosophila* database http://flybase.org/), SGD (the *Saccharomyces* Genome Database http://www.yeastgenome.org/), and RDP (the Ribosomal Database Project http://rdp.cme.msu.edu/). Many sites present their data in the context of a genome map browser, e.g. the University of California, Santa Cruz, bioinformatics group's genome browser, http://genome.ucsc.edu/; and the Ensembl project, http://www.ensembl.org/, jointly hosted by the Wellcome Trust Sanger Institute and

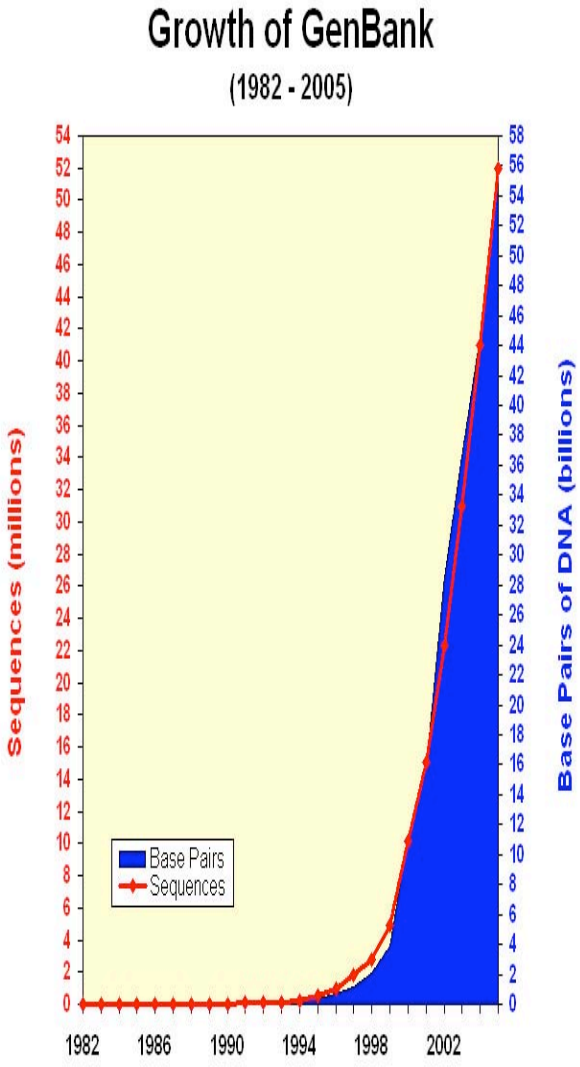
the European Bioinformatics Institute. Map browsers attempt to tie together as many data types as possible using a physical map of a particular genome as a framework.

Two other types of databases are commonly accessed in bioinformatics: reference and three-dimensional structure. Reference databases run the gamut from OMIM (Online Mendelian Inheritance In Man, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM), that catalogs human genes and phenotypes, particularly those associated with human disease states, to PubMed access of MedLine bibliographic references (the National Library of Medicine's citation and author abstract bibliographic database of over 4,800 biomedical research and review journals, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed). Other databases that could be put in this class include things like proprietary medical records databases and population studies databases.

Finally, the Research Collaboratory for Structural Bioinformatics (RCSB http://home.rcsb.org/, a consortium of three institutions: the State University of New Jersey, Rutgers; the San Diego Supercomputer Center at the University of California, San Diego; and the University of Wisconsin-Madison) supports the three-dimensional structure Protein Data Bank (PDB http://www.rcsb.org/pdb/). RCSB PDB is now a member of a World Wide PDB organization (http://www.wwpdb.org/), along with centers in Japan and Europe, whose mission "is to maintain a single Protein Data Bank Archive of macromolecular structural data that is freely and publicly available to the global community." The National Institute of Health maintains "Molecules To Go" at http://molbio.info.nih.gov/cgi-bin/pdb as a very easy to use interface to PDB. Other three-dimensional structure databases include the Nucleic Acid Databank at Rutgers (NDB http://ndbserver.rutgers.edu/) and the proprietary Cambridge small molecule Crystallographic Structural Database (CSD http://www.ccdc.cam.ac.uk/products/csd/).

Genome projects from the world over have kept the data coming at alarming rates. GenBank has staggering growth statistics (http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html). These statistics, starting with the birth of GenBank in 1982 at just over 600 sequences, are shown on the following page:

Year	BasePairs	Sequences
1982	680338	606
1983	2274029	2427
1984	3368765	4175
1985	5204420	5700
1986	9615371	9978
1987	15514776	14584
1988	23800000	20579
1989	34762585	28791
1990	49179285	39533
1991	71947426	55627
1992	101008486	78608
1993	157152442	143492
1994	217102462	215273
1995	384939485	555694
1996	651972984	1021211
1997	1160300687	1765847
1998	2008761784	2837897
1999	3841163011	4864570
2000	11101066288	10106023
2001	15849921438	14976310
2002	28507990166	22318883
2003	36553368485	30968418
2004	44575745176	40604319
2005	56037734462	52016762
2006	69019290705	64893747
2007	83874179730	80388382



It doubles in size about every 18 months! GenBank release 167, August 2008, has 95,033,791,652 bases, from 92,748,599 sequences, and this doesn't include the 118,593,509,342 bases worth of preliminary data in NCBI's WGS (Whole Genome Shotgun) database. Of those GenBank sequences 49 Archaea, 574 Bacteria, and almost 3,000 virus and viroid completely finished genomes are represented. Twenty-two to 241 Eukaryote genomes are present, depending on your definition of complete versus map versus draft (not even NCBI agrees with itself on this point!). Among them are cryptomonads, *Guillardia theta*, flagellates, *Leishmania major*, apicomplexan, *Plasmodium falciparum*, red algae, *Cyanidioschyzon merolae*, microsporidium, *Encephalitozoon cuniculi*, baker's yeast, *Saccharomyces cerevisiae*, fission yeast, *Schizosaccharomyces pombe*, nematode, *Caenorhabditis elegans*, mosquito, *Anopheles gambiae*, honeybee, *Apis mellifera*, fruit fly, *Drosophila melanogaster*, sea squirt, *Ciona intestinalis*, zebrafish, *Danio rerio*, chimp, *Pan troglogdytes*, human, *Homo sapiens*, mouse, *Mus musculus*, rat, *Rattus norvegicus*, thale cress, *Arabidopsis thaliana*, oat, *Avena sativa*, soybean, *Glycine max*, barley, *Hordeum vulgare*, tomato,

Lycopersicon esculentum, rice, *Oryza sativa,* bread wheat, *Triticum aestivum*, and corn, *Zea mays*. (conflicting genome statistics from NCBI: http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html, http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi, http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi, and http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi**).**

Your project molecular system choices

As I mentioned last week, you need to decide on a particular molecule with which to perform this and the remaining eight directed computer exercise tutorials. So that I can provide the necessary data, and to provide a diverse, yet level, playing field, this choice must be made off a list that I provide of four different 'hot' interest molecules. My list will contain molecular systems for which at least one experimentally solved protein structure and that protein's genomic DNA sequence is known. They will all be from organisms possessing exons and introns, to make the gene finding tutorial fair, regardless of choice. My apologies are offered to prokaryote biologists — sorry, but I think that you should know about splice-site recognition also. You will gain experience in all aspects of biocomputing covered in the course in a project-oriented fashion using the same natural progression as would be used in an actual experimental setting.

This approach should appeal to a wide cross-section of students working in diverse areas. Furthermore, you'll contrast predictive data derived from sequence analysis with known structural data. Parts may conflict, but common truths will also be found. In this way the strengths and weaknesses of each approach can be better understood, and a greater empathy is found for the tremendous problems encountered in the all-too-common case of a newly sequenced gene product with no discernable structural homologues. With this approach to computational molecular biology, you will "come full swing," gaining an appreciation for the full biocomputing spectrum available.

The directed exercise tutorial sequence lasts for the first two thirds of the semester, ten weeks. Scheduled lab sessions are devoted after that to working on and conferring with us about your semester final research project. Select the project molecule that interests you the most to use for the biocomputing practice tutorials, from the list below. Choose one that most closely fits the general type of work that you plan on doing in your academic or professional career, or just one that you'd like to know more about. Take your pick off of the following project molecular systems list:

1) Primitive, non-vascular plant (Viridiplantae ! Tracheophyta) ribulose bisphosphate carboxylase/oxygenase (RuBisCO), the nuclear encoded, small subunit only. This is a crucial enzyme in the Calvin cycle of photosynthesis, and, some would claim, the most abundant enzyme on earth.

2) Vertebrate (Vertebrata) c-H-Ras, also known as P21 Harvey ras proto-oncogene transforming protein. This incredibly 'hot' molecule is critically important in oncogenesis.

3) Vertebrate (Vertebrata) basic fibroblast growth factor, also known as heparin-binding growth factor 2 and prostatotropin. This is 'hot' cytokine is also relevant to cancer research and angiogenesis.

4) Fungal (Fungi) Cu/Zn superoxide dismutase (gene name *sod*). This is a cytopalsmic, oxireducatase type, free radical scavenger. Aren't free radicals implicated in both cell aging and cancer, isn't that the big deal with antioxidant vitamins and 'eating your fruits and vegetables?'

Exploring biological molecular databases

World Wide Web database sites

Activate and log on to the computing workstation you are sitting at. The first section of today's tutorial will deal with databases accessible through the World Wide Web using a browser. Therefore, launch whatever WWW browser is available on your local machine.

Here's some of my favorite WWW sites for biological molecular databases:

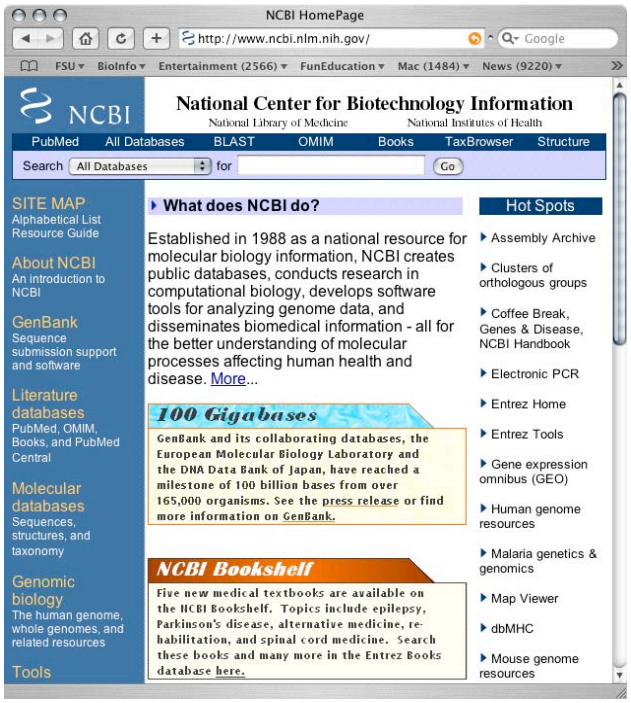
Site	URL (Uniform Resource Locator)	Content
National Center for Biotechnology Information	http://www.ncbi.nlm.nih.gov/	databases/analysis/software
Protein Identification Resource	http://www-nbrf.georgetown.edu/	protein sequence database
Protein Data Bank	http://www.rcsb.org/pdb/	3D mol' structure database
Molecules To Go	http://molbio.info.nih.gov/cgi-bin/pdb/	3D protein/nuc' visualization
IUBIO Biology Archive	http://iubio.bio.indiana.edu/	database/software archive
Univ. of Montreal Evolutionary Genomics	http://megasun.bch.umontreal.ca/	database/software archive
Japan's GenomeNet Server	http://www.genome.ad.jp/	databases/analysis/software
European Molecular Biology Laboratory	http://www.embl-heidelberg.de/	databases/analysis/software
European Bioinformatics Institute	http://www.ebi.ac.uk/	databases/analysis/software
The Sanger Institute	http://www.sanger.ac.uk/	databases/analysis/software
Swiss Expert Protein Analysis System	http://www.expasy.org/	databases/analysis/software
Stanford Genomic Resource	http://genome-www.stanford.edu/	various genome projects
J. Craig Venter Institute	http://www.tigr.org/	microbial genome projects
HIV Sequence Database	http://www.hiv.lanl.gov/	HIV epidemeology seq' DB
Ribosomal Database Project	http://rdp.cme.msu.edu/	databases/analysis/software
PUMA2 at Argonne National Laboratory	http://compbio.mcs.anl.gov/puma2/cgi-bin/index.cgi	metabolic reconstructions

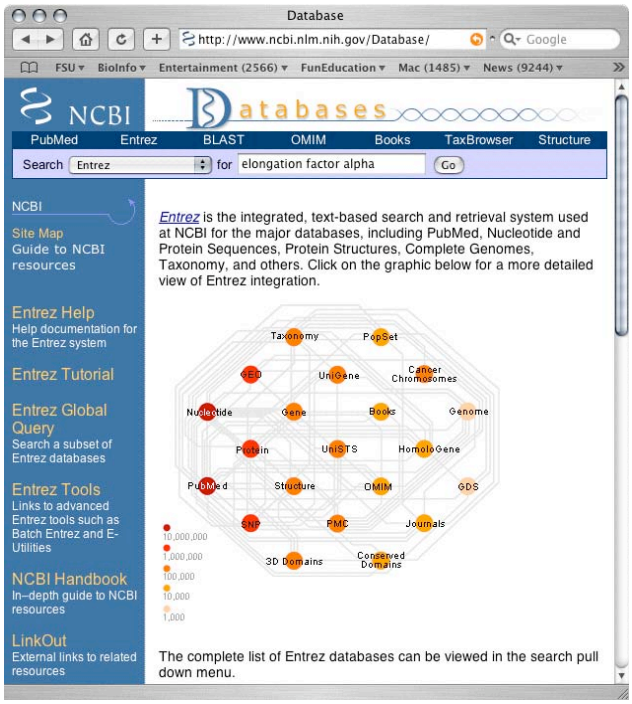
First go to NCBI's site, **http://www.ncbi.nlm.nih.gov/**. Many, perhaps all of you, have been to this site in the past. It is one of the busiest non-pornography World Wide Web sites in the world. Notice that the home page presents a text based database search right up front. This is actually NCBI's Entrez database search system.

Entrez is a very powerful relational style database access tool available from most of NCBI's Web pages. It provides two very important concepts that make it quite powerful — database 'linking' and entry 'neighboring.' Linking allows you to find corresponding entries in other databases, for instance finding a nucleotide entry links to the translated protein entries and the protein entries link to any solved structures and all entries link to references in PubMed, the online access to MedLine, a database of scientific journal articles maintained by the National Library of Medicine. Entrez links over thirty different databases together — ranging from molecular sequences and structures to literature references and disease associations to evolution and

population structure. Neighboring allows you find entries similar to the entry originally found. These similarities are based on precompiled database searches performed with the BLAST algorithm (way more on BLAST and other sequence similarity searching algorithms in the weeks to come!) for sequences, and on precompiled lists based on shared keywords for reference databases.

The NCBI home page follows below on the left in a screen snapshot; your display should look similar:





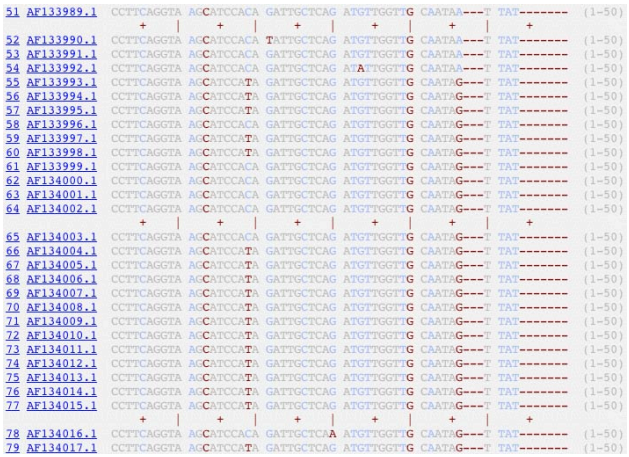
Click on the "**Molecular databases**" link in the sidebar to see all of the databases that Entrez interlinks. Type a general description of your choice from the course project molecule list into the "**for**" text box and then press "**Go**." Remember that my examples will all use elongation factor 1 alpha, but that you are to use the same project molecule from the list on page 8/9 throughout the tutorial series. My "NCBI Databases" browser window looks like the following graphic, opposite-right:

The result lists the number of entries in each of the Entrez databases that contain all of the words that you typed into the "for" box anywhere in the entries' annotation. Because the molecular systems that we are using as examples are very well studied, and because of all the redundancy in the databases, the list will likely be huge; mine includes almost 50,000 nucleotide entries alone! This is often a problem with initial Entrez searches — you get way too many sequences.

My "Search across databases" results follow in the screen snapshot at the left on the following page:

The links listed from "**PubMed**," "**OMIM**," "**Nucleotide**," "**Protein**," and "**Structure**" should be fairly obvious, others may not be quite so easy to understand. The "**Genome**" database can be particularly confusing because of conflicting usage of the word "complete." Check out the various database links, using your **browser back button** to return to the "**cross-database search**" result page after exploring each database a bit.

Notice the "**PopSet"** database; it can be very helpful for comparative studies at the population, sub-species, or closely related species level. The database consists of datasets, many of them fully aligned, used by researchers working on population genetics and speciation questions. A portion of one of the PopSet entries for EF1α is shown at the top of the opposite column on the right:

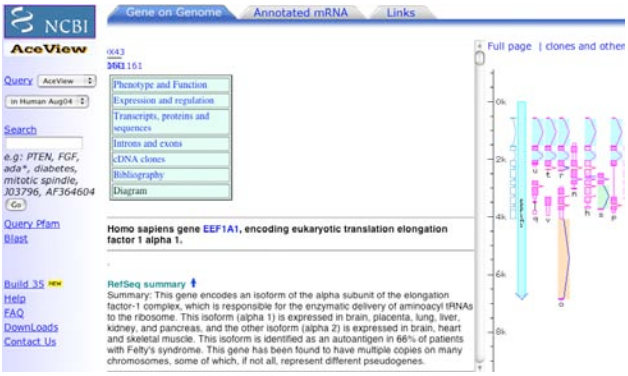


The Entrez "Gene" database is incredibly informative. This database completely supercedes their LocusLink database as of March 2005, and contains all of the previous database's information and more. Pick the link that goes to the "**Gene**" database on your search page. The third hit in my Gene report is one of the two human EF1α paralogues. Select relevant entries in your search to see their Gene records. Click on the entries and read through their complete "Gene" records. The information and links in this record are amazing, providing a Web hyperlink portal to a whole slew of other databases.

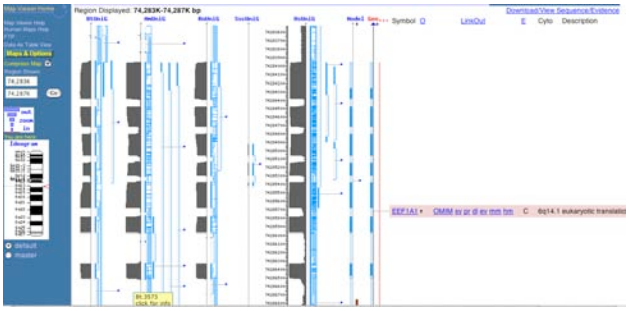
One of your major objectives at this point should be to learn as much about your particular molecular system as you can — it'll make all subsequent work easier. This is especially true if you are dealing with a biological molecular system that pertains to your own academic research. The more background knowledge you have, the more success you'll have understanding the analyses. These databases can go a long way toward achieving that knowledge. My EF1α "Gene" entry is shown in the screenshot displayed on the following page, top left:

One of the more important features of most Entrez report pages can be somewhat inconspicuous. Notice the "**Links**" list on the far right side. Other pages merely present a "Links" button. Each link in the menu will get you directly to the corresponding entry in its respective database. What a great way to gather data related to anything at NCBI!

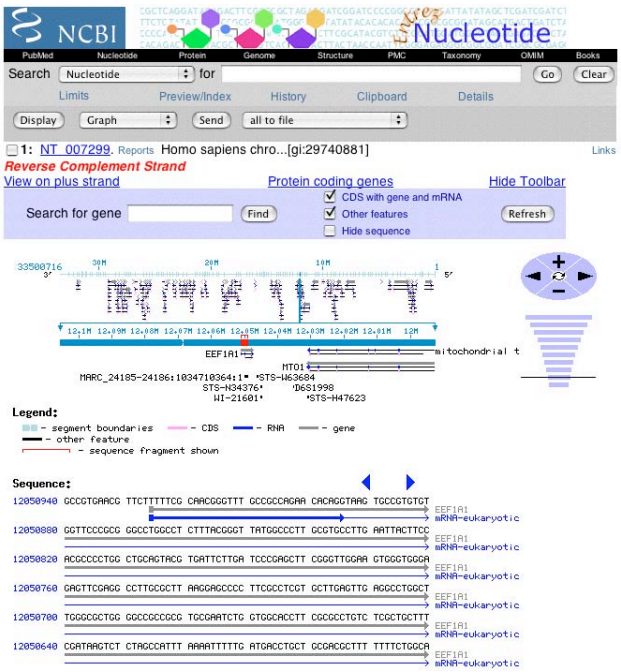
Pick the link that takes you to "**AceView**." This relatively new and very interesting system isn't available for all NCBI data, but when it's there it integrates a map viewer and an annotation summary. Unlike other genome analysis systems, AceView transcripts are not predictions; they are created from real EST and mRNA data aligned to the appropriate genome. Alternative splicing products are, therefore, explicitly defined. Take a few moments to explore the links in the "AceView" display, if there is one for your molecule. A screenshot of the top portion of my "AceView" EF1α report is shown at the top of the opposite column:



Use your 'back' button to return to your lab project "**Gene**" entry, and then use the "**Links**" menu "**Map Viewer**" choice to generate a view of NCBI's genome browser map centered about your gene. The Entrez Genome Map Viewer presents the chromosomal context of the gene compared to other model organism homologues. NCBI's Map Viewer display of EF1α is shown below:



Explore the links on your "**Map Viewer**" page. Clicking on your gene name and then picking "**sv**" from the resultant list will load a nice 'sequence view' graphic of the gene along with its flanking 3' and 5' regions. This representation is shown at top-left on the following page:



This portrayal can be changed to "Display" "GenBank" format (or any other desired format from the drop-down menu) if desired, and downloaded to your computer with the "Send to" "File" buttons, if needed. But, most sequences will already be available locally in the GCG system!

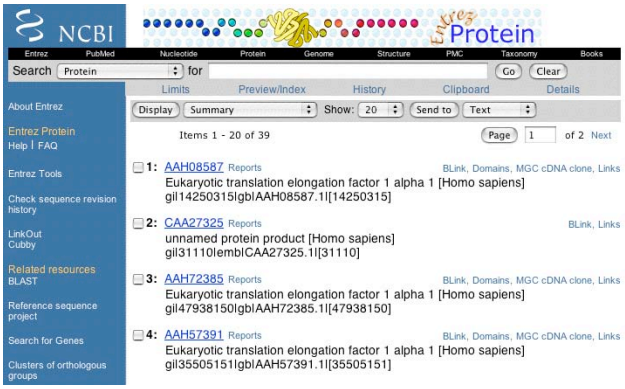
Obviously, when using Entrez, there are many different routes to the same data, and it doesn't really matter which route one takes. Some routes will be information rich, others terser. The point is all the interconnectedness of the data. Unfortunately, in some ways, the huge amount of data, remember I found almost 50.000 nucleotide entries with my preliminary search, can make finding a particular sequence entry challenging. What can you do if you aren't able to find the 'tree for the forest?'

NCBI provides a powerful way to restrict your search beyond just restricting it to particular databases, because of this problem. Press the "**Limits**" button just below the "**Search for**" text boxes on most Entrez pages. This function allows you to restrict the search in several different ways, the "Field," "Exclude," and "Limit" menus perhaps being the most helpful. Notice the default is "All Fields;" in other words, your desired text can appear anywhere in the entry's annotation. Available fields will depend on the nature of the report you have loaded. "Gene" reports have a handy "Limit by Taxonomy" menu.

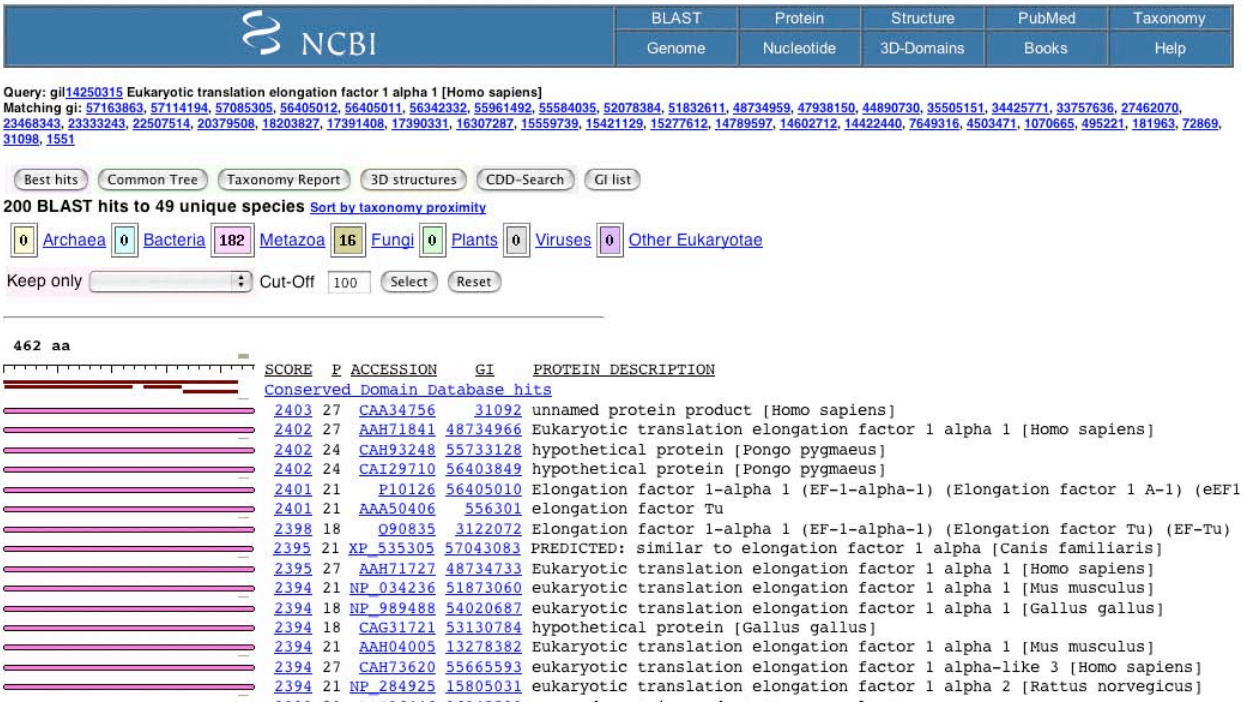
Load an entry from the "**Nucleotide**" database, click on "**Limits**," and then switch "**All Field**" to "**Title**." Also check "**exclude**" for all of the available categories, and then press "**Go**" again. Now the search is restricted to only those entries whose "Title" line contains your text and also excludes some problematic subsets of the Entrez nucleotide database. "Title" is also known as "Definition" and is a one-line description of the entry that usually contains the type of English words that would normally be used to name the entry. The "Protein Name," "Gene Name," and especially "Keyword" Fields are often problematic due to naming discrepancies. In my case no entries were found at all if I restricted the "Field" to either "Protein" or "Gene" "Name" using the text "elongation factor alpha," yet the "Title" search found just almost 7,000, a lot better than 50,000! But that's still a ton.

If you need to restrict your search to a particular 'critter,' the "History" function is great. Repeat your "**Limited to**" search, only this time restrict the search to your "**Organism**" of interest, typing the organism's proper genus name in the "**for**" field. You'll find all of the entries from that organism in that database. I found over a million sequences in the Nucleotide database from "Homo." The trick is to use that "**History**" function to link the two searches together. Note that you need to use the pound sign (#) before the search numbers and that Boolean operators need to be capitalized. "Limits." "Preview/Index." and "History" all make Entrez extremely powerful.

Pick "**Protein**" from a "**Links**" menu to load your project protein entries. You'll see the protein sequences that correspond to your project; mine is shown adjacent. Protein entries in Entrez have an additional link that is incredibly helpful for comparative studies. That link is "BLink" — precompiled BLAST similarity search results for each protein. Go ahead and select one of your entry's "**BLink**" buttons.



The resulting page has a wealth of information; all based on precompiled BLAST database searches against NCBI's "Protein" database. You'll see a graphical representation of the most similar protein sequences aligned to the particular query sequence that you selected. By default the order is determined by BLAST score similarity ranking; however, an often-helpful feature is to "Sort by taxonomy proximity." Taxonomic groupings are color-coded and alignments are graphically represented. My example follows below:



The "**SCORE**" column leads you to pairwise alignments and the "**ACCESSION**" column gets you directly to each homologue. Explore the links for a while, but not too long, there's lots more to do this afternoon. As with most Web stuff, it is very easy to get sidetracked and end up spending way more time than you intended!

Entrez sequence entries can be saved in their default GenBank/GenPept format (or other desired formats from the drop-down menu) to your computer by specifying "Send to" "File" (but, as you'll see later today, most sequences are available locally in the GCG system so there is seldom need to download sequences), or they can be copied and pasted into other applications.

For further online assistance be sure to visit and read over NCBI's extensive Entrez Help page at **http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helpentrez** for much more information on maximizing its power, and check out their tutorial at **http://www.ncbi.nlm.nih.gov/Entrez/tutor.html** for a guided tour through the Entrez system. Especially investigate the "Limits," "History," and "Preview/Index" functions further. And NCBI's human genome guide, **http://www.ncbi.nlm.nih.gov/genome/guide/human/** provides a fantastic starting point for human genomics research. Particularly note "**Comparative Genomics**" near the bottom of the page where you can find "**Homologene**" and the "**Homology Map resources**."

For the lab report find a human homologue in your project molecule system and tell me what human disease states, if any, that particular gene affects by finding the relevant Online Mendelian Inheritance in Man (**OMIM**) link. If you've chosen RuBisCO, there'll be no human homologue, but use any relevant links to tell me if there are any plant disease states associated with it. Also tell me what the chromosomal location(s) of the gene(s) is(are), human or *Arabidopsis*, as the case may be. Please do this assignment outside of assigned lab time.

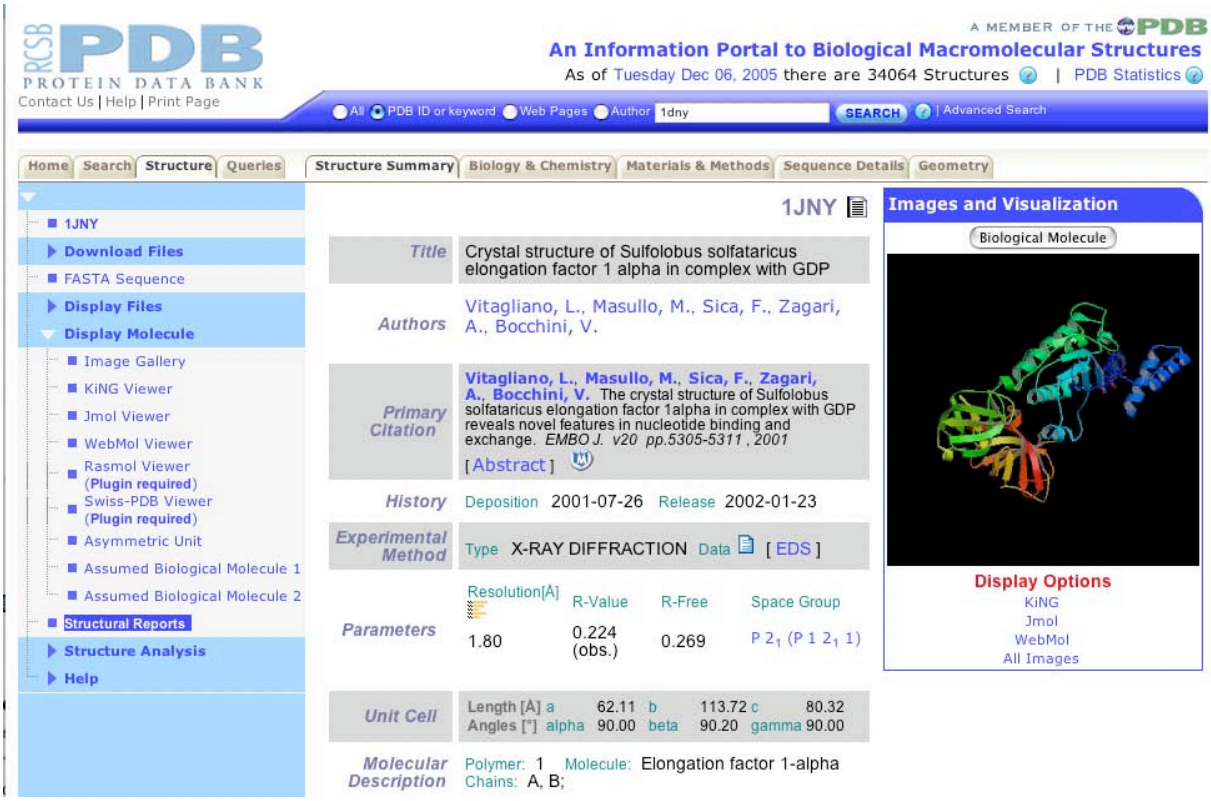
Next, let's move on to 3D structure databases available on the Web. Return to NCBI if you've left their site. Select one of your project choice protein entries from a relevant Entrez page, or use the main "cross-database search" result page; either way pick a "**Structure**" link to find solved structures. My initial Entrez search found sixteen. As in most text based searching, they may not all be relevant, but they all have the search terms you used somewhere in their annotation. I'll pick on entry "1JNY," EF1α from *Sulfolobus solfataricus*. Click your entry's name to go to its "MMDB Structure Summary." Many objects in the new schematic are links; explore them a bit. A graphic representation of a structural alignment can be shown in a "VAST Structure Neighbors" window. You can generate actual alignments from that page, or you can view the 3D structure with NCBI's Cn3D structure viewer (from most MMDB links). You should be able to visualize this structure without having to download any software since Cn3D has been installed on the SCS common use computers and on the Conradi Biology teaching lab computers.

Click on the graphic of the structure or on the "**View 3D Structure**" button to download the structure and visualize it with Cn3D. You may have to configure your Web browser giving it the proper mime type (chemical/ncbi-asn1-binary) and executable path (/usr/common/i686-linux/Cn3D) the first time you use Cn3D to open the file, or you can manually open it. We'll discuss the particulars of this in lab. Cn3D will launch with your structure in its window. My 1JNY Cn3D display is shown to the right in the adjacent screen snapshot:

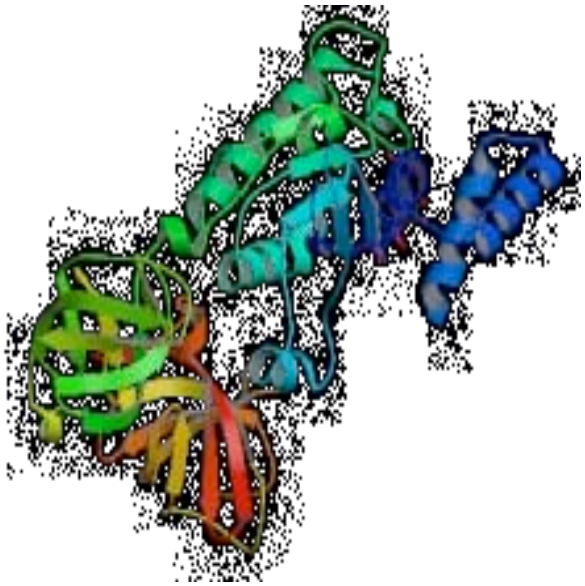
This particular structure consists of two chains — you should be able to see the symmetry. Your choice may have one or more chains; it doesn't matter. The default graphic representation is called "Worms," several others are available under the "Style" "Rendering Shortcuts" menu. Also notice that the default coloring is based on secondary structure, but can also be changed to other schemes, if desired. You can 'grab' the molecule at any point and rotate it in space. You can increase the molecule's size by zooming in. There's complete help on all the "Commands" as well as Web help from NCBI. A powerful Cn3D feature is its ability to display more than one molecule and their corresponding sequence alignment. Check it out!

The Protein Data Bank is the major biological molecular structure database in the world. Connect to their Web site, **http://www.rcsb.org/pdb/**. Links to deposit or download data, to software tools, and for general browsing are all available. PDB provides a search query form on their home page. You can use a PDB access code found previously at NCBI, or, if you don't know the code, you can use general search terms. Check out their help pages and their tutorial for more information. The PDB home page is shown below:

Enter the same PDB code that you used previously at NCBI and press "**SEARCH**." As with other biological molecular databases, searches can be performed in more sophisticated manners, but we'll just use the defaults this time. Since you directly searched for a PDB accession code, the "Structure Explorer" "Structure Summary" page will be presented, otherwise a list of candidates will be seen. However, you need to be very careful with the list approach, as not all results may be the molecule you are looking for. This is always a concern with general text searches. My search for 1JNY produced the following window:

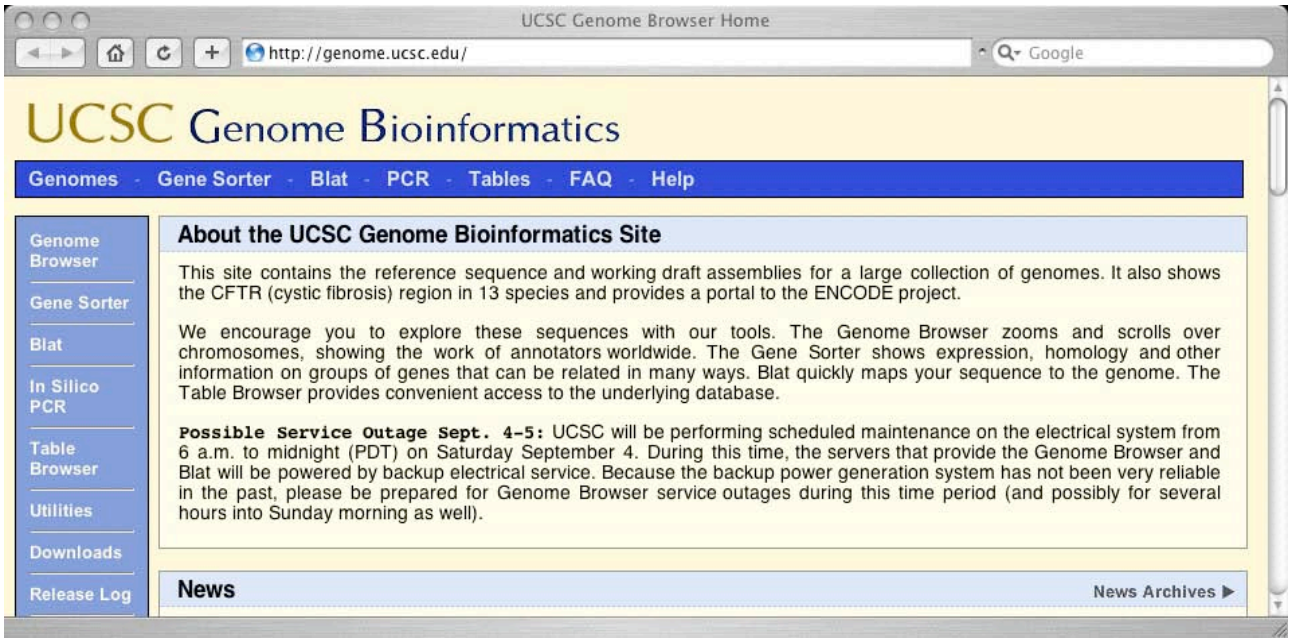


"**Display Options**" lists several Java-driven interactive, three-dimensional representations, as well as a few still image. Explore some of the options, including the "**KING**" viewer. All of these interactive viewers require Java. Opening up the "**Display Molecule**" menu shows other viewers: Rasmol and Swiss-PDB Viewer. You can try these as well. You may be asked what to do with the PDB file by your browser, as it did with NCBI's Cn3D file, if this is the first PDB format file you've ever downloaded on the particular machine that you are using. We'll explore in lab. One of my still image examples is shown here to the right:



We will explore three-dimensional structure further later on in the course, so I'm not going to discuss any more details related to 3D visualization and modelling. Plus we have a bunch more to do today!

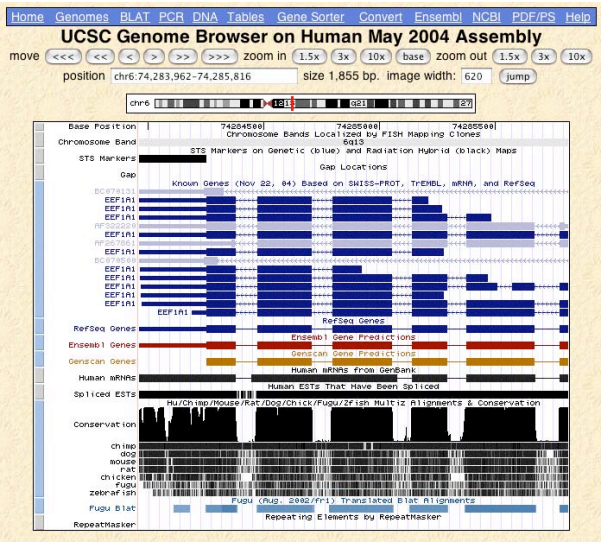
However, before moving on to the way that things are done in the Wisconsin Package, let's briefly check out some of the other biological molecular database resources on the Web. For a genome mapping perspective connect to the University of California, Santa Cruz, bioinformatics group's excellent human genome browser, **http://genome.ucsc.edu/**. The welcome page explains the project and is shown in the screenshot below:



Searching for "elongation factor 1 alpha" with their "**Genome Browser**" and following the relevant link produces the genome map shown here opposite, below to the right.

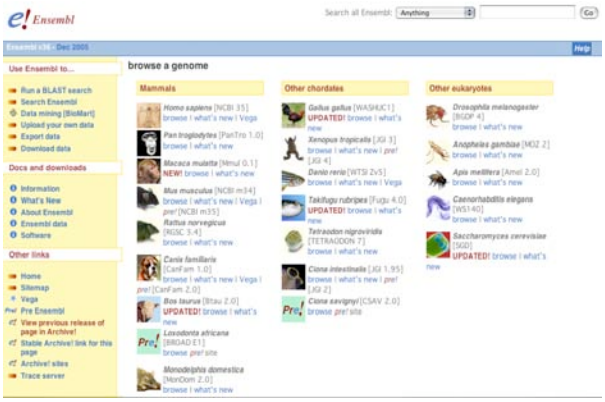
Every object on this map is a link, and there's a whole lot of them. Start clicking away; it's very impressive! Begin by clicking on a gene name. You'll get a page that summarizes present knowledge about the particular molecular system selected, including expression profiles and 3D structure cartoons, if available. Zoom all the way in to individual bases and amino acids, and all the way out to the entire chromosome, with the scale factor buttons near the top. Scroll along the chromosome with the "move" buttons.

In particular, note the lower third of the display is comparative genomics — many homologous loci all aligned to each other. Click anywhere within the similarity profile to get to the actual alignment.

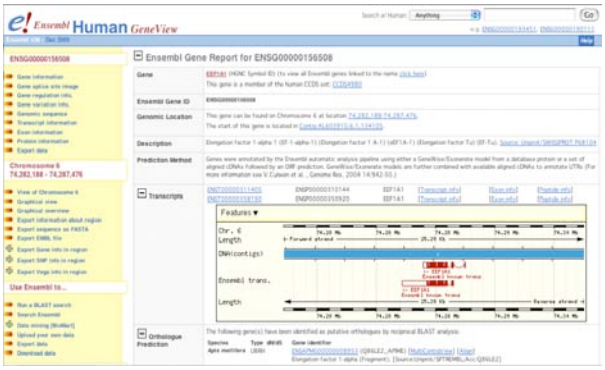


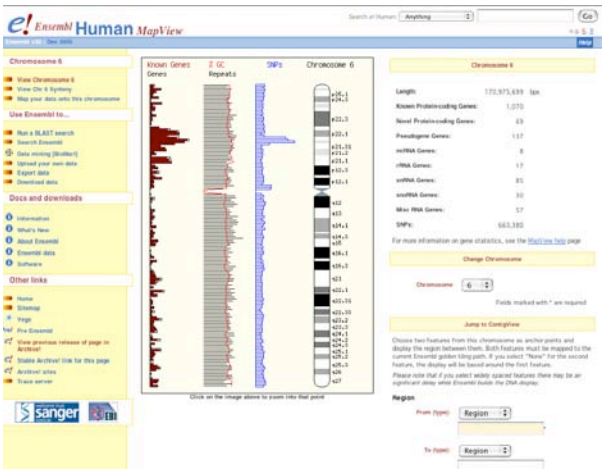
The Ensembl project, **http://www.ensembl.org/**, jointly hosted by the Welcome Trust Sanger Institute and the European Bioinformatics Institute, has built a comparable system in the United Kingdom. It is incredibly powerful.

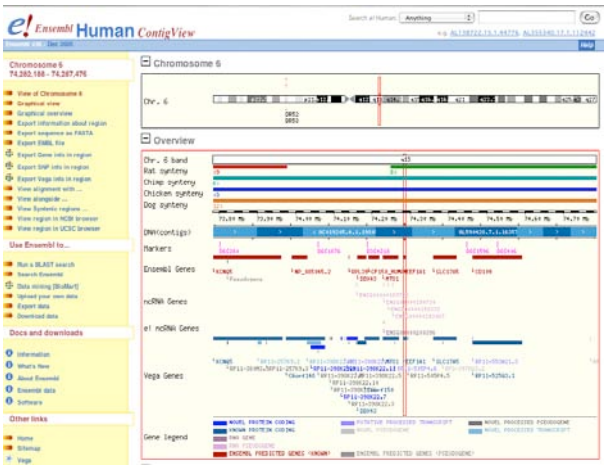
The Main Ensembl home page should look similar to the screenshot shown directly below:



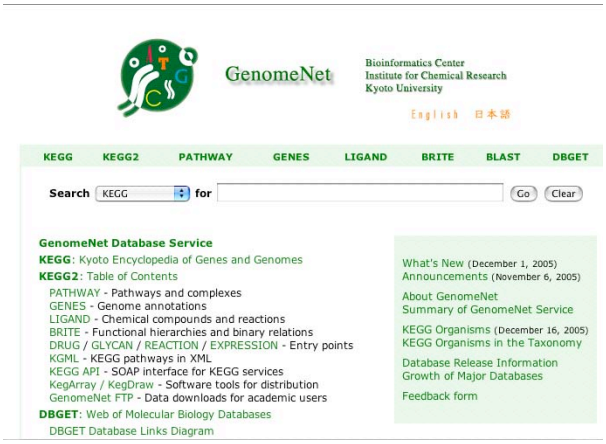
Searching for “elongation AND factor AND alpha” on their system found a slew of entries. Note that the Ensembl search engine requires the use capitalized Boolean logic connectors between search terms. Following the human entry for “Gene ID” that corresponded to “EF1A1” produced the “TextView” screen snapshot shown at top opposite. Several other Ensembl “–Views” help with the visualization: “ContigView,” “MapView,” “AlignSliceView,” “MultiContigView,” “CytoView,” and “FamilyView.” “FamilyView” even offers a Java powerd “JalView” multiple sequence alignment editor! I’ve included a couple of these other “–Views” screen snapshots in the right column, middle and bottom:



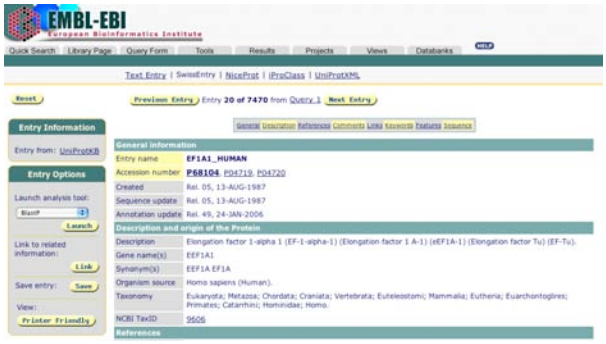




The GenomeNet, **http://www.genome.jp/**, at Kyoto University, Japan is another popular route, a good ‘one-stop-spot’ to genomic data, especially if you’re based in Asia. Their home page is shown below on the following page, top-left:



query form right up front, just like NCBI. However, it doesn't use Entrez; it uses SRS instead. See if you can repeat the search that you preformed at NCBI to see how the systems compare. My eventual EF1A1 protein results follow below in the next screen snapshot:



Finally, go to the European Bioinformatics Institute (EBI) at **http://www.ebi.ac.uk/** to see their Web-based Sequence Retrieval System (SRS) (Etzold and Argos, 1993). The EBI home page has a search

You should discover about the same number of sequences this way as you did at NCBI. Similar field restrictions as NCBI's are also available, though I think they are a bit more difficult to use.

That's probably enough time to spend on Web access to the databases for now, but don't shut down your browser — we'll need it for one more step below. Please explore further on your own time. There is still one more very important access route that we need to investigate today.

HPC — FSU's High Performance Cluster: biocomputing services and the Wisconsin Package

Launch an X-terminal program window with the appropriate icon from the desktop or from one of the menus. Regardless of what local machine you are using, you need an X-tunneled ssh session to display X windows on your local machine. On Linux machines type the following command into your terminal window (the -x has to be capitalized and replace "user" with your new HPC account name, from last week):

```
$ ssh –X user@submit.hpc.fsu.edu
```

The "**-x**" option is necessary to allow 'X-tunneling' and set up your X environment. This is the only encrypted, secure way to make X connections, and is required by HPC, if you want to use any resources that require X windows. If this is the first time that you've connected to HPC from your local machine, you'll likely get something similar to the following screen trace, asking if you really want to do what you're doing. Answer the question with "**yes**" spelled all the way out:

```
The authenticity of host 'submit.hpc.fsu.edu (144.174.80.99)' can't be
established.
RSA key fingerprint is 99:6d:e4:a3:0e:16:23:27:eb:31:1f:3d:91:af:8a:cc.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'submit.hpc.fsu.edu' (RSA) to the list of known hosts.
```

Next you'll be asked for your HPC account password. You assigned this to yourself last week! Note that passwords are not displayed in terminal windows as you type them. The OS checks your username and password, and if correct, runs your default shell program and any startup scripts, and then returns the system prompt. This user name and password is separate and distinct from the one earlier used to get on the SCS Computer Classroom system, though you are welcome to use the same password for both.

On pre-OS X Macs or MS Windows machines find and use the appropriate icons to launch an X11 emulator and ssh, either on the desktop or in menus. After ssh opens, use or build the appropriate menu command to connect to submit.hpc.fsu.edu. The X tunneling ssh option needs to be turned on. Further details of X11 will not be covered in this tutorial. There are too many variables depending on your local machine. If this isn't enough, ask me for further assistance. I'm also available for personal help in your own laboratories. If you are having trouble connecting to and using HPC from there, just contact me at **stevet@bio.fsu.edu**.

We need to make a few crucial changes to our new HPC accounts in order to easily use the bioinformatics tools, and, especially, the GCG package on HPC. Return to your Web browser and go to the HPC Software page. Click on "**Important information about this software category**" under the "**Bioinformatics Suite**" heading to see what I'm talking about. Remember the "**nano**" editor and those funny 'dot' files from last week? Well, we need to use an editor now to modify our "**.bashrc**" file. Issue the command "**nano .bashrc**" and your "**.bashrc**" file will appear in your terminal window. This file helps to set up your bash shell user environment. Remember, "**nano**" doesn't know anything about the mouse. You have to use the keyboard arrow keys to move around. Move your cursor to the bottom of the file with the < down arrow > key, open up an empty line with the < return/enter > key, and type the following line there:

```
. /usr/local/profile.d/bio.sh
```

This step will automatically initialize the GCG package and set up all of the "$PATH"'s to all of the bioinformatics tools on HPC every time that you log in. After making this modification issue the following command to run your new ".bashrc" file (or logout from HPC and log back in). You won't have to do this on subsequent logins, as ".bashrc" is sourced by default every time you log in:

```
$ source .bashrc
```

You'll see a screen trace similar to the following after sourcing your new ".bashrc" file:

```
                    Welcome to GCG
                  Version 11.1.3-UNIX
                  Installed on linux

          Copyright (c) 1982 - 2008, Accelrys Inc.
                   All rights reserved.

    Published research assisted by this software should cite:
        GCG Version 11.1, Accelrys Inc., San Diego, CA
Databases available:
        GenBank              Release  164.0   (02/2008)
        GenPept              Release  164.0   (02/2008)
```

```
Refseq                  Release  27.0    (01/2008)
UniProt                 Release  12.8    (02/2008)
PROSITE                 Release  20.26   (02/2008)
Pfam                    Release  22.00   (06/2007)
Restriction Enzymes (REBASE)     802     (02/2008)


Technical support: contact Steve Thompson, stevet@bio.fsu.edu
      Online help: $ genhelp or http://gcg.scs.fsu.edu/
```

The screen trace shows the version numbers for the Accelrys GCG Package and of all its online databases that are installed on the HPC, as well as how to get help. Now it's time to see the primary sequence analysis bioinformatics package used in this course. That package is the Accelrys GCG Wisconsin Package.

The Genetics Computer Group

The Wisconsin Package for Sequence Analysis began as a service project in 1982 in Oliver Smithies' lab in the Genetics Department at the University of Wisconsin, Madison. It spun off that effort into a University Research Park location becoming an independent private company, the Genetics Computer Group (GCG), in 1990. Then in 1997 the Oxford Molecular Group of Great Britain, a chemical informatics company, acquired GCG. The drug discovery and development firm Pharmacopeia next purchased GCG, and the other Oxford Molecular holdings, in late 2000. In the summer of 2001, it, along with Pharmacopeia's other software holdings, were all placed under the new corporate name Accelrys, Inc., which became a subsidiary of Pharmacopeia. And then in Spring 2004 Pharmacopeia formally separated from Accelrys. Finally, in June 2008, the legacy was killed. Accelrys decided GCG was not making the company enough money and 'retired' the product to concentrate on their drug-design holdings. For more information on these issues see http://www.PetitionOnline.com/gcg/petition.html and http://bio.fsu.edu/~stevet/GCG.death.pdf.

The Wisconsin Package became the global 'industry-standard' in sequence analysis software within the twenty+ years of its existence. It provides a comprehensive suite of over 150 integrated DNA and protein analysis programs, from database, pattern, and motif searching; fragment assembly; mapping; and sequence comparison; to gene finding; protein and evolutionary analysis; primer selection; and DNA and RNA secondary structure prediction. Unfortunately almost one third of these programs no longer work in the HPC's Cent OS 5 environment, but the remainder work fine. The package's programs work together in a "toolbox" fashion. Much like a carpenter's toolbox where using the right tool correctly in the right order can build a house; several relatively simple programs properly used in succession can lead to sophisticated sequence analysis results with the Wisconsin Package. Furthermore, the programs are 'internally compatible.' By this I mean that once you learn how to use one program, how the programs 'look and feel,' then you pretty much know how to run all of the programs, since all 'act' similarly, and, most importantly, the output from many programs can be used as input for other programs. This is how you use the programs in a logical succession. The Wisconsin Package only runs on computers running the UNIX operating system, but it can be accessed from any networked terminal. The package has been used for over 20 years at more than 950 institutions worldwide; even with its 'retirement,' learning it here may prove useful at other institutions as well.

SeqLab is included as a part of the GCG Wisconsin Package standard distribution. This powerful X Windows based GUI is a 'front-end' to the package. It provides an intuitive alternative to the UNIX command line by allowing menu-driven access to most of GCG's programs. SeqLab makes running the Wisconsin Package much easier by providing a common editing interface from which most programs can be launched and alignments can be manipulated. SeqLab originated way before it had anything to do with GCG. Steve Smith was working on bacterial ribosomal RNA phylogenies with Walter Gilbert and Carl Woese. Steve realized the vital need for a comprehensive multiple sequence editor. Nothing existed at the time that satisfied him, so he invented one. In addition to providing the vital editing function, it also served as a menuing system to external functions such as the PHYLIP molecular evolution programs and Clustal alignments. He called it the "Genetic Data Environment: GDE" (1994). Many people were very impressed and he made it freely available. Coincidentally GCG realized the need for some sort of a 'point-and-click' environment for their system. They were losing business, only being able to provide a command line interface. Therefore, they started trying to develop a GUI for the Wisconsin Package and released it in 1994. They called it the "Wisconsin Package Interface," WPI for short. Few were impressed. It only provided a menu to their programs, hardly anything more than the "-check" command line option they've always had. So they did a natural and very smart thing. They hired Steve Smith away from Millipore, where he had recently moved, into their company, so that he could merge his GDE with their WPI. The late 1996 offspring, without the acronyms (GDE + WPI = SeqLab) was SeqLab. Even though it's far from perfect, once you gain an appreciation for SeqLab's power and ease of use, I don't think you'll be satisfied with any other sequence analysis system. I know I'm not.

Using the Wisconsin Package — specifying sequences and logical terms!

Before launching into SeqLab, let's go over an important, central 'idea' of the Wisconsin Package. One of the most difficult aspects of the Package for new users to get used to is how to tell the programs what sequences you want to work with. GCG calls this "specifying sequences" and it's crucial to understanding the way their programs work. Once you've become comfortable with these concepts, so many of the frustrations commonly encountered with the Package will disappear. So, to answer the always perplexing GCG question "What sequence(s)?" . . . the four ways of specifying sequences, in order of increasing power and complexity:

1. The sequence is in a local GCG single sequence format (SSF) file in your UNIX account. This file can be anywhere in your account as long as you supply an appropriate path such that a program can find the file. The file can have any name but it is best to use extensions that tell you what type of molecule it is, e.g. "`.seq`" and "`.pep`" (e.g. "`my.pep`" or "`~/subdir/my.seg`"). The GCG program SeqConv+ is available for specific data format conversions, and SeqLab's Editor Mode can directly "Import" native GenBank, FastA, and ABI style trace format files. Probably the simplest format is Pearson's FastA format (Pearson and Lipmann, 1988). It allows one line of annotation, after a required 'greater than' (>) and name:

```
>EXAMPLE  example Pearson FastA format - documentation one line only
ACTGACGTCACATACTGGGACTGAGATTTACCGAGTTATACACAGATTTAATAGCATGCGATCCCATGGGA
```

The SeqConv+ program can change it into a GCG format single sequence file:

```
!!NA_SEQUENCE 1.0

EXAMPLE  Length: 71  October 23, 2008 09:26  Type: N  Check: 6762  ..

       1  ACTGACGTCA CATACTGGGA CTGAGATTTA CCGAGTTATA CACAGATTTA

      51  ATAGCATGCG ATCCCATGGG A
```

2. The sequence is in a local GCG database in which case you 'point' to it by using any of the GCG database logical names. These names make sense and are either the name of the database or an abbreviation thereof. Subcategory logical names based on GenBank divisions can be used for nucleotide databases, such as Bacterial. A colon (**:**) always sets the logical name apart from either an accession code or a proper identifier name or a wildcard expression, and unlike the rest of UNIX, the whole thing is case insensitive. Several examples follow: GenBank:EctufBT, gb:x57091, SwissProt:EFTu_Ecoli, and uni:p02990 all refer to the elongation factor Tu in *E. coli*, DNA in the first two, protein in the last two. If you know that the database uses consistent naming conventions, then you can use a wild card to specify all of a particular type of sequence. This works particularly well in SwissProt because of its consistent naming conventions; e.g. SW:EFTu_* specifies all of the EF-Tu sequences in SwissProt, SW:*_Ecoli specifies all of the *E. coli* sequences in SwissProt.

Because most sequences are available in the local GCG databases, it is seldom necessary to put individual sequences in your account. Exceptions include sequences that you modify, such as making an alignment or engineering a vector, and very new sequences that are not in the GCG databases. There's no need to fill your account with data available on the same server outside your account.

Logical terms for the GCG Wisconsin Package at FSU

Sequence databases, nucleic acids:

GENBANKPLUS:*	all of GenBank plus EST, HTC, and GSS	PHAGE:*	GenBank phage
GBP:*	all of GenBank plus EST, HTC, and GSS	PH:*	GenBank phage
GENBANK:*	all of GenBank except EST, HTC, and GSS	PLANT:*	GenBank plant
GB:*	all of GenBank except EST, HTC, and GSS	PL:*	GenBank plant
BACTERIAL:*	GenBank bacterial	PRIMATE:*	GenBank primate
BA:*	GenBank bacterial	PR:*	GenBank primate
EST:*	GenBank EST (Expressed Sequence Tags)	RODENT:*	GenBank rodent
GSS:*	GenBank GSS (Genome Survey Sequences)	RO:*	GenBank rodent
HTC:*	GenBank High Throughput cDNA	STS:*	GenBank (sequence tagged sites)
HTG:*	GenBank High Throughput Genomic	SYNTHETIC:*	GenBank synthetic
INVERTEBRATE:*	GenBank invertebrate	SY:*	GenBank synthetic
IN:*	GenBank invertebrate	TAGS:*	GenBank EST, HTC, and GSS
OTHERMAMMAL:*	GenBank other mammalian	UNANNOTATED:*	GenBank unannotated
OM:*	GenBank other mammalian	UN:*	GenBank unannotated
OTHERVERTEBRATE:*	GenBank other vertebrate	VIRAL:*	GenBank viral subdivision
OV:*	GenBank other vertebrate	VI:*	GenBank viral subdivision
PATENT:*	GenBank patent	REFSEQNUC:*	NCBI RefSeq transcriptomes
PAT:*	GenBank patent	RS_RNA:*	NCBI RefSeq transcriiptomes

Sequence databases, amino acids:

UNIPROT:*	all of Swiss-Prot and all of SPTREMBL	SPTREMBL:*	Swiss-Prot preliminary EMBL translations
UNI:*	all of Swiss-Prot and all of SPTREMBL	SPT:*	Swiss-Prot preliminary EMBL translations
SWISSPROTPLUS:*	all of Swiss-Prot and all of SPTREMBL	GENPEPT:*	all of GenBank's CDS translations
SWP:*	all of Swiss-Prot and all of SPTREMBL	GP:*	all of GenBank's CDS translations
SWISSPROT:*	all of Swiss-Prot (fully annotated)	REFSEQPROT:*	NCBI RefSeq proteomes
SWISS:*	all of Swiss-Prot (fully annotated)	RS_PROT: *	NCBI RefSeq proteomes
SW:*	all of Swiss-Prot (fully annotated)		

3. The sequence is in a GCG format multiple sequence file, either an MSF (multiple sequence format) file or an RSF (rich sequence format) file. The difference is that MSF files contain only the sequence names and sequence characters, whereas RSF files contain sequence names and data, plus sequence annotation; i.e. they are "richer." As in GCG single sequence format, it is always best to retain the suggested GCG extensions, "`.msf`" or "`.rsf`," in order for you to easily recognize what type of file they are without having to look, though it is not required and they could just as well be named "`Joe.Blow`."

To specify sequences contained in a GCG multiple sequence file, supply the file name followed by a pair of braces, "`{}`," containing the desired sequence specification. For example, to specify all of the sequences in an alignment of elongation 1α and Tu factors, you could use a naming system like the following: "`ef1a-tu.msf{*}`." Furthermore, one can point to individual members of the alignment or subgroups by specifying their name within the braces, e.g. "`EF1a-Tu.rsf{eftu_ecoli}`" to point just to the *E coli* sequence or "`EF1a-Tu.rsf{eftu_*}`" to point at all of the EFTu's as long as you use a sequence naming convention that retains this convention.

4. Finally, the most powerful method of specifying sequences, and the 'way' that SeqLab works, is in a GCG "list" file. This file can have any name though it is convenient to use the GCG extension "`.list`" to help identify them in your directory. It is merely a list of any other sequence specifications and can even contain other list files within it. List files can be created by hand with an editor and they are produced by many of the GCG programs, such as all the search programs. This is how the output from one program can become input to another. The convention to use a GCG list file in a program is to precede it with an 'at' sign (`@`). Furthermore, you can supply attribute information within list files to specify something special about the sequence. This is especially helpful with length attributes that can restrict an analysis to specific portions of a sequence and are shown in the example below:

```
!!SEQUENCE_LIST 1.0

A GCG list file of EF 1 alpha/Tu sequences.  Five specifications are listed.
Two periods, side-by-side, always separate documentation from data.   ..

my-special.pep       begin:24    end:134
SwissProt:EfTu_Ecoli
Ef1a-Tu.msf{*}
~/some/other/directory/another.rsf{ef1a_*}
@another.list
```

Using SeqLab

Now that you understand something about the way that the Wisconsin Package 'thinks' let's take a look at the SeqLab GUI. But first of all, it would help to keep your files organized by creating and using a new directory for each week's data. Therefore, 'make' a new directory and 'change directory' into it by issuing the following two commands in your active HPC X-tunneled ssh terminal session:

```
$ mkdir lab2
$ cd lab2
```

Next use the GCG command "**fetch**" (without the quotes) to retrieve a sample list file that I have prepared and made available in the GCG system. I've copied some of my elongation factor data to the public GCG directories for your use, and this file points to that data:

```
$ fetch sample.list
```

Display your new "**sample.list**" file:

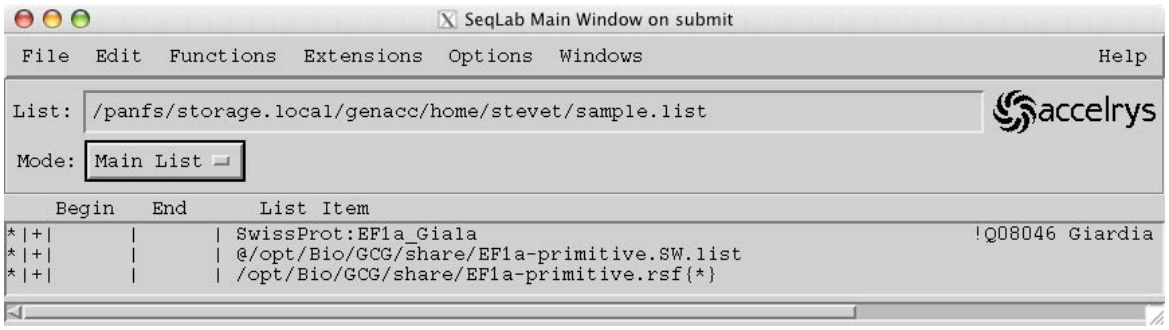
```
$ more sample.list
!!SEQUENCE_LIST 1.0
..
SwissProt:EF1a_Giala Begin:1 End:396 !Q08046 Giardia lamblia Elongation Factor 1-alpha
@/opt/Bio/GCG/share/EF1a-primitive.SW.list
/opt/Bio/GCG/share/EF1a-primitive.rsf{*}
```

There's three different GCG specifications for EF-1α sequences in this sample list file. The first points to *Giardia lamblia* EF-1α from the local GCG SwissProt database; the second is a list file (notice the "@") of EF-1α sequences from several 'primitive' eukaryotes in the local SwissProt database made by the GCG reference searching program LookUp; the third points to a Rich Sequence Format (RSF) file (notice the "{*}") of a multiple sequence alignment of the molecules in that list.

Now issue the command "**seqlab &**" (again, without the quotes) in your terminal window to fire up the SeqLab interface:

```
$ seqlab &
```

The ampersand, "**&**," is not necessary but really helps out by launching SeqLab as a background process so that you retain control of your initial terminal window where you can issue OS commands. This should produce two new windows, the first an introduction with an "**OK**" box; check "**OK**." You should now be in SeqLab's "**Main List**" mode with the empty default list file, "**working.list**," open in the main window. If SeqLab opened in "**Mode: Editor**," switch to "**Main List**." Next, go to the "**File**" menu and select "**Open List . . .**" and then use the "**Open List File**" window to select the "**sample.list**" file from your account. Press "**OK**" to see something similar to the screen snapshot at the top of the following page:



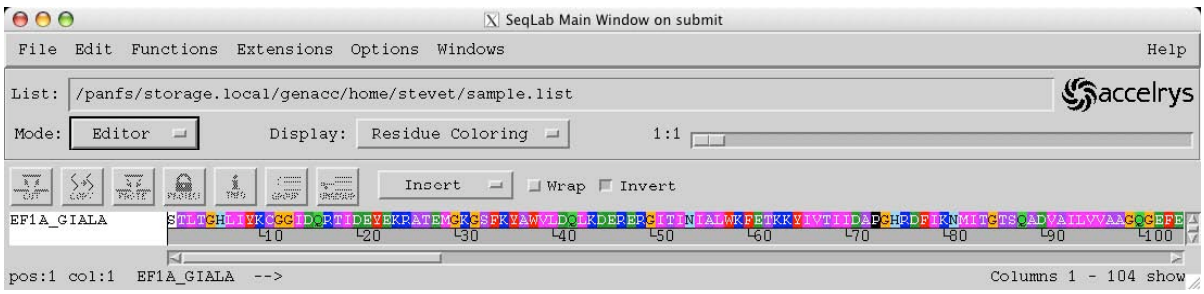
SeqLab Preferences and Help

Before going any further, go to the "**Options**" menu and select "**Preferences . . .**" A few of the options should be checked there to insure that SeqLab runs its most intuitive manner. The defaults are usually fine, but I want you to see what's available. Remember, buttons are turned on when they're pushed in and shaded. First notice that there are three different "**Preferences**" settings that can be changed: "**General**," "**Output**," and "**Fonts**;" start with "**General**." The "**Working Dir . . .**" setting will be the directory from which SeqLab was initially launched. This is where all SeqLab's working files will be stored; it can be changed in your accounts if desired, however, it's fine to leave it as "**`lab2`**" is for now. Be sure that the "**Start SeqLab in:**" choice has "**Main List**" selected and that "**Close the window**" is selected under the "**After I push the "Run" button:**" choice. Next select the "**Output**" Preference. Be sure "**Automatically display new output**" is selected. Finally, take a look at the "**Fonts**" menu. If you are dealing with very large alignments, then picking a smaller Editor font point size may be desirable in order to see more of your alignment on the screen at once. <**Click**> "**OK**" to accept any changes.

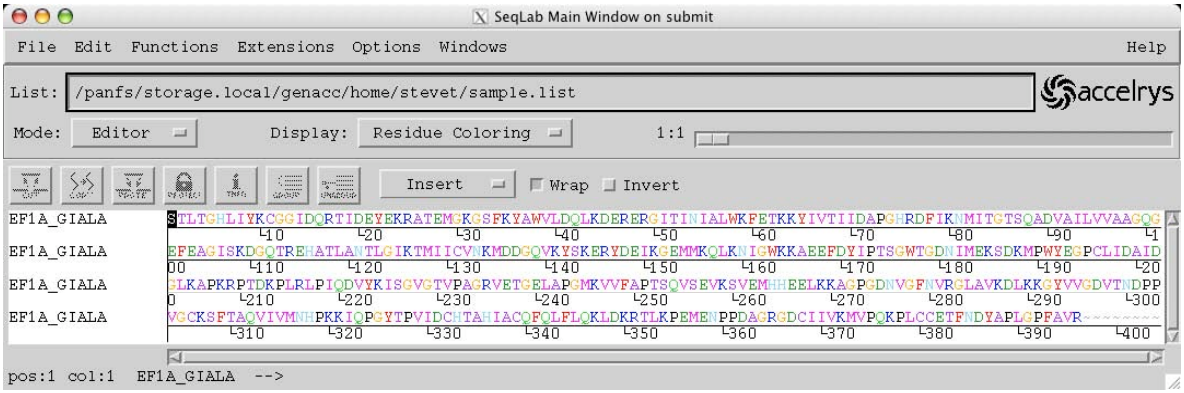
Help is always available when using the Wisconsin Package. Every SeqLab program window and SeqLab's Main window have "Help" buttons. Furthermore, issuing the command "`genhelp`" in a command line terminal window launches a text driven Help system, and, probably best of all, the URL **http://gcg.sc.fsu.edu/** links to a Web version of the Help system. Explore GCG's Help system on your own out-of-lab time.

Exploring SeqLab

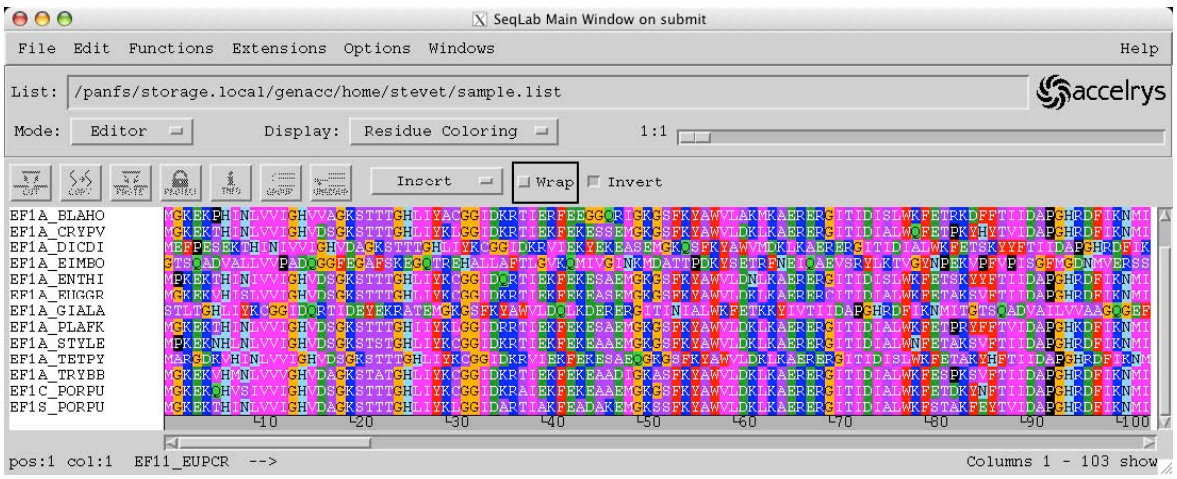
Select the top entry in your sample list file, "**`SwissProt:EF1a_Giala`**" and switch "**Mode:**" to "**Editor**." SeqLab now displays the EF-1α protein sequence from *Giardia lamblia* in its Main Editor Window. Your display should look similar to my screen snapshot below:



This individual sequence comes from the GCG SwissProt database (remember logical_term:ID). It's named by its official SwissProt entry name (ID identifier) and appears in the Editor window with the amino acid residues color-coded. The nine color groups are based on a UPGMA clustering of the BLOSUM62 amino acid scoring matrix, and approximate physical property categories for the different amino acids. Turning off "Invert" causes the letters to assume the colors and the background to go white. Turning "Wrap" on causes the sequence to wrap vertically in the display. Use whichever combination of settings you prefer; my wrapped, non-inverted display follows below:

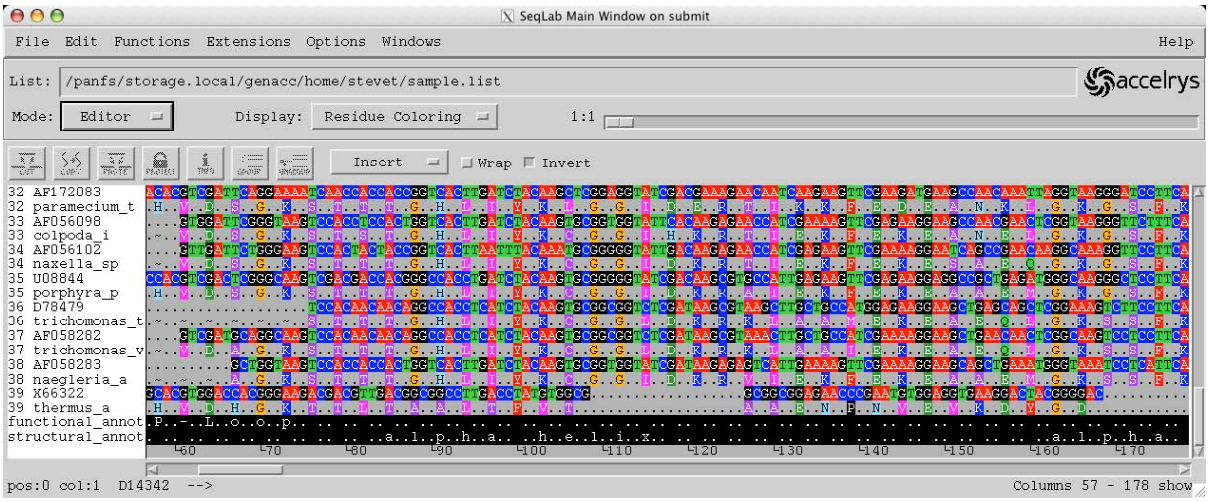


I prefer the default non-wrapped, inverted display. Quickly <**double-click**> on the entry's name (or single <click> the "INFO" icon with the sequence entry name selected) to see the database reference documentation for it. (This is the same information that you can get with the GCG command "`typedata -ref`" at the command line.) You can also change sequences' names and add any documentation that you want in this window. "**Close**" the "**Sequence Information**" window after reading it. Switch "**Mode:**" back to "**Main List**" to see the other sequences specified in our sample list file. Select the second entry in your list, "**`EF1a-primitive.SW.list`**." This is a list file within a list file. Notice the 'at' sign (@) in front of it in "`sample.list`." Remember the 'at' sign is necessary in the Wisconsin Package for specifying a list file. <**Double-clicking**> the entry 'opens' the file to show you all of its members; <**double-click**> it again to 'close' the list. Insure that "**`EF1a-primitive.SW.list`**" is selected and then switch "**Mode:**" back to "**Editor**." You should now see an unaligned dataset similar to the figure below:



The file "`EF1a-primitive.SW.list`" specifies a number of EF-1α proteins from 'primitive' eukaryotic organisms, i.e., those eukaryotes that exclude fungi, metazoans, and true plants, all from SwissProt. Those are the entries seen in the Editor display now. The previous *Giardia* sequence is a member of this dataset. This is an example of the type of dataset that can be specified by the list file output of GCG programs, including all the database similarity and reference searching programs. Change "**Mode:**" back to "**Main List**" so that we can see the last GCG sequence specification in your sample list.

Select the third entry there, the "**`EF1a-primitive.rsf{*}`**" multiple sequence alignment. Remember the {*} specifies all of the sequences within the Rich Sequence Format (RSF) file. Be sure "**`EF1a-primitive.rsf{*}`**" is selected and then switch "**Mode:**" back to "**Editor**" again. Expand the window to an appropriate size by 'grabbing' the bottom-right corner of its 'frame' and 'pulling' it out as far as desired. Any portion of, or the entire alignment, is now available for analysis by the GCG programs. My display follows:



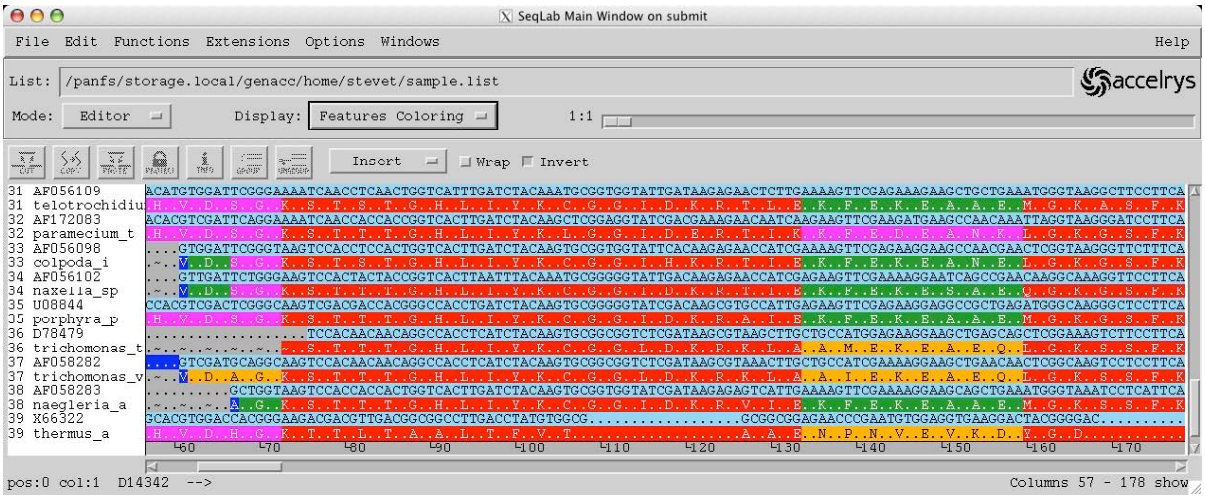
With this alignment loaded in the Editor explore the interface a bit. Nearly all GCG programs are accessible through the "Functions" menu (but several no longer work). You can select various entries' names and then go to the "Functions" menu to perform different analyses on them, but don't take the time to do that today. You can select sequences in their entirety by <clicking> on their names or you can select any position(s) within sequences by 'capturing' them with the mouse or by using the "Edit" menu "Select Range. . ." function. You can select a range of sequence names by <shift><clicking> the top-most and bottom-most name desired, or <ctrl><click> sequence entry names to select noncontiguous entries. (However, a bug in the Linux version of SeqLab prevents this from working correctly. The 'work-around' is to use the right mouse button, not the left along with the control key.) The "pos:" and "col:" indicators show you where your cursor is located on a sequence without including and with including any gaps respectively.

Notice this alignment contains both protein and DNA sequence data. SeqLab doesn't restrict you to using one or the other data types at a time. I've named each protein sequence with the Genus and species abbreviation for the organism it came from, and the DNA sequences are named by their GenBank Accession codes. It also contains on-screen text annotation, the lines entitled "functional_annotation" and

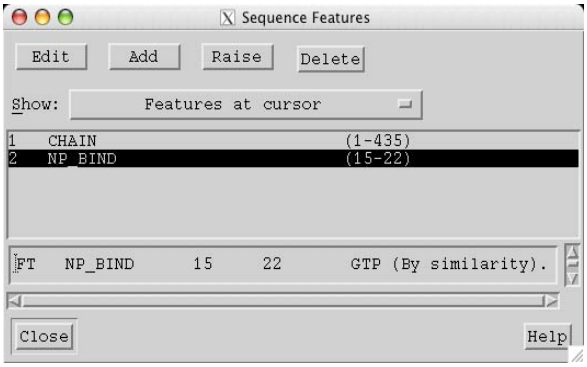
"structural_annotation." Use the scroll bars to move around within the alignment. The scroll bar at the bottom allows you to move through the sequences linearly (unless "wrap" is set); the one at the side allows you to scroll through all of the entries vertically.

Place your cursor anywhere within the alignment data. Press the <**space bar**> to insert gaps and move the sequence to the right. Periods (**.**) appear in the sequence to represent alignment gaps. Press <**delete**> to remove the gaps. A very powerful manual alignment function can be thought of as the 'abacus' function. To do this select a region that you want to slide flanked by gaps and then press the <**shift**> key as you move the region with the right or left arrow key. You can slide residues greater distances by prefacing the command keystrokes with the number of spaces that you want them to slide. Notice that as you move a DNA sequence its corresponding protein translation comes along. This is called "grouping" and is indicated by the group numbers associated with the entry names. The "GROUP" function allows you to manipulate 'families' of sequences as a whole — any change in one will be propagated throughout them all. Here I have each DNA sequence grouped to its corresponding protein translation. Select those sequences that you want to behave collectively and then <click> on the "GROUP" icon above your alignment to "GROUP" them. You can have as many groups as you want. You are not allowed to delete sequence characters unless you change their "Protections." This prevents you from accidentally changing your sequence data. <Click> on the padlock icon to produce a "Protections" window. Notice that the default protection allows you to modify "Gap Characters" and "Reversals" only. Check "All other characters" to allow you to "CUT" regions out of an alignment and/or delete individual residues and then <click> "OK" to close the window.

Change "**Display:**" from "**Residue Coloring**" to "**Feature Coloring**." The display now shows a color schematic of each entry's feature information based on its database Feature Table, and on various GCG analyses that can produce RSF feature files. My "Features Coloring" display follows below:

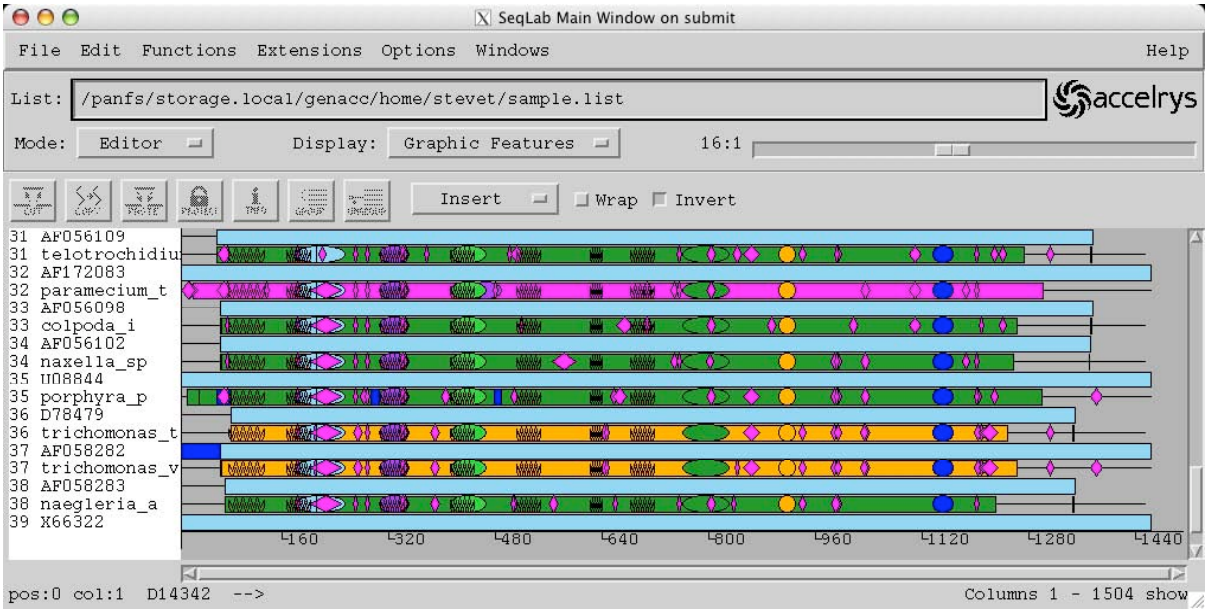


Quickly <**double-click**> on one of the various colored regions of the sequences (or use the "Features" choice under the "Windows" menu). This will produce a new window that describes the features located at the cursor, as seen to the right on the following page:



Select the feature to show more details and to select that feature in its entirety. All the features are fully editable through the "Edit" check box in this panel and new features can be added with several desired shapes and colors through the "Add" check box. "**Close**" the "**Feature Editor**" and "**Sequence Features**" windows, if you've opened them.

Next change "**Display:**" to "**Graphic Features**." "Graphic Features" represents features using the same colors as above, but in a 'cartoon' fashion. The "1:1" scroll bar near the upper right-hand corner allows you to 'zoom' in or out on the sequences. Move it to "2:1" and beyond and notice the difference in the display. Here's a subset of my alignment at a "**16:1**" zoom factor using the "Graphic Features" "Display" option:

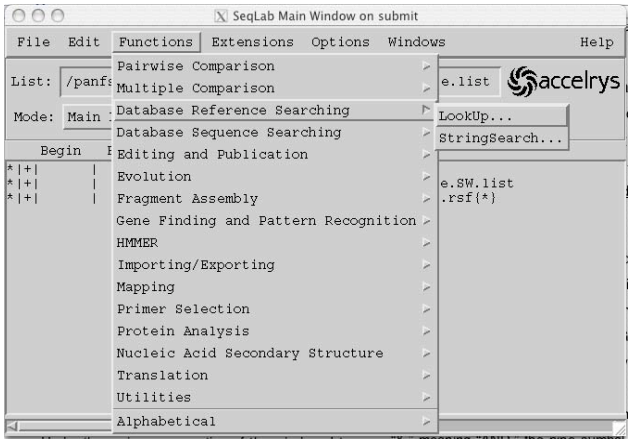


Switch "**Mode:**" back to "**Main List**" where we can explore some other ways of getting sequences into SeqLab. I mentioned some of the advantages of using a centralized, local biocomputing server system earlier and won't belabor the issue here. I will stress one point again though. Perhaps the biggest advantage of local biocomputing databases is convenience — there is often no need to download and reformat data! This can be a huge time saver, particularly if you are dealing with large datasets. So how do you access the sequence databases in the Wisconsin Package? As we saw earlier today, GCG offers a system of 'logical' terms that point to the data. Remember that these terms all support either proper sequence identifier names or accession codes or wild cards. Therefore a term like GB:* points to all of GenBank (less the Tags category), whereas SW:*_Ecoli points to all of the *Escherichia coli* entries in Swiss-Prot (since that particular database uses a consistent taxon naming convention). Within SeqLab the "Database Browser" allows you to peruse the databases using these conventions.

From SeqLab's "**Main Window**" launch the "**Database Browser**" found under the "**Windows**" menu. You should see something like the screenshot on left below. Notice the categories from the "Logical Terms" table are duplicated here in the "Database Browser." Click on a category to place the specification in the "Database Specification" text box and replace the wild card asterisk with your desired sequence identifier name or accession code. Pressing <return> or the "Show Matching Entries" button will then return the particular entry that you are looking for. You can then either "View Sequence" or "Add to Main Window" to place the entry in SeqLab's Main Window, List Mode or Editor Mode as the case may be.

But what if you don't have any clue about the sequence's proper identifier? What if you hadn't already searched at NCBI and don't particularly want to? How can you perform the same type of text based reference search in GCG as we did with NCBI's Entrez? The solution is "LookUp." LookUp is a Sequence Retrieval System (SRS) derivative (Etzold and Argos, 1993). It creates an output list file that can be used as input for other GCG programs. I'll use it here to create another representative list of elongation factor entries from the 'primitive' eukaryotes. This example will illustrate many of LookUp's rules and allow you to construct your own query for your project molecule.

"**Close**" the "Database Browser" window. Go to the "**Functions**" "**Database Reference Searching**" menu and choose "**LookUp. . .**" to launch the Wisconsin Package's sequence database annotation searching program. You should see a menu similar to that shown on the left below:



<**Click**> on "**LookUp. . .**" Be sure that "**Search the chosen sequence libraries**" is checked In the new "LookUp" window, and then select "**UniProt**" as the library to search. You can search as many libraries as you want, either protein or nucleotide. It just doesn't make sense to use both protein and nucleotide databases at the same time, since no subsequent programs can analyze both data types concurrently.

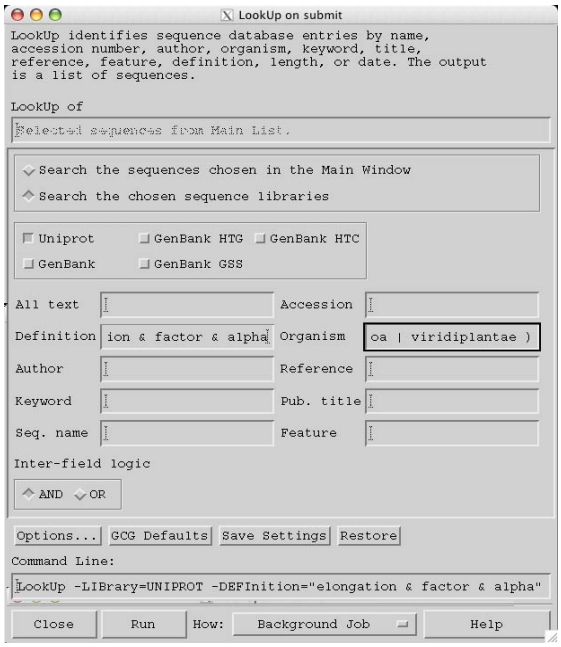
Under the main query section of the window, I type the words and symbols "elongation & factor & alpha" following the category "**Definition**" (the same as Entrez's "Title" field) and the words and symbols "eukaryota !

(fungi | metazoa | viridiplantae)" in the "**Organism**" category (separate every symbol with a space). Note the Boolean symbols rather than words between individual query terms. Obviously, and this is very important, do not use my exact query phrase — your query needs to be constructed so as to discover your desired project molecule, and not my example.

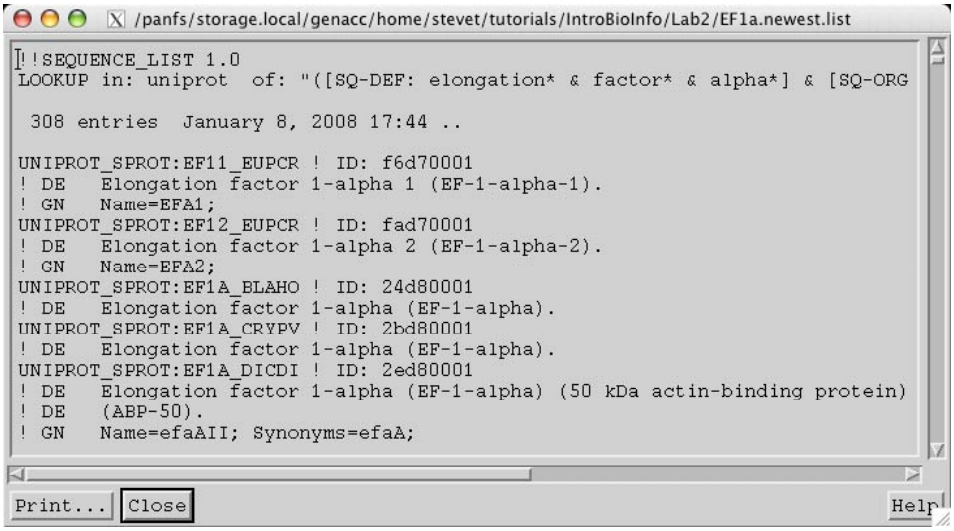
The Boolean symbols connect individual query strings because the databases are indexed using individual words for most fields. The "Organism" field is an exception; it will accept 'Genus species' designations as well as any single word supported level of taxonomy, e.g. "vertebrata." The Boolean operators supported by LookUp are the ampersand, "&," meaning "AND," the pipe symbol, "|," to denote the logical "OR," and the exclamation point, "!," to specify "BUT NOT."

Other LookUp query construction rules are case insensitivity, parenthesis nesting, "*" and "?" wildcard support, and automatic wildcard extension (e.g. "transcript" will find "transcriptional" and "transcript"). The wildcard extension can be turned off by ending your phrase with a pound symbol, "#."

My query should find most of the elongation factor alpha's from the 'primitive' eukaryotes in the current version of the UniProt database on HPC and will provide a good dataset example for my tutorials. Your "LookUp" window should look similar to my display shown opposite to the right, but with your query terms, not mine:



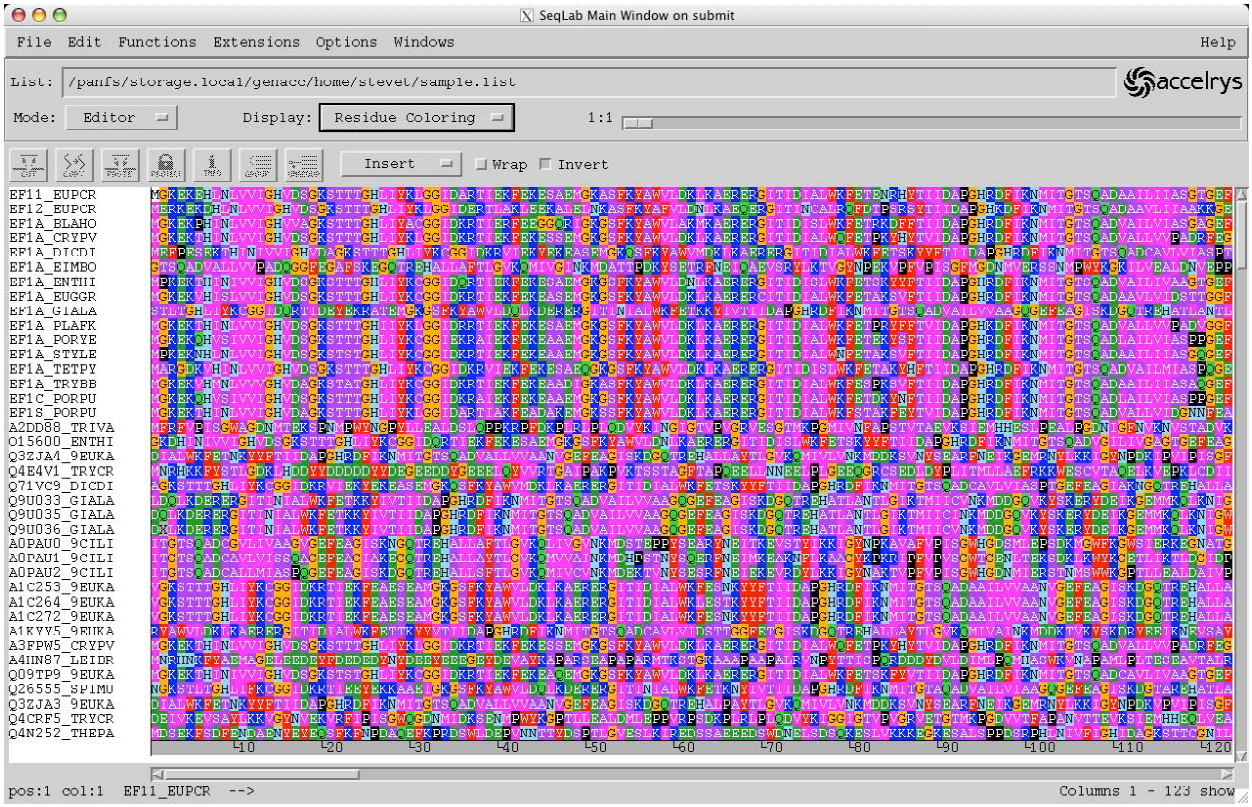
Next press the "**Run**" button. The program will display the results of the search; scroll through the output and then "**Close**" the window. The beginning of the LookUp output file from my example follows below:



I found 308 entries in UniProt that met my restrictions. As mentioned previously, be careful that all of the sequences included in the output from any text-searching program are appropriate. Improper nomenclature and other database inconsistencies can always cause problems. If you find inappropriate sequences upon reading the LookUp output, you can either edit the output file to remove them, or "CUT" them from the SeqLab Editor display after loading the list. Another option, if you use an editor, is to comment out the undesired sequences by placing an exclamation point, "!," in front of the unwanted lines. Exclamation points delineate comments in all GCG data fies.

Select the LookUp output file in the "**SeqLab Output Manager**." This is a very important window and will contain all of the output from your current SeqLab session. Files may be displayed, printed, saved in other locations or with other names, and deleted from this window. Press the "**Save As. . .**" button and give the LookUp output file a more appropriate name. Be sure not to change the directory specification, only changing that portion after the final slash. Next press the "**Add to Main List**" button in the "**SeqLab Output Manager**" and "**Close**" the window afterwards. This will add the results of the LookUp search to our previous "`sample.list`." Go to the "**File**" menu next and press "**Save List**." Next, be sure that your LookUp output file is selected in the "**SeqLab Main Window**" and then switch "**Mode:**" to "**Editor**." This will load the file into the SeqLab Editor where we could perform further analyses on the entries, if so desired.

All of your project sequences now appear in the Editor window. Expand the window to an appropriate size. The display should look similar to the graphic below after loading your list file (but with your project molecular system, not my example):



Other ways to get sequences into SeqLab include the “Add sequences from” “Sequence Files. . .” choice under the “File” menu. GCG format compatible sequences or list files are accessible through this route. You can also directly load sequences from the online GCG databases with the “Databases. . .” choice under the “Add sequences” menu if you know their proper identifier name or accession code. When using the “Add Sequences from” “Sequence Files. . .” function, the “Filter” box can be quite helpful. By default all of the files in your working directory are displayed due to a “*” wild card filter. You may want to use some other term in the “Filter” box to restrict what is displayed. Press the “Filter” button to display only those files that you ‘filtered.’ Select the file that you want from the “Files” box, and then check the “Add” and then “Close” buttons at the bottom of the window to put the desired file into your current list, if you’re in List Mode, or directly into the Editor, if you’re in “Editor Mode.” Furthermore, in “Editor Mode” two additional choices are available. You can “Import” sequences from GenBank or FastA format files or from ABI style binary trace files. And you can use the “File” menu “New Sequence” choice to create empty slots to hold brand new entries, either “DNA,” “RNA,” “Protein,” or “Text,” where you can either type in data or copy and paste it from a different window.

It’s probably a good idea to save the sequences in the display at this point and multiple times down the road as you work on a dataset. Do this occasionally the whole time you’re in SeqLab just in case there’s an interruption of service for any reason. Go to the “**File**” menu and choose “**Save As**.” Accept the default “**`.rsf`**” extension and its directory location, but give it some file name that makes sense to you. Remember, RSF (Rich Sequence Format) contains all the aligned sequence data as well as all the reference and feature annotation associated with each entry, and is SeqLab’s default sequence format.

That’s enough SeqLab for now. Switch “**Mode:**” back to “**Main List**” and then use the “**File**” menu to “**Save List**.” Finally, return to SeqLab’s “**File**” menu and choose “**Exit**” to leave the interface. You may be asked to “Save” the RSF file and the working list, if you’ve made any changes in either since you last saved them. If so, check “OK” and the windows will go away.

Sequence format issues

Before leaving GCG topics and logging off HPC we should discuss one more thing. What happens if you get sequences from sources other than the GCG online databases; how do you get them into the Wisconsin Package? This is the realm of sequence format. I’ve already mentioned that SeqLab has the ability to “Import” some alternate formats, and that route is always worth trying, but what if that doesn’t work? Well GCG has the SeqConv+ format conversion program that works in most of these cases. It’s under SeqLab’s “Functions” menu in the “Importing/Exporting” category. Or, perhaps even easier, just launch it at the command line with “`seqconv+`.” SeqConv+ recognizes a slew of different sequence formats and allows the back and forth conversion between them.

However, what if somebody has e-mailed you some data that is in no recognized sequence format and is just plain text, i.e. just a string of A’s, C’s, T’s, and G’s? One way of dealing with this situation is to ‘cut-and-paste’ it into SeqLab’s Editor. However, cutting and pasting can be problematic — you may accidentally miss part of

the sequence, or your mouse buttons may not be mapped correctly to emulate the proper X windowing cut and paste actions. A far more reliable route is to edit the sequence, either with a UNIX editor like nano on HPC, or with your favorite word processor (as long as you save it as text only!) followed by file transfer.

This was briefly discussed earlier when I described specifying single sequence format (SSF) to GCG. In review, open a plain ASCII text sequence file with an editor, open a blank line above the sequence data and place the 'greater than' symbol (>) there, immediately followed by the sequence's name. Only sequence data is allowed below that line. In general, do not use spaces or punctuation other than hyphens or underscores in the sequence's name. Save your revision and exit the editor. If the editing was done on a different platform than the HPC, you'll have to transfer the file to the HPC with scp. Either way, the next step is to run the GCG program SeqConv+ on the file. This will take the FastA format that you produced with the editing and convert it into valid GCG single sequence format (SSF).

To see how this works open the HPC's "**nano**" editor to create a test sequence in your terminal window:

```
$ nano test.fsa
```

Enter the sequence's name and some sample amino acid sequence data following the directions in the paragraph above to create a FastA format test peptide sequence. The sequence can be as long or short as you wish and can contain any mix of valid amino acid characters. Press < **control** > "**x**" to save the entry and exit nano. Next run GCG's SeqConv+ command on the entry to convert it to a valid GCG SSF file:

```
$ seqconv+ -in=test.fsa -out=test.pep -format=ssf
```

Check out the new GCG SSF entry with the more command:

```
$ more test.pep
```

Obviously this would not be the most efficient way to create a new sequence whenever you needed to. Instead you should probably directly import automated sequencing machine trace data or raw ASCII sequence text directly into SeqLab, or you could type sequence characters right into the SeqLab editor, or you could even use GCG's SeqEd command line sequence editor, all producing valid GCG format without the need to run SeqConv+. However, I make you do it here so that you can see how it works. There are many occasions when you may need to use it in the future. Log off HPC and exit your ssh terminal session upon finishing this portion of the lab.

Homework assignment

Sometime before next week's lab send me this week's lab report. As with last week's, the lab report is Web based. Open a Web browser window and go to the Course Lab home page. This should still be your SCS Classroom account Web home page from last week. Select this week's lab report link, "**Lab Report #2**." Use the form to tell me what project molecular system from my list you've chosen for this and the next eight labs. Also tell me what you discovered on your own time from the Web portion of the lab regarding human or

Arabidopsis homologues and associated disease states and chromosomal locations. Finally 'copy and paste' your reformatted file "**`test.pep`**" in its entirety into the form, and send it to me. Next week we jump right into our lab projects by taking the LookUp list that we created today and using it to design PCR primers for our hypothetical genomic search for an unknown gene.

Conclusion

You should now have a basic understanding of how the GCG Wisconsin Package for sequence analysis works, and of how to use their SeqLab GUI. We will be using SeqLab all semester long — get used to it. Furthermore, you've been exposed to a vast array of bioinformatics databases. They come in a bewildering variety of forms, and access methods vary tremendously depending on the database. Becoming familiar with appropriate methods for the various databases, and learning what each database has to offer, both its strengths and weaknesses, will make your biocomputing experiences much less frustrating. This lab tutorial just barely touches the surface of all that is available, particularly on the World Wide Web, and I encourage you to explore these resources much further on your own time. Contact me with any questions you may have about what additional types of biological molecular data is available and where that data can be found.

References

Bairoch, A. (1991) The Swiss-Prot Protein Sequence Data Bank. *Nucleic Acids Research* **19**, 2247–2249.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M., (1977) The Protein Data Bank: A computer-based archival file for macromodel structures. *Journal of Molecular Biology* **112**, 535–542.

Bilofsky, H.S., Burks, C., Fickett, J.W., Goad, W.B., Lewitter, F.I., Rindone, W.P., Swindell, C.D., and Tung, C.S. (1986) The GenBank™ Genetic Sequence Data Bank. *Nucleic Acids Research* **14**, 1–4.

Dayhoff, M.O., Eck, R.V., Chang, M.A. and Sochard, M.R. (1965) *Atlas of Protein Sequence and Structure*, Vol. 1. National Biomedical Research Foundation, Silver Spring, MD, U.S.A.

Etzold, T. and Argos, P. (1993) SRS — an indexing and retrieval tool for flat file data libraries. *Computer Applications in the Biosciences* **9**, 49–57.

George, D.G., Barker, W.C., and Hunt, L.T. (1986) The Protein Identification Resource (PIR). *Nucleic Acids Research* **14**, 11–16.

Genetics Computer Group (GCG®), (Copyright 1982-2008) *Program Manual for the Wisconsin Package®*, version 11, http://www.accelrys.com/ Accelrys Inc., San Diego, California, U.S.A.

Hamm, G.H. and Cameron, G.N. (1986) The EMBL Data Library. *Nucleic Acids Research* **14**, 5–10.

Hogue, C.W., Ohkawa, H. and Bryant, S.H. (1996) A dynamic look at structures: WWW-Entrez and the Molecular Modeling Database. *Trends in Biochemical Sciences* **21**, 226–229.

National Center for Biotechnology Information (NCBI) *Entrez and CN3D*, public domain software distributed at: http://www.ncbi.nlm.nih.gov/ National Library of Medicine, National Institutes of Health, Bethesda, Maryland, U.S.A.

Online Mendelian Inheritance in Man, OMIM™. (1996) http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim Center for Medical Genetics, Johns Hopkins University, Baltimore, Maryland, U.S.A. and National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland, U.S.A.

Pearson, P., Francomano, C., Foster, P., Bocchini, C., Li, P., and McKusick, V. (1994) The Status of Online Mendelian Inheritance in Man (OMIM) medio 1994. *Nucleic Acids Research* **22**, 3470–3473.

Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence analysis. *Proceedings of the National Academy of Sciences U.S.A.* **85**, 2444–2448.

Sayle, R.A. and Milner-White, E.J. (1995) RasMol: biomolecular graphics for all. *Trends in Biochemical Sciences* **20**, 374–376.

Schuler, G.D., Epstein, J.A., Ohkawa, H., and Kans, J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods in Enzymology* **226**, 141–162.