Last modified August 15, 2018

Elementary Sequence Analysis

Brian Golding, Dick Morton and Wilfried Haerty

Department of Biology
McMaster University
Hamilton, Ontario
L8S 4K1

These notes are in Adobe Acrobat format (they are available upon request in other formats) and they can be obtained from the website http://helix.biology.mcmaster.ca/courses.html. Some of the programs that you will be using in this course and which will be run locally can be found at http://evol.mcmaster.ca/p3S03.html.

The "blue text" should designate links within this document while the "red text" designate links outside of this document. Clicking on the latter should activate your web browser and load the appropriate page into your browser. If these do not work please check your Acrobat reader setup. The web links are accurate to the best of our knowledge but the web changes quickly and we cannot guarantee that they are still accurate. The links designated next to the JAVA logo,  , require that JAVA be installed on your computer.

These notes are used in Biology 3S03. The purpose of this course is to introduce students to the basics of bioinformatics and to give them the opportunity to learn to manipulate and analyze DNA/protein sequences. Of necessity only some of the more simple algorithms will be examined.

The course will hopefully cover . . .

- databases of relevance to molecular biology.
- some common network servers/sites that provide access to these databases.
- methods to obtain sequence analysis software and data.
- methods of sequence alignment.
- methods of calculating genetic distance.
- methods of phylogenetic reconstruction.
- methods for detecting patterns and codon usage.
- methods for detecting gene coding regions.

The formal part of the course will consist of two approximately one hour lectures each week. Weekly assignments will be be provided to practice and explore the lecture material. In addition there will be an optional tutorial to help students with these assignments or other problems. These assignments will be 40% of your grade and three, in class quizzes will make up the remainder.

We would appreciate any comments, corrections or updates regarding these notes.

Golding@McMaster.CA Morton@McMaster.CA HaertyW@McMaster.CA

This document was entirely constructed with pdfLaTeX. Enormous kudos to all those that make this great software free to everyone.

Table of Contents in Brief

In order to speed download, I place here links to the individual chapters in pdf format. The contents of these are shown on the following 'Contents' pages but note that the links will function only for the individual chapter included here.

Contents

Chapter 4

Databases

4.1 Introduction

Molecular biology has undergone amazing advances in the last twenty years. We can now sequence DNA and proteins in most any laboratory in the country. Indeed it is sometimes even given to undergraduate students as an laboratory exercise. Most universities don't do this but this is because of the use of radioisotopes rather than the difficulty of the technique. The ability to rapidly and easily sequence DNA has also lead to a shift in the way that science is now done. With automated, massively parallel sequencing machines we would no longer give undergraduates an exercise to sequence something as this is the realm of robots and not humans.

It has become easier to simply sequence the gene (sometimes a whole genome) as more preliminary information can be often be gained this way than to carry out a sophisticated and well thought out experiment. These advances have lead to the establishment of genome projects. In the past there were large scale projects to set up laboratories to sequence DNA in an efficient way rather than having each laboratory do the sequencing *in house*. The largest of these projects was the human genome project to which the United States government alone committed $3,000,000,000.00 (that is three billion!). Other governments of the world also supported their own projects. Additionally, more organisms than just humans are being examined and many have been sequenced. Within the last half decade, the cost of sequencing and the cost of the sequencers has fallen so drastically that genome projects can be done by individual laboratories. Indeed, the estimated cost to sequence a single human genome is approaching $1000 or less.

Computer technology has also undergone an amazing advance in the last twenty years. It is now unusual for scientists not to have an extremely fast, multi-processor computer on their desk. Additionally, though somewhat contrary to popular use, these computers are not simply fancy typewriters. They have many capabilities beyond word processing and can deal with a large amount of information. They are also capable of doing analyses that are beyond the computational ability of any scientist.

One of the other major advances in computer technology has been in connectivity. Computers are now connected to networks that permit access to other computers all over the world. This of course allows students to chat with one another but it also means that the data generated by these sequencing machines (and other biological instruments) are available to anyone anywhere in the world. This permits anyone with a computer to access databases of all kinds - if they know how. The purpose of this section is to provide you an entry point to this knowledge.

None of the genome projects, nor most of the other projects that create databases, would have been funded if their research was kept private. Indeed an openness about research results has been a long standing principle that has guided science. It is oft quoted that "the experiment is not finished until it has been published". Publication has been the traditional method of permitting worldwide access to research results. However, the retrieval of this information can be a labour intensive practice that required great skill in the days of paper publications. Here, I am mainly referring to simple factual data rather than experiments that require interpretation. To accumulate this factual data and to make use of it is often difficult. With computer databases, however, this data is as accessible to you as it is to the expert that compiled it. You can bring the data directly to your desktop in its entirety, cut/paste the pieces you want, and analyze it according to your fancy. An article by

W. Gilbert in NATURE suggests that this combination of advances will lead to a shift in the way science will be done in the future.

Towards a paradigm shift in biology.
W. Gilbert NATURE 349:99 1991.

The steady conversion of new techniques into purchasable kits and the accumulation of nucleotide sequence data in the electronic data banks leads one practitioner to cry, "Molecular biology is dead - Long live molecular biology!".

There is a malaise in biology. The growing excitement about the genome project is marred by a worry that something is wrong - a tension in the minds of many biologist reflected in the frequent declaration that sequencing is boring, and yet everyone is sequencing. What can be happening? Our paradigm is changing.

Molecular biology, from which has sprung the attitude that the best approach is to identify a relevant region of DNA, a gene, and then to clone and sequence it before proceeding, is now the underpinning of all biological science. Biology has been transformed by the ability to make genes and then the gene products to order. Developmental biology now looks first for a gene to specify a form in the embryo. Cellular biology looks to the gene to specify a structural element. And medicine looks to genes to yield the body's proteins or to trace causes for illnesses. Evolutionary questions - from the origin of life to the speciation of birds - are all traced by patterns on DNA molecules. Ecology characterizes natural populations by amplifying their DNA. The social habits of lions, the wanderings of turtles and the migrations of human populations leave patterns on their DNA. Legal issues of life or death can turn on DNA fingerprints.

And now the genome project contemplates working out the complete DNA pattern and listing every one of the genes that characterize all of the model species that biologist study - ourselves even included.

At the same time, all of these experimental processes - cloning, amplifying and sequencing DNA - have become cook-book techniques. One looks up a recipe in the Maniatis book, or sometimes simply buys a kit and follows the instructions in the inserted instructional leaflet. Scientists write letters bemoaning the fact that students no longer understand how their experiments really work. What has been the point of their education?

The questions of science always lie in what is not yet known. Although our techniques determine what questions we can study, they are not themselves the goal. The march of science devises ever newer and more powerful techniques. Widely used techniques begin as breakthroughs in a single laboratory, move to being used by many researchers, then by technicians, then to being taught in undergraduate courses and then to being supplied as purchased services - or, in their turn, superseded.

Fifteen years ago, nobody could work out DNA sequences, today every molecular scientists does so and, five years from now, it will all be purchased from an outside supplier. Just this happened with restriction enzymes. In 1970, each of my graduate students had to make restriction enzymes in order to work with DNA molecules; by 1976 the enzymes were all purchased and today no graduate student knows how to make them. Once one had to synthesize triphosphates to do experiments; still earlier, of course, one blew one's own glassware.

Yet in the current paradigm, the attack on the problems of biology is viewed as being solely experimental. The 'correct' approach is to identify a gene by some direct experimental procedure - determined by some property of its product or otherwise related to its phenotype - to clone it, to sequence it, to make its product and to continue to work experimentally so as to seek an understanding of its function.

The new paradigm, now emerging, is that all the 'genes' will be known (in the sense of being resident in databases available electronically), and that the starting point of a biological investigation will be theoretical. An individual scientist will begin with a theoretical conjecture, only then turning to experiment to follow or test that hypothesis. The actual biology will continue to be done as "small science" - depending on individual insight and inspiration to produce new knowledge - but the reagents that the scientist uses will include a knowledge of the primary sequence of the organism, together with a list of all previous deductions from that sequence.

How quickly will this happen? It is happening today: the databases now contain enough information to affect the interpretations of almost every sequence. If a new sequence has no match in the databases as they are, a week later a still new sequence will match it. For 15 years, the DNA databases have grown by 60 per cent a year, a factor of ten every five years. The human genome project will continue and accelerate this rate of increase. Thus I expect that sequence data for all of the model organisms and half of the total knowledge of the human organism will be available in five to seven years, and all of it by the end of the decade.

To use this flood of knowledge, which will pour across the computer networks of the world, biologists not only must become computer- literate, but also change their approach to the problem of understanding life.

The next tenfold increase in the amount of information in the databases will divide the world into haves and have-nots, unless each of us connects to that information and learns how to sift through it for the parts we need. This is not more difficult than knowing how to access the scientific literature as it is at present, for even that skill involves more than a traditional reading of the printed page, but today involves a search by computer.

We must hook our individual computers into the worldwide network that gives us access to daily changes in the database and also makes immediate our communications with each other. The programs that display and analyze the material for us must be improved - and we must learn how to use them more effectively. Like the purchased kits, they will make our life easier, but also like the kits we must understand enough of how they work to use them effectively.

The view that the genome project is breaking the rice bowl of the individual biologist confuses the pattern of experiments done today with the essential questions of the science. Many of those who complain about the genome project are really manifesting fears of technological unemployment. Their hard-won PhDs seem suddenly to be valueless because they think of themselves as being trained to a single marketable skill, for a particular way of doing experiments. But this is not the meaning of their education. Their doctorates should be testimonials that they had solved a novel problem, and in so doing had learned the general ability to find whatever new or old techniques were needed; a skill that transcends any particular problem.

To indicate how far this shift has occurred, a famous author had the temerity to publish an article in Cell with the title "Sequence first. Ask questions later".

There is now a new concept of public data. Everyone that desires access can retrieve this data. This includes not only scientists and medical practitioners but also private companies and members of the general public. The data is also raising a large number of ethical problems that have not been fully considered.

These advances have combined to create a new field of science. This is called bioinformatics (along with its relative – medical informatics). It is, basically, a mixture of computer science, mathematics, and biology. It combines aspects from all three fields to study the methods and the problems associated with the task of bringing information to a researcher, sorting this mass of information in a meaningful way, and then analyzing it.

Our concern in this section will be focused on the databases of relevance to molecular biology. However, you should be aware that this is but the tip of the iceberg – there are many databases of many natures. There are other biological databases such as some of a biochemical nature, one on enzyme kinetics, some of a more general nature, and some just plain weird.

The major databases for molecular biology are centered around the molecular sequence databases. The genome projects supplying these databases promise to yield the greatest mass of data that biology has ever seen. The human genome alone covers 3 billion nucleotides. In February of 2001, the completion of the human genome draft sequence was jointly announced by the private company Celera Genomics and the publically funded Human Genome Project. This represents an enormous accomplishment and will probably represent the biggest achievement since the discovery of the structure of DNA.

Figure 4.1: The completion of a first draft of the human genome was announced in February of 2001

But a single human genome (and a mosaic of several individuals at that) was only the beginning. In June 2012 there were over 1000 distinct human individuals completely sequenced as part of a large project; the 1000 human genome project. But this is just 1000 genomes. In 2016, a single article published in PNAS reported the "Deep sequencing of 10,000 human genomes"; just one paper.

And there is no reason to stop at humans. There are many other eukaryotes whose genome has been sequenced and even more in the pipeline. Currently in the public domain there are many more whole eukaryotic genome shotgun component projects nearly completion. These genomes include fish, nematodes, insects, birds, mouse, rats, cows, plants, and many more to come.

This mass of data also presents many problems – how do you store all of this information, how do you access it, and move it? The rate of accumulation of sequence data is exponentially growing. This has been partly due to the fact that the technology to carry out DNA sequencing has rapidly advanced. Today, almost the entire job can be carried out by robots – from an input of tissue, the robots can automatically extract the DNA, amplify regions of interest, and prepare sequence cocktails. These are then loaded onto the gels of automatic sequencing machines. These machines will run the gels, a laser will scan the gels and calculate the DNA sequence in the case of Sanger sequencing or in the case of the other machines the similar automated processes occur with sequence reads automatically entered in computer files. Finally, the sequence is often automatically passed on to computer clusters for preliminary analysis and these computers might automatically assemble the fragments, search/compile databases, or other analyzes. In addition, as the cost goes down, the number of laboratories that routinely sequence DNA has increased.

The result of this increased activity is shown in Figure 4.2. Some of this data is annotated but since 2004 EMBL has included in its data releases nucleotides of mostly unannotated whole genome shotgun data. Over 692 billion of the nucleotides in the database come from this and other raw data sources. The current (June 2016) official EMBL release 128 yields over 1700 billion nucleotides and is the data that is plotted Figure 4.2. The rate of growth of the data has been close to exponential for many years. Obviously an exponential growth cannot continue (physical laws prevent this). However, I have stated this every year since I (BG) first taught this course in 1990. I think that finally (after many years) I have been proven correct and physical laws have taken hold (hmmmm, maybe not; comparable figures from other years are 1993, 1995, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017). (How can one tell if it is exponential? Well a simple way is to see if the data is a straight line when plotted on a

log scale. Take a look at a log plot and see if you think that these data are linear). Perhaps a definite slow down in the rate within EMBL is seen but the rate is still exponential at a somewhat slower pace.

The slow down is, however, not so much due to the mundane reason that such things are physically impossible. No indeed, the advances in sequencing technology have more than kept pace as evidenced by the previous chapter. Rather it is because the cost of sequencing has been reduced and a new type of DNA sequencing project is emerging to become more and more common. Resequencing of an already known genomes was an unthinkable cost just a few years ago. But by July 2008, 36 strains of *Saccharomyces cerevisiae* had been completely sequenced, fifteen strains of mice, and three nematodes. Such resequencing data is often not entered into the databases and indeed, for some bacterial genomes the resequence data from thousands of genomes might even be considered to be too cheap to be worth the effort of finishing the data and trying to get it submitted.

More large sequence projects beyond the human genome are in progress. The sequencing of the entire genomes of 1000 humans is completed and the data is now public record with a devoted NCBI browser. There are now large (multi-institutional) projects to sequence 10,000 eukaryotic genomes; projects to sequence human centenarians; projects to sequence the entire human microbiome (all the bacteria, fungi and microbes associated with humans); a total of 100,000 bacterial genomes (!!).

We now enter the new realm of population genomics. No longer is it suitable to sequence the entire genome of a species. It is now feasible and academically intriguing to sequence all of the genomes of a population or of a community. So for example, the NCBI pathogens database (https://www.ncbi.nlm.nih.gov/pathogens/) contains data on the genomes of bacterial pathogens; for Salmonella there are 139,754 entries!

Such data are unlikely to find their way into the databases in their entirety. But regardless, all of this mass of data is open for analysis and is a rich research field. It is now stored (at NCBI; below) in a separate archive called the Sequence Read Archive (SRA). In August 9, 2018 it contained 19,469,647,048,078,497 total bases (19 peta-bases). Even the name of this database has had to change. With the growth of sequencing technology, the original name "Short Read Archive" was no longer appropriate with the growing lengths of reads. It's growth is shown in Figure 4.3. Note that, on this scale, the database didn't exist just a few years ago.

There are three major nucleotide sequence databases. These are EMBL (European Molecular Biology Laboratory), NCBI (the U.S. National Center for Biotechnology Information) and DDBJ (the DNA Data Bank of Japan). Each of these databases attempt to collect all of the known nucleic acid (DNA/RNA) sequences. The sequences were collected from published sources and most journals now require submission of the sequences to a database before publication is permitted. Many sequences are directly deposited into the databases and will not be published in any other form. In addition to the sequences, the databases also contain many other useful bits of data, including (but not limited to) organism, tissue, function, and bibliographic information.

All three of these organizations are in electronic contact with each other and exchange sequence information daily. Hence, you need not worry that one database might not have the sequence of interest but a search of some other database would have it (at least in theory, anyway).

The following sections are intended to give you a flavour of the database contents.

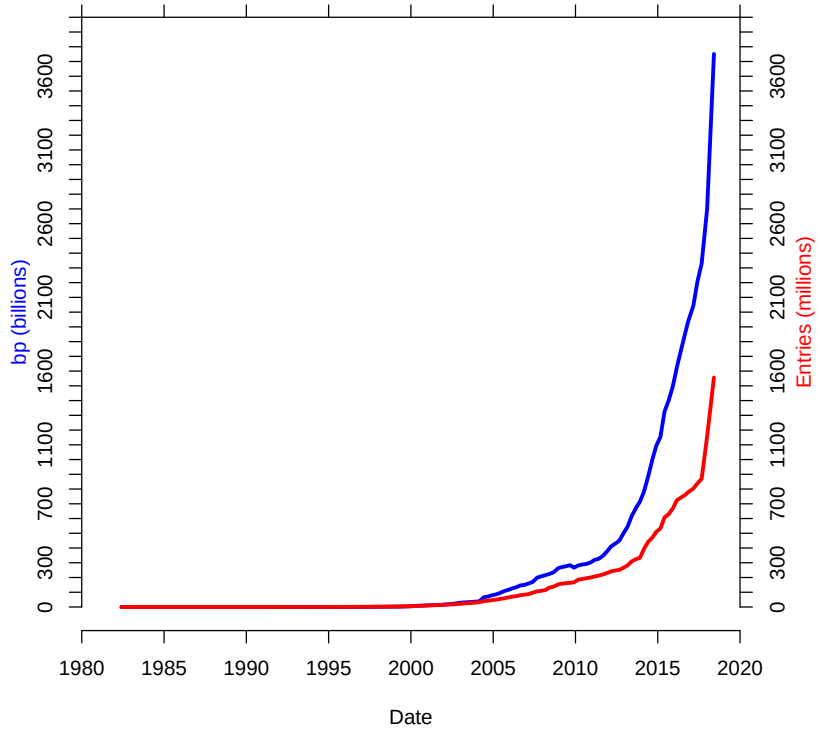


Figure 4.2: The growth of the EMBL database

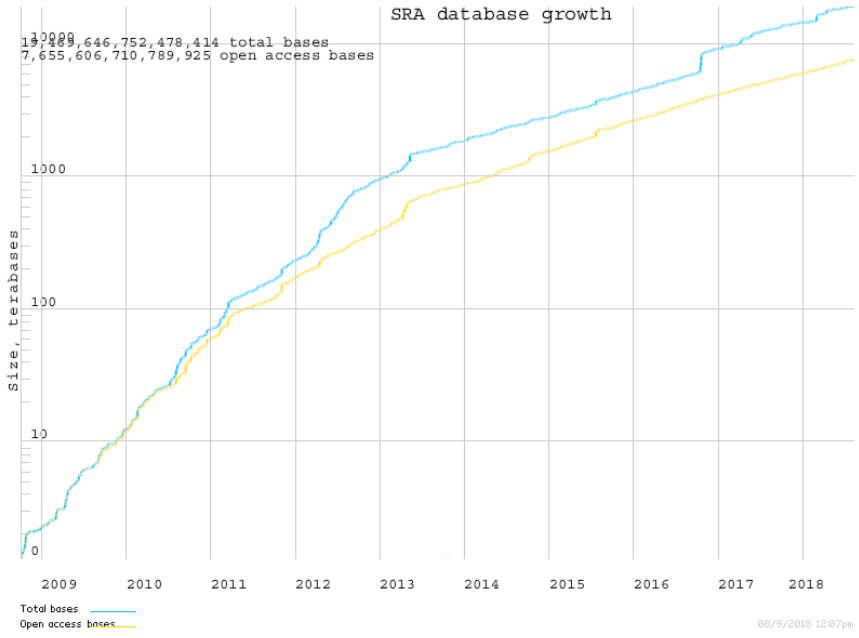


Figure 4.3: The growth of the SRA database

4.2 N.C.B.I.

The easiest way to explain what is contained in the database is to examine an actual entry from the data base. This is shown below. This example contains the nucleotide sequence of the first exon of the human lung adenocarcinoma (PR310) c-K-ras oncogene.

Note that the actual sequence information provided at the end of the entry may be, as in this case, only a small fraction of the total data entry. NCBI organizes its entries onto several lines each of which begin with a special header. The first header and that which always begins the entry is the LOCUS name. This provides a identifying code word (in uppercase) to be associated with this entry. It also gives the length of the sequence and the date the sequence was entered or last modified.

Example Entry #1: GenBank/NCBI for HUMCKRASA

```
LOCUS       HUMCKRASA     450 bp ss-mRNA            PRI       15-SEP-1990
DEFINITION  Human PR310 c-K-ras protein mRNA, 5' end.
ACCESSION   M35504
KEYWORDS    c-K-ras protein; c-myc oncogene.
SOURCE      Human (patient PR310) lung carcinoma, cDNA to mRNA.
  ORGANISM  Homo sapiens
            Eukaryota; Animalia; Chordata; Vertebrata; Mammalia; Theria;
            Eutheria; Primates; Haplorhini; Catarrhini; Hominidae.
REFERENCE   1  (bases 1 to 450)
  AUTHORS   Yamamoto,F., Nakano,H., Neville,C. and Perucho,M.
  TITLE     Structure and mechanisms of activation of c-K-ras oncogenes in
            human lung cancer
  JOURNAL   Prog. Med. Virol. 32, 101-114 (1985)
  STANDARD  full automatic
FEATURES             Location/Qualifiers
     CDS             1..>450
                     /note="PR310 c-K-ras oncogene"
                     /codon_start=1
                     /translation="MTEYKLVVVGAGGVGKSALTIQLIDNHFVDEYDPTIEDSYRKQV
                     VIDGETCLLDILDTAGHEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKR
                     VKDSEDVPMVLVGNKCDLPSRTVDTKQAQDLARSYGIPFIQTSAKTRQ"
   source            1..450
                     /organism="Homo sapiens"
  pept        1  >   450     PR310 c-K-ras oncogene
BASE COUNT      155 a     71 c    106 g    118 t
ORIGIN
        1 atgactgaat ataaacttgt ggtagttgga gctggtggcg taggcaagag tgccttgacg
       61 atacagctaa ttgacaatca ttttgtggac gaatatgatc caacaataga ggattcctac
      121 aggaagcaag tagtaattga tggagaaacc tgtctcttgg atattctcga cacagcaggt
      181 catgaggagt acagtgcaat gagggaccag tacatgagga ctggggaggg ctttctttgt
      241 gtatttgcca taaataatac taaatcattt gaagatattc accattatag agaacaaatt
      301 aaaagagtta aggactctga agatgtacct atggtcctag taggaaataa atgtgatttg
      361 ccttctagaa cagtagacac aaaacaggct caggacttag caagaagtta tggaattcct
      421 tttattcaaa catcagcaaa gacaagacag
//
```

The next line contains a short DEFINITION of the sequence that is contained in the entry. An ACCESSION number is a unique identifying sequence for this data entry. Note that only the accession number will necessarily be constant across nucleotide databases (NCBI, EMBL, DDBJ). The accession numbers are unique among these three nucleotide databases but are not necessarily unique between other databases (e.g. between protein and nucleotide databases). LOCUS names are variable and can be changed. The LOCUS names are often changed as nomenclature is changed or as sequences are merged into larger entries. The ACCESSION number and the LOCUS names are two character strings that can be easily used to access and retrieve sequence entries from NCBI.

Following these, come KEYWORDS that identify the particular entry. The SOURCE line describes how the sequence was cloned/sequenced and the ORGANISM line describes the species/construct from which the sequence originates. Following this, is a description of REFERENCES that deal with this entry. Note that this would include only the original papers describing the sequencing and not any other subsequent papers that might analyze the sequence. Multiple references will be given when different labs have sequenced the same DNA or when different publications describe different parts of the sequence. Throughout the NCBI database, numbers in square brackets indicate items in the REFERENCE list. The STANDARD describes any checks on the accuracy of the sequence.

The FEATURES section will describe things such as coding sequence start/stop, leader sequence start/stop, presence of signal sequences, locations of exons/introns, repeats, polymorphisms, and so on. There may also be comments in this

section that can be useful. It may describe the some of the interesting facts that may go along with this sequence. This might include why it was sequenced, how it relates to other sequences in the database, some unusual features of the sequence, etc.

The BASE COUNT line gives the proportions of each nucleotide in the sequence. The ORIGIN line gives details of where the sequence starts relative to restriction sites (or other location markers) that aided the cloning.

Finally the sequence follows in lower case, in groups of 10 and with the number of the first nucleotide given on the left.

The above example entry is a particularly short sequence. This is not the norm for NCBI entries. Most entries contain a longer sequence as shown in the example below.

Example Entry #2: GenBank/NCBI - GORHBBPG

```
LOCUS       GORHBBPG     7055 bp    DNA             PRI       13-JUL-1993
DEFINITION  Gorilla beta-globin and eta-globin pseudogenes and an Alu repeat.
ACCESSION   K02543 M18037
NID         g177056
KEYWORDS    Alu repeat; globin; hemoglobin; pseudogene.
SOURCE      Lowland gorilla (Gorilla gorilla gorilla; SF-4) blood DNA, clone
            Ggo lambda-1059-1.1 [1].
  ORGANISM  Gorilla gorilla
            Eukaryota; Animalia; Chordata; Vertebrata; Mammalia; Theria;
            Eutheria; Primates; Haplorhini; Catarrhini; Pongidae.
REFERENCE   1  (bases 1613 to 3763)
  AUTHORS   Chang,L.Y. and Slightom,J.L.
  TITLE     Isolation and nucleotide sequence analysis of the beta-type globin
            pseudogene from human, gorilla and chimpanzee
  JOURNAL   J. Mol. Biol. 180, 767-784 (1984)
  MEDLINE   85134894
REFERENCE   2  (bases 1613 to 3763)
  AUTHORS   Koop,B.F., Goodman,M., Xu,P., Chan,K. and Slightom,J.L.
  TITLE     Primate eta-globin DNA sequences and man's place among the great
            apes
  JOURNAL   Nature 319, 234-238 (1986)
  MEDLINE   86118664
REFERENCE   3  (bases 1 to 7055)
  AUTHORS   Miyamoto,M.M., Slightom,J.L. and Goodman,M.
  TITLE     Phylogenetic relations of humans and African apes from DNA
            sequences in the pseudo-eta-globin region
  JOURNAL   Science 238, 369-373 (1987)
  MEDLINE   88018021
COMMENT     [3]  revises [1],[2].
            Computer-readable sequence for [3] kindly provided by M.M.Miyamoto,
            6-FEB-1988.

            NCBI gi: 177056
FEATURES             Location/Qualifiers
     source          1..7055
                     /organism="Gorilla gorilla"
                     /isolate="SF-4"
                     /sub_species="gorilla"
                     /sequenced_mol="DNA"
                     /tissue_type="blood"
     repeat_region   1067..1391
                     /note="Alu repeat"
     repeat_region   1814..1852
                     /note="direct degenerate repeat copy A"
     repeat_region   1853..1889
                     /note="direct degenerate repeat copy B"
     mRNA            1935..3667
                     /note="pseudo-beta-globim mRNA"
     CDS             join(1988..2078,2200..2422,3274..3400)
                     /gene="pseudo-beta-globin"
                     /pseudo
                     /codon_start=1
     exon            2200..2422
                     /gene="pseudo-beta-globin"
                     /pseudo
                     /number=2
BASE COUNT     2215 a   1315 c   1439 g   2086 t
ORIGIN      1 bp upstream of EcoRI site.
        1 gaattcctgg ttggctgatg gaagatgggg caactgttca ctggtatgca gggttttaga
       61 tgtatgtacc taaggatatg aggtatggca atgaacagaa attcttttgg gaatgagttt
      121 tagggccatt aaaggacatg acctgaagtt tcctctcagg ccagtcccca caactcaata
```

```
 181 taaatgtgtt tcctgcatac agtcaaagtt gccacttctt tttcttcata tcatcgatct
 241 ctgctcttaa agataatctt ggttttgcct caaactgttt gtcactacaa actttcccca
 301 tgttcctaag taaaacagat aactgcctct caactatatc aagtagacta aaatattgtg
 361 tctctaatat cagaaattca gctttaatat attgggttta actctttgaa atttagagta
 421 tccttgaaat acacatgggg gtgatttcct aaactttatt tcttgtaagg atttatctca
 481 ggggtaacac acaaaccagc atcctgaacc tctaagtatg aggacagtaa gccttaagaa
 541 tataaaataa actgttattc tctctgccgg tgcaagtgcg ccctgtctat tcctgaaatt
 601 gctcgtttga gacgcatgag acgtgcagca catgagacac gtgcagcagc ctgtggaata
 661 ttgtcagtga agaatgtctc tgcctgatta gatataaaga caagttaaac acagcattag
 721 actatagctc aagcctgtgc cagacacaaa tgacctaatg cccagcatgg gccatggaat
 781 ctcctatcct cttgcttgaa cagagcagca cacttctccc ccaacactat tagatgttct
 841 ggcataattt tgtagatatg taggatttga catggactat tgttcaatga ttcagaggaa
 901 atctcctttg ttcagataag tacactgact actaaatgga ttaaaaaaca cagtaataaa
 961 acccagtttt ccccttattt ccctagtttg tttcttattc tgctttcttc caaattgatg
1021 ctggatagag gtgtttattt ctattctaaa aagtgatgaa attggccggg cgcggtggct
1081 cacacctgta atcccagcac tttgggaggc tgaggtgggc ggatcacgag gtcaggagat
1141 caagaccatc ctggctaaca tggtgaaacc ccatctctac taaaaataca aaaaaattag
1201 ccagagacgg tggcgggtgc ctgtagtccc agctactcgg gaggctgagg caggagaatg
1261 gtgtgaacct gggaggcaga gcttgcagtg agcagagatc gtgccactgc acactccagc
1321 ctgggtgaca aagcaagact ccatctcaaa aaaaaaaaaa aaaagaaaga aagaaagaaa
1381 gaaaaataaa ggtgatgaaa ttgtgtattc aatgtagtct caagagaatt gaaaaccaag
1441 aaaggctgtg gcttcttcca cataaaacct ggatgaataa caggataaca cgtcgttaca
1501 ttgtcacaac tcctgatcca ggaattgatg gctaagatat tcgtaattct tatccttttc
1561 agttgtaact tattcctatt tgtcagcatt caggttatta gcggccgctg gcgaagtcct
1621 tgagaaataa actgcacact ggacggtggg gatagcgtag gaaaatggag gggaaggaag
1681 taaagttcca aattaagcct gaacagcaaa gttcccctga gaaggccacc tggattctat
1741 cagaaactcg aatgtccatc ttgcaaaact tccttgccca aaccccaccc ctggagtcac
1801 aacccaccct tgaccaatag attcatttca ctaagagaag caaagggctg gtcaatggat
1861 tcatttcact gggagaggca aagggctggg ggccagagag gagaagtaaa aagccacaca
1921 tgaagcagca atgcaggcat gtttctggct catctgtgat caccaggaaa ctcccagatc
1981 tgacactgta gtgcatttca ctgctgacaa gaaggctgct gccaccagcc tgtgaagcaa
2041 ggttaaggtg agaaggctgg aggtgagatt ctgggcaggt aggtactgga agccggggca
2101 aggtgcagaa aggcagaaag tgtttctgaa agagggatta gcccattgtc ttacatagtc
2161 tgactttgca cctgctctgt gattatgact atcccacagt ctcctggttg tctacccatg
2221 gacctagagg tactttgaaa gttttggata tctgggctct gactgtgcaa taatgggcaa
2281 ccccaaagtc aaggcacatg gcaagaaggt gctgatctcc ttcggaaaag ctgttatgct
2341 cacggatgac ctcaaaggca cctttgctac gctgagtgac ctgcactgta acaagctgca
2401 cgtggaccct gagaacttcc tggtgagtac taagtacact cacactttct tctttaccct
2461 tagatatttg cactatgggc acttttgaaa gcagaggtgg ctttctcttg tgttatgagt
2521 cagctgtggg atataatatt tcagcagtgg gattttgaga gttatgttgc tgtaaataac
2581 ataactaaaa tttggtagag caaggactac gaataatgga aggccactta ccatttgata
2641 gctctgaaaa acacatctta taaaaaattc tggccaaaat caaactgagt gtttttggat
2701 gagggaacag aagttgagat agagaaaata acatctttcc tttggtcagc gaaattttct
2761 ataaaaatta atagtcactt ttcttcatag tcctggaggt tagaaaaaga tcaactgaac
2821 aaagtagtgg gaagctgtta aaaagaggat tgtttccctc ctaatgatga tggtatactt
2881 ttgtacgcat ggtacaggat tctttgttat gagtgtttgg gaaaattgta tgtatgtatg
2941 tatgtgatga ctggggactt atcctatcca ttactgttcc ttgaagtact attatcctac
3001 tttttaaaag gacgaagtct ctaaaaaaaa tgaaacaatt aatcacaata tgctggggta
3061 gtgagttggc atagcaagta agagaaggat aggacacaat gggaggtgca gggctgccag
3121 tcatattgaa gctgatatct agcccataat ggtgagagtt gctcaaactc tggtcaaaaa
3181 ggatgtaagt gttatatcta tttactgcaa gtccagcttg aggccttcta ttcactatgt
3241 accattttct ttttatcttc actccctccc cagctcttag gcaacgtgat attgattgtt
3301 ttggcaaccc acttcagcga ggagtttacc ctacagatac aggcttcctg gcagtaacta
3361 acaaatgctg tggttaatgc tgtagcccac aagaccactg agtcccctgt ccactatgtt
3421 tgtacctact ggtccactat gtttgtacct atgtccccaa atctcatctc ctttagatgg
3481 gggaggttgg ggagaagagc agtatcctgc ctgctgattc agttcctgca tgataaaaat
3541 aaaataaaga aatatgctct ctaagaaata tcattgtatt ctttttctgt ctttatattt
3601 taccctgatt cagccaaaag gacgcaccat ttctgatgga aatgagaatg ttggagaatg
3661 ggagcttaag gacagagaag atactttctt gcaatcctgc aagaaaagag agaacttgtg
3721 ggttgattta gtggggtagt tactcctagg aaggggaaat cgtctctaga ataagacaat
3781 gtctttacag aaagggaggt caatggaggt actctttgga gatgtaagag gattgttggt
3841 agtgtgtaga ggtatgttag gactcaaatt agaagttctg tataggctat tatttgtatg
3901 aactcaggat acagctcatt tggtgactgc agttcacttc tacttatttt gaacaacata
3961 ttttttatga cttataatga agtggggatg gggcttccta gagaccaatc aagggccaaa
4021 ccttgaactt tctcttaacg tcttcaatgg tattaataga gaattatctc taaggcatgt
4081 gaactggctg tcttggtttt catctgtact tcatctgcta cctctgtgac ctgaaacata
4141 tttataattc cattaagctg tacatatgat agatttatca tatttatttt ccttaaagga
4201 tttttgtaag aacgaattga attgatacct gtaaagtctt tatcacacta cccgataaat
4261 aataaatctc tttgttcagc tctctgtttc tataaatatg tacaagtttt attgttttta
4321 gtggtagtga ttttattctc tttctatata tatatataca catatgtgtg cattcataaa
4381 tatatacaat ttttatgaat aaaaaattat tagcaatcaa tattgaaaac cactgatttt
4441 tgtttatgtg agtaaacagc agattaaaag gctgagattt aggaaacagc acgttaagtc
4501 aagttgatag aggagaatat ggacatttaa aagaggcagg atgatataaa attagggaaa
4561 ctggatgcag agaccagatg aagtaagaaa aatagctatt gttttgagca aaaatcactg
4621 aagtttctcg catatgagag tgacataata aatagggaaa tgtagaaaat tgattcacat
4681 gtatatatat atatagaact gattagacaa agtctaactt gggtatagtc agaggagctt
4741 gctgtaatta tattgaggtg atggataaag aactgaagtt gatggaaaca atgaagttaa
4801 gaaaaaaaat cgagtaagag accactgtgg cagtgattgc acagaactgg aaaacactgt
4861 gaaacagaga gtcagagatg acagctaaaa tccctgcctg tgaatgaaaa gaaggaaatt
4921 tattgacaga acagcaaatg cctacaagcc ccctgtttgg atctggcaat gaacatagtc
```

```
     4981 attctgtggc aatcacttca aactcctgta cccaagaccc ttaggaagta tgtagcaccc
     5041 tcaaacctaa aacctcaaag aaagaggttt tagaagatat aatatccttt cttctccagt
     5101 ttcattaatc cccaagcctc tttctcaaag tatttcctct atgtgtccac cccaaagagc
     5161 tcacctcacc atatctcttg agtgggagca catagatagg cggtgctacc atctgacagc
     5221 ttctgaaatt cctttgtcat atttttgagt ccccactaat aacccacaaa gcagaataaa
     5281 taacagttgc tcatgtacaa taatcactca actgctgtct tgtagcatac attaattaag
     5341 cacattcttt gaataattac tgtgtccaaa caatcacact ttaaaatctc acacttatgc
     5401 tatcccttgc ccttctgaat gtcactctgt attttaaatt aagagaggag ggttgaattt
     5461 cctgtgttac ttattgttca tttctcgatg aggagttttc acattcacct ttactggaaa
     5521 acacataagc acacatctta caggaaaaat ataccaaact gacatgtagc atgaatgctt
     5581 gtgcatgtag tcatataaaa tcttgtagca atgtaaacat tctctgatat acacatacag
     5641 atgtgtctat atgtctacac aatttcttat gctccatgaa caaacattcc atgcacacat
     5701 aagaacacac actgttacag atgcatactt gagtgcattg acaaaattac cccagtcaat
     5761 ctagagaatt tggatttctg catttgactc tgttagcttt gtgcatgctg ttcatttact
     5821 ctgggtgatg tctttccctc attttgccct gtctatcttg tactcatact ttaagtccta
     5881 acttatatgc tatctcaatt aagaagctat ttttttttaa attttaactg ggcttaaagc
     5941 cctgtctata aactctgcta caattatggg ctctttctta taatatttag tgtttttcct
     6001 actaatgtac ttaatctgct cattgtatct tcctaccact aaattttaac ctcatttatg
     6061 gtagagacat tgtcttgtaa actcttattt ccctagtatt tggagatgaa aaaaaaagat
     6121 taaattatcc aaaattagat ctctctttтc tacattatga gtattacact atccatagag
     6181 aagtttgttt gagacctaaa ctgaggaacc tttggttcta aaatgactat gtgatatctt
     6241 agtatttgta ggtcatgagg ttccttcctc tgcctctata aggctgttcc ctcaatctcc
     6301 cttgctatag tttgattagt caacaagcat gtgtcatgca tttattcaca tcagaatttt
     6361 catacactaa taagacatag tatcagaagt cagtttatta gttatatcag ttagggtcca
     6421 tcaaggaaag gacaaaccat tatcagttac tcaacctaga attaaataca gctcttaata
     6481 gttaattatc cttgtattgg aagagctaaa atatcaaata aaggacagtg cagaaatcta
     6541 gatgttagta acatcagaaa acctcttccg ccattaggcc tagaagggca gaaggagaaa
     6601 atgtttatac caccagagtc cagaaccaga gaccataacc agaggtccac tggattcagt
     6661 gagctagtgg gggctccttg gagagagcca gaactaggtg tctaatgggg gtatcaaaat
     6721 atcagccata aaaaaccata aaaaagactg tctgctgtag gagatctgtt cagagagaga
     6781 gagagacaag aaataatctt gcttatgctt tccctcagcc agtgtttacc actgcagaat
     6841 gtacatgcaa ctgaaagggt gaggaaacct gggaaatgtc agttcctcaa atacagagaa
     6901 cactgaggga aggatgagaa ataaatttga aagcagacat gaatggtaat tgacagaagg
     6961 aaactaggat gtgtccagta aatgaataat tacagtgtgc agtgattatt gcaatgatta
     7021 atgtattgat aagataatat gaaaacacag aattc
//
```

This is still a very short entry. Many of the entries originate from complete genomes (or exceptionally long contigs that when joined together represent a complete genome). These can of course be many millions of nucleotides long.

As of May 2015, there were one hundred seventy eight eukaryotic organisms listed as completely sequenced at `http://www.ebi.ac.uk./genomes/eukaryota.html`. These include the "lab-rat" yeast *Saccharomyces cerevisiae* and other fungi, single celled protists, the nematodes *Caenorhabditis elegans* (and *C. briggsae*), the plants *Arabidopsis thaliana* and *Oryza sativa*, twelve complete fruit fly genomes (including the ubiquitous *Drosophila melanogaster*), the mosquito *Anopheles gambiae*, fish (*Danio rerio* and *Tetraodon nigroviridis*), the chicken *Gallus gallus*, mouse *Mus musculus*, the rat *Ratuus norvegicus*, man's best friend *Canis familiaris*, the Rhesus macaque *Macaca mulatta*, the chimpanzee *Pan troglodytes* and *Homo sapiens*. Many particularly interesting genomes have been done such as that of the duck-billed platypus (*Ornithorhynchus anatinus*; Nature 2008 453:175-83). However this listing is horribly out of date, many completed genomes are not listed at all and the listing has not been updated since May 2015. The humans maintaining the database entries can't keep up with the automated sequencing machines.

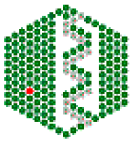
The prokaryotic genomes are sequenced more rapidly since they are much smaller and as of August 2015, there have been at least 202 archaeal and 3316 bacterial genomes completely sequenced (the first being *Haemophilus influenzae* in 1995 coming in at 1,830,140 bp.) and there are many more being sequenced and many new genomes completely sequenced now don't even make it into the databases.

In the past, full genome sequences were reported as major milestones of achievement in journals such as SCIENCE and NATURE. In the last 1.5 months of 1997, the journal NATURE published five issues. Of these, three issues reported the completion of three different bacterial genomes (*Bacillus subtilis, Archaeoglobus fulgidus*, and *Borrelia burgdorferi*). Today, only a few genomes are still reported in high profile journals with others published in more specialized journals and some are simply published online. The rate of genome completion is rapidly speeding up and as more genomes are completed, the news worthiness of a single genome (or even dozens) tends to diminish. But their utility increases with each one determined.

In addition to the complete genomes, there is a long list of viruses and organelles (e.g. see also the OGMP - organelle genome megasequencing program) that have been completely sequenced including the CMV DNA virus (300,000 bp), the Epstein-Barr virus genome (172,282 bp), the AIDS virus (9,737 bp), human mitochondria (16,569 bp), human leukaemia virus type I (9,032 bp), lambda (48,502 bp), PhiX174 (5,386 bp) and more.

For more information about NCBI go to their web site.

4.3 E.M.B.L.



The same entry as in Example #1 above is also present at EMBL (as they all should be). It's format is shown in Example #3.

Example Entry #3: EMBL - HSCKRA01

```
ID   HSCKRA01   standard; RNA; HUM; 450 BP.
XX
AC   M35504;
XX
SV   M35504.1
XX
DT   26-NOV-1990 (Rel. 25, Created)
DT   04-MAR-2000 (Rel. 63, Last updated, Version 3)
XX
DE   Human PR310 c-K-ras protein mRNA, 5' end.
XX
KW   c-K-ras oncogene; c-myc proto-oncogene.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Primates; Catarrhini; Hominidae; Homo.
XX
RN   [1]
RP   1-450
RX   MEDLINE; 85271309.
RA   Yamamoto F., Nakano H., Neville C., Perucho M.;
RT   "Structure and mechanisms of activation of c-K-ras oncogenes in human lung
RT   cancer";
RL   Prog. Med. Virol. 32:101-114(1985).
XX
DR   GOA; Q14014; Q14014.
DR   SPTREMBL; Q14014; Q14014.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..450
FT                   /db_xref="taxon:9606"
FT                   /organism="Homo sapiens"
FT   CDS             1..>450
FT                   /codon_start=1
FT                   /db_xref="GOA:Q14014"
FT                   /db_xref="SPTREMBL:Q14014"
FT                   /note="PR310 c-K-ras oncogene"
FT                   /protein_id="AAA35689.1"
FT                   /translation="MTEYKLVVVGAGGVGKSALTIQLIDNHFVDEYDPTIEDSYRKQVV
FT                   IDGETCLLDILDTAGHEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVK
FT                   DSEDVPMVLVGNKCDLPSRTVDTKQAQDLARSYGIPFIQTSAKTRQ"
XX
SQ   Sequence 450 BP; 155 A; 71 C; 106 G; 118 T; 0 other;
     atgactgaat ataaacttgt ggtagttgga gctggtggcg taggcaagag tgccttgacg        60
     atacagctaa ttgacaatca ttttgtggac gaatatgatc caacaataga ggattcctac       120
     aggaagcaag tagtaattga tggagaaacc tgtctcttgg atattctcga cacagcaggt       180
     catgaggagt acagtgcaat gagggaccag tacatgagga ctggggaggg ctttctttgt       240
     gtatttgcca taaataatac taaatcattt gaagatattc accattatag agaacaaatt       300
     aaaagagtta aggactctga agatgtacct atggtcctag taggaaataa atgtgatttg       360
     ccttctagaa cagtagacac aaaacaggct caggacttag caagaagtta tggaattcct       420
     tttattcaaa catcagcaaa gacaagacag                                        450
//
```

Note that the entry contains the same information but in a slightly different form. In this case, the data is more structured with defined prefixes at the beginning of every line. This difference can be useful if you wish to write your own code to analyze some features of this data.

The EMBL databases have moved from Heidelberg, Germany to Hinxton Hall (Cambridge, England). But many of the people doing protein analysis still exist at Heidelberg and throughout the EMBL organization. For more information about the EMBL database check out their web site or send e-mail to `netserv@ebi.ac.uk`.

4.4 D.D.B.J.

For the particular entry chosen in Example #1, the DDBJ format is essentially equivalent (there are minor differences). It is ...

Example Entry #4: DDBJ - HUMCKRASA

```
LOCUS       HUMCKRASA     450 bp ss-mRNA            PRI       15-SEP-1990
DEFINITION  Human PR310 c-K-ras protein mRNA, 5' end.
ACCESSION   M35504
KEYWORDS    c-K-ras protein; c-myc oncogene.
SOURCE      Human (patient PR310) lung carcinoma, cDNA to mRNA.
  ORGANISM  Homo sapiens
            Eukaryota; Animalia; Metazoa; Chordata; Vertebrata; Mammalia;
            Theria; Eutheria; Primates; Haplorhini; Catarrhini; Hominidae.
REFERENCE   1  (bases 1 to 450)
  AUTHORS   Yamamoto,F., Nakano,H., Neville,C. and Perucho,M.
  TITLE     Structure and mechanisms of activation of c-K-ras oncogenes in
            human lung cancer
  JOURNAL   Prog. Med. Virol. 32, 101-114 (1985)
  STANDARD  simple staff_entry
FEATURES             Location/Qualifiers
     CDS             1..>450
                     /note="PR310 c-K-ras oncogene"
                     /codon_start=1
BASE COUNT      155 a     71 c    106 g    118 t
ORIGIN
        1 atgactgaat ataaacttgt ggtagttgga gctggtggcg taggcaagag tgccttgacg
       61 atacagctaa ttgacaatca ttttgtggac gaatatgatc caacaataga ggattcctac
      121 aggaagcaag tagtaattga tggagaaacc tgtctcttgg atattctcga cacagcaggt
      181 catgaggagt acagtgcaat gagggaccag tacatgagga ctggggaggg ctttctttgt
      241 gtatttgcca taaataatac taaatcattt gaagatattc accattatag agaacaaatt
      301 aaaagagtta aggactctga agatgtacct atggtcctag taggaaataa atgtgatttg
      361 ccttctagaa cagtagacac aaaacaggct caggacttag caagaagtta tggaattcct
      421 tttattcaaa catcagcaaa gacaagacag
//
```

The DDBJ began in 1986 and is operated from grants from the Japanese Ministry of Education, Science and Culture. For more information go to their web site.

4.5 SwissProt

These are the three major nucleotide databases, but there are also a large number of protein sequence databases. Again many of these databases are very large. For example, release 2018_07 of UniProtKB/Swiss-Prot (18 July 2018) contains 557,992 annotated entries containing a total of 200,270,360 amino acid residues. There are more than 1,269 entries having proteins larger than 2500 residues including the absolutely massive human nebulin protein of 6669 amino acids, human nesprin 1 protein of 8797 amino acids, the Caenorhabditis elegans mesocentin protein of 13100 amino acids and the mouse titin protein of 35213 amino acids (currently the largest protein in SWISS-PROT). The entries in this database are similar to the nucleotide databases of EMBL. Two examples are shown below.

Example Entry #5: SWISS-PROT - ACYO_HUMAN

```
ID   ACYO_HUMAN     STANDARD;      PRT;    98 AA.
AC   P07311;
DT   01-APR-1988 (REL. 07, CREATED)
DT   01-APR-1988 (REL. 07, LAST SEQUENCE UPDATE)
DT   01-NOV-1995 (REL. 32, LAST ANNOTATION UPDATE)
DE   ACYLPHOSPHATASE, ORGAN-COMMON TYPE ISOZYME (EC 3.6.1.7)
DE   (ACYLPHOSPHATE PHOSPHOHYDROLASE) (ACYLPHOSPHATASE, ERYTHROCYTE
DE   ISOZYME).
OS   HOMO SAPIENS (HUMAN).
OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC   EUTHERIA; PRIMATES.
RN   [1]
RP   SEQUENCE.
RX   MEDLINE; 87101109.
RA   LIGURI G., CAMICI G., MANAO G., CAPPUGI G., NASSI P., MODESTI A.,
RA   RAMPONI G.;
RL   BIOCHEMISTRY 25:8089-8094(1986).
CC   -!- FUNCTION: ITS PHYSIOLOGICAL ROLE IS NOT YET CLEAR.
CC   -!- CATALYTIC ACTIVITY: AN ACYLPHOSPHATE + H(2)O = A FATTY ACID ANION
CC       + ORTHOPHOSPHATE.
CC   -!- TISSUE SPECIFICITY: ORGAN-COMMON TYPE ISOZYME IS FOUND IN MANY
CC       DIFFERENT TISSUES.
CC   -!- SIMILARITY: HIGH, WITH ORGAN-COMMON TYPE ACYLPHOSPHATASES. LESS
CC       WITH MUSCLE TYPE ACYLPHOSPHATASES.
DR   PIR; A25587; QPHUE.
DR   HSSP; P00818; 1APS.
DR   PROSITE; PS00150; ACYLPHOSPHATASE_1.
DR   PROSITE; PS00151; ACYLPHOSPHATASE_2.
KW   HYDROLASE; ACETYLATION; MULTIGENE FAMILY.
FT   MOD_RES       1      1       ACETYLATION.
SQ   SEQUENCE   98 AA;  11130 MW;  51080 CN;
     AEGNTLISVD YEIFGKVQGV FFRKHTQAEG KKLGLVGWVQ NTDRGTVQGQ LQGPISKVRH        60
     MQEWLETRGS PKSHIDKANF NNEKVILKLD YSDFQIVK                              98
//
```

Example Entry #6: SWISS-PROT - CD25_YEAST

```
ID   CC25_YEAST     STANDARD;      PRT;  1589 AA.
AC   P04821;
DT   13-AUG-1987 (REL. 05, CREATED)
DT   01-JAN-1988 (REL. 06, LAST SEQUENCE UPDATE)
DT   01-NOV-1995 (REL. 32, LAST ANNOTATION UPDATE)
DE   CELL DIVISION CONTROL PROTEIN 25.
GN   CDC25 OR CTN1.
OS   SACCHAROMYCES CEREVISIAE (BAKER'S YEAST).
OC   EUKARYOTA; FUNGI; ASCOMYCOTINA; HEMIASCOMYCETES.
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 87131091.
RA   BROEK D., TODA T., MICHAELI T., LEVIN L., BIRCHMEIER C., ZOLLER M.,
RA   POWERS S., WIGLER M.;
RL   CELL 48:789-799(1987).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE; 86220116.
RA   CAMONIS J.H., KALEKINE M., GONDRE B., GARREAU H., BOY-MARCOTTE E.,
RA   JACQUET M.;
RL   EMBO J. 5:375-380(1986).
RN   [3]
RP   DOMAINS.
RX   MEDLINE; 89181526.
RA   MUNDER T., MINK M., KUNTZEL H.;
RL   MOL. GEN. GENET. 214:271-277(1988).
RN   [4]
RP   FUNCTION.
RX   MEDLINE; 91203884.
RA   JONES S., VIGNAIS M.L., BROACH J.R.;
RL   MOL. CELL. BIOL. 11:2641-2646(1991).
CC   -!- FUNCTION: PROMOTES THE EXCHANGE OF RAS-BOUND GDP BY GTP. THIS
CC       PROTEIN POSITIVELY CONTROLS THE LEVEL OF CELLULAR CAMP AT START,
CC       THE STAGE AT WHICH THE YEAST CELL DIVISION CYCLE IS TRIGGERED.
CC   -!- SIMILARITY: CONTAINS A COPY OF THE SH3 DOMAIN.
CC   -!- SIMILARITY: TO OTHER GUANINE-NUCLEOTIDE RELEASING FACTORS OF THE
CC       CDC25 FAMILY.
```

```
DR   EMBL; X03579; X03579.
DR   EMBL; M15458; M15458.
DR   PIR; A26596; RGBYC5.
DR   HSSP; P00519; 1ABL.
DR   LISTA; SC00152; CDC25.
DR   SGD; L0000263; CDC25.
DR   PROSITE; PS00720; GDS_CDC25.
DR   PROSITE; PS50002; SH3.
KW   GUANINE-NUCLEOTIDE RELEASING FACTOR; CELL DIVISION; CELL CYCLE;
KW   MITOSIS; TRANSMEMBRANE; SH3 DOMAIN.
FT   TRANSMEM   1452   1473       POTENTIAL.
FT   DOMAIN       58    128       SH3.
FT   CONFLICT    497    497       I -> Y (IN REF. 2).
FT   CONFLICT    954    963       PVGHHEPFKN -> LSVIMNLSR (IN REF. 2).
SQ   SEQUENCE   1589 AA;  179091 MW;  13488958 CN;
     MSDTNTSIPN TSSAREAGNA SQTPSISSSS NTSTTTNTES SSASLSSSPS TSELTSIRPI        60
     GIVVAAYDFN YPIKKDSSSQ LLSVQQGETI YILNKNSSGW WDGLVIDDSN GKVNRGWFPQ       120
     NFGRPLRDSH LRKHSHPMKK YSSSKSSRRS SLNSLGNSAY LHVPRNPSKS RRGSSTLSAS       180
     LSNAHNAETS SGHNNTVSMN NSPFSAPNDA SHITPQSSNF NSNASLSQDM TKSADGSSEM       240
     NTNAIMNNNE TNLQTSGEKA GPPLVAEETI KILPLEEIEM IINGIRSNIA STWSPIPLIT       300
     KTSDYKLVYY NKDLDIYCSE LPLISNSIME SDDICDSEPK FPPNDHLVNL YTRDLRKNAN       360
     IEDSSTRSKQ SESEQNRSSL LMEKQDSKET DGNNNSINDD DNNNENNKNE FNEAGPSSLN       420
     SLSAPDLTQN IQSRVVAPSR SSILAKSDIF YHYSRDIKLW TELQDLTVYY TKTAHKMFLK       480
     ENRLNFTKYF DLISDSIVFT QLGCRLMQHE IKAKSCSKEI KKIFKGLISS LSRISINSHL       540
     YFDSAFHRKK MDTMNDKDND NQENNCSRTE GDDGKIEVDS VHDLVSVPLS GKRNVSTSTT       600
     DTLTPMRSSF STVNENDMEN FSVLGPRNSV NSVVTPRTSI QNSTLEDFSP SNKNFKSAKS       660
     IYEMVDVEFS KFLRHVQLLY FVLQSSVFSD DNTLPQLLPR FFKGSFSGGS WTNPFSTFIT       720
     DEFGNATKNK AVTSNEVTAS SSKNSSISRI PPKMADAIAS ASGYSANSET NSQIDLKASS       780
     AASGSVFTPF NRPSHNRTFS RARVSKRKKK YPLTVDTLNT MKKKSSQIFE KLNNATGEHL       840
     KIISKPKSRI RNLEINSSTY EQINQNVLLL EILENLDLSI FINLKNLIKT PSILLDLESE       900
     EFLVHAMSSV SSVLTEFFDI KQAFHDIVIR LIMTTQQTTL DDPYLFSSMR SNFPVGHHEP       960
     FKNISNTPLV KGPFHKKNEQ LALSLFHVLV SQDVEFNNLE FLNNSDDFKD ACEKYVEISN      1020
     LACIIVDQLI EERENLLNYA ARMMKNNLTA ELLKGEQEKW FDIYSEDYSD DDSENDEAII      1080
     DDELGSEDYI ERKAANIEKN LPWFLTSDYE TSLVYDSRGK IRGGTKEALI EHLTSHELVD      1140
     AAFNVTMLIT FRSILTTREF FYALIYRYNL YPPEGLSYDD YNIWIEKKSN PIKCRVVNIM      1200
     RTFLTQYWTR NYYEPGIPLI LNFAKMVVSE KIPGAEDLLQ KINEKLINEN EKEPVDPKQQ      1260
     DSVSAVVQTT KRDNKSPIHM SSSSLPSSAS SAFFRLKKLK LLDIDPYTYA TQLTVLEHDL      1320
     YLRITMFECL DRAWGTKYCN MGGSPNITKF IANANTLTNF VSHTIVKQAD VKTRSKLTQY      1380
     FVTVAQHCKE LNNFSSMTAI VSALYSSPIY RLKKTWDLVS TESKDLLKNL NNLMDSKRNF      1440
     VKYRELLRSV TDVACVPFFG VYLSDLTFTF VGNPDFLHNS TNIINFSKRT KIANIVEEII      1500
     SFKRFHYKLK RLDDIQTVIE ASLENVPHIE KQYQLSLQVE PRSGNTKGST HASSASGTKT      1560
     AKFLSEFTDD KNGNFLKLGK KKPPSRLFR                                        1589
//
```

Again, various features are on individual lines – the identification line (ID) giving a locus name, the accession number line (AC), the date of entry (DT), a description of the entry (DE), a line specifying the organism (OS), the organism's phylogenetic classification (OC), lines describing the reference number, author and location (RN, RA, RL), the comment lines (CC), a database reference line (DR) to cross reference the entry to other database entries, the keyword line (KW), the feature tables (FT) and the sequence header (SQ) giving length in aa, molecular weight and a checking number defined in A. Bairoch, J. Biochem. 203: 527 (1983).

In addition to these protein databases, there are databases devoted to particular families of proteins and to particular organisms. In addition there are protein databases constructed from translations of the nucleotide databases – NCBI's is called GenPept and EMBL's is termed TrEMBL (release 2018_07 of UniProtKB/TrEMBL (18 July 2018) combines the translated EMBL nucleotide database, the Genbank database and Swiss-Prot to yield 120,243,849 sequence entries comprising 40,506,871,635 amino acids). The best access for the SwissProt database is through the UniProtKB web site (or the ExPASy (Expert Protein Analysis System) web site).

Another protein sequence database of interest is PIR (Protein Information Resource) sponsored by NBRF (National Biomedical Research Foundation) at Georgetown University. This database of protein sequences is completely cross-referenced to known nucleic acid sequences, has data on x-ray crystallography and active site determination, and is fully annotated. The last release of this database was on Dec 2004 as it is now integrated into UniProt.

In an effort to combine the information in these disparate protein databases, the UniProt database was constructed. It joins the information contained in Swiss-Prot, TrEMBL, and PIR. UniProt (Universal Protein Resource) claims to be the world's most comprehensive catalog of information on proteins. It is divided into three parts: UniProt Knowledgebase (UniProtKB) is the central access point for extensive curated protein information, including function, classification, and cross-reference. The UniProt Reference Clusters (UniRef) databases combine closely related sequences into a single record. The UniProt Archive (UniParc) is meant to be a comprehensive repository for all protein sequences.

Table 4.1: Sequence by Organism (millions of nucleotides)

Species	EMBL rel 124 Jun 2015	EMBL rel 104 Jul 2010	EMBL rel 83 Aug 2005	EMBL rel 63 Jun 2000	EMBL rel 44 Aug 1995	IG rel 63 Jun 1990
Environment Samples	94835	11356	112	-	-	-
Human	61467	33894	4158	825	-	-
Primate	-	-	-	-	41	2
Mus musculus	23220	14660	2819	-	-	-
Other Rodent	60189	18453	116	85	29	2
Other Mammalian	292145	98710	396	23	8	2
Other Vertebrate	171916	1102	26	9	2	-
Invertebrate	159685	34023	717	326	36	2
Plant	240331	34221	1519	204	17	3
Fungi	36687	6244	251	72	23	-
Organelle	-	-	323	56	11	2
Prokaryotes	175144	11192	1033	189	43	2
Viral	2510	990	268	82	28	2
Bacteriophage	204	42	16	4	2	1
Unclassified	5861	2808	2	2	5	0
Transgenic	859	-	-	-	-	-
Synthetic	1595	633	31	9	5	1
Constructed	864785	-	-	-	-	-
ESTs	42497	36234	14544	1641	100	-
GSSs	25423	18322	7155	838	-	-
HTC	629	658	422	-	-	-
HTG	25528	24132	11490	3756	-	-
HTG0	-	-	510	-	-	-
Patents	15272	8191	1450	-	-	-
Standard	72562	-	-	-	-	-
STSs	640	634	492	51	6	-
TSAs	65576	-	-	-	-	-
WGS	1078521	172426	-	-	-	-

4.6 Organization of the entries

The entries can also be grouped according to their organismal affiliation. To give you an example of how this data is distributed by organism consider the data in Table 4.1 (this table excludes whole genome shotgun data). Several groups require further explanation. The large Unannotated group contains sequence entries that have not yet had things like the FEATURE table or other niceties added to the entry. The Synthetic group contains things such as plasmids, Yacs, etc that have been constructed by researchers. The ESTs (**e**xpressed **s**equence **t**ags) are short sequences derived from cDNAs. That is messenger RNA is reverse transcribed to cDNA, and the cDNA is partially sequenced from the the poly-A addition site. The STSs (**s**equence **t**agged **s**ites) and the GSSs (**g**enome **s**urvey **s**equences) are short sequences that can serve to map various regions. The HTC stands for **h**igh **t**hroughput **c**DNA sequences. The HTG stands for **h**igh **t**hroughput **g**enomic sequences. These are contig sequences greater than 2kb that have not yet been fully released from large scale sequencing laboratories. The HTG0 stands for **h**igh **t**hroughput **g**enomic sequences but with one-to-few pass reads of a single clone and the WGS stands for **w**hole **g**enome **s**hotgun sequences. The sequences will normally be checked, assembled and annotated by these groups before official release. Beware these sequences may therefore contain more than their share of sequencing errors.

These data can be compared to previous years - the columns give the corresponding EMBL data from 2015, 2010, 2005, 2000, and 1995 while the last column on the right gives GenBank/Intelligenetics data from release 63 in June of 1990. The fastest growing groups (excluding the entries at the bottom) were plants, invertebrates and human/mammals. The slowest

Table 4.2: The Most Sequenced Organisms (DDBJ:release 92, March 2013 vs. GB:Rel. 104 1997)

	Species	Nucleotides	Nucleotides	Species
1	*Homo sapiens*	16905195941	551650819	*Homo sapiens*
2	*Mus musculus*	9979520722	133404600	*Mus musculus*
3	*Rattus norvegicus*	6523318591	113659949	*Caenorhabditis elegans*
4	*Bos taurus*	5387379177	46566672	*Arabidopsis thaliana*
5	*Zea mays*	5066583840	29198918	*Drosophila melanogaster*
6	*Sus scrofa*	4889169499	28567944	*Saccharomyces cerevisiae*
7	Unknown	4814890256	17345841	*Escherichia coli*
8	*Danio rerio*	3119551273	13826989	*Rattus norvegicus*
9	marine metagenome	2149495444	9715721	*Bacillus subtilis*
10	uncultured bacterium	2101091113	9198129	Human immunodeficiency virus type 1
11	*Vitis vinifera*	1554924461	8932632	*Oryza sativa*
12	*Hordeum vulgare*	1452393117	7522232	*Fugu rubripes*
13	*Strongylocentrotus purpuratus*	1435236537	6398746	*Schizosaccharomyces pombe*
14	*Macaca mulatta*	1256365166	5099500	*Gallus gallus*
15	*Xenopus tropicalis*	1249967444	4509620	*Mycobacterium tuberculosis*
16	*Oryza sativa*	1205152636	4350356	*Toxoplasma gondii*
17	*Nicotiana tabacum*	1198652893	4312388	*Brugia malayi*
18	*Arabidopsis thaliana*	1151239675	4164811	*Bos taurus*
19	*Triticum aestivum*	1136410179	3829215	*Synechocystis* sp.
20	*Drosophila melanogaster*	1123080930	3160564	*Xenopus laevis*

growing groups were fungi sequences. It is amazing to see the growth reported in just a few years.

The top twenty organisms (excluding chloroplast and mitochondrial sequences) as of Jun 2012 are shown in Table 4.2. Note that humans have already been done several times over. Also note that bacterial sequences can no longer make it among the top twenty. Even with multiple sequencing of their genomes, they do not contain the quantity of sequence that is present in eukaryotes. However, there are three unusual entries that now make it into the top 20; the "marine metagenome", "uncultured bacterium" and "unknown". These are sequences that arise from what is known as environmental sequencing. In these analysis, DNA is simply isolated from sea water or from soil or from some other natural environment. It is isolated as raw DNA without concern from whatever organism it originates from and is simply sequenced. These sequences can then be built into contigs and whole organisms can be sequenced without ever knowing what you have actually sequenced! Since some organisms are more dominant over others in sea water, their genomes are sequenced many times.

Besides the data itself, most of the databases also maintain various index files. These include indices of authors, journals, organism, etc. all cross-referenced by accession number or locus name. Again these can be very helpful to analyze the entries of the databases.

4.7 Other Major Databases

There are many other databases of note and not all can be covered here. One of database of importance, though at first glance might be considered boring, is the GO database. This project is, in my mind, somewhat akin to the huge project that eventually became the Oxford Dictionary of the English Language. This is not simply a dictionary but rather attempts to be the definitive reference for a language that constantly changes and evolves. For any word, both present or past, it includes the meaning, history, pronunciation, usage, quotations, variants, etymology and so on. The GO database is (will become) the dictionary of gene nomenclature. The example from their website states "if you were searching for new targets for antibiotics, you might want to find all the gene products that are involved in

bacterial protein synthesis, and that have significantly different sequences or structures from those in humans. But if one database describes these molecules as being involved in 'translation', whereas another uses the phrase 'protein synthesis', it will be difficult for you - and even harder for a computer - to find functionally equivalent terms". The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. This is quickly becoming a standard that must be used in the annotations of new genomes.

The PDB (Protein Data Bank), sponsored by Rutgers University, contains the 3-D atomic coordinates from x-ray diffraction or NMR studies. The database also contains secondary and other structural features such as bond connectivity data. The individual database entries are usually directly suitable for entry into 3-D rendering programs.

The PROSITE database is very useful. It lists the distinct structural motifs in proteins. This includes amino acid post-translational modifications, topogenic sequences, domains of specific biological function (e.g. DNA binding domains), enzyme active sites and signature patterns that are specific to a family or group of proteins. For example, it lists the Kringle domain signature as (Y,F)-C-R-N-P-D; a triple-looped, disulfide cross-linked region found mostly in serine proteases and plasma proteins. For more information on this database, send e-mail to the EMBL databases.

Some databases are built on filtered data and are 'value-added' derivatives of these basic sequence databases. For example the COG database holds aligned clusters of orthologous proteins. There are proteins collected from 43 completed genomes and then compared "all-against-all" to yield 3307 clusters of related proteins (as of August 2002). One of the tools associated with this database is the `COGnitor` program, which will assign query proteins to pre-existing clusters (and hence usefully identify its functional category).

Besides these sorts of databases, there are also databases containing different types of information. The GDB (Genome Database) contains mapping information of the human genome project. It contains information on the location of genes, DNA segments, expressed sequence tags (EST's), clinical phenotypes, polymorphisms and alleles, probes, CEPH reference family data markers, etc. As part of this database, Victor McKusick's original **Mendelian Inheritance in Man** (OMIM) has been made available as an online, freely available computer database. This database lists clinical disorders or traits in man, gene names, clinical observations, inheritance patterns, allelic variations, chromosome locations, linkages and so on. The GDB and OMIM and most of the other molecular biology databases are all cross-linked. These databases are maintained by the Welch medical library at Johns Hopkins.

Similar to the GDB, is the database for the mouse genome MGI: Mouse Genome Informatics. This again contains mapping information of much the same sort as the human database. However, it also includes homologies for mice, humans and 23 other species. Thus, if you are interested in a gene on chromosome 11 in mice, you can find out where it has been located in some other species (the references to the papers showing this, what other genes are similarly located and so on).

Pick an organelle and again you will find specialized databases. For just the mitochondria try the comprehensive MITOP web site, or Human Mitochondrial Protein database a database of Mendelian Inheritance and the Mitochondrion db (mitochondrial nuclear genes) or MITOMAP a database for the human mitochondrial genome. If a whole organelle is too large for your tastes, how about picking something smaller like a database for just part of the mitochondria, the hypervariable control region at HvrBase or how about the weird on wonderful inteins. Don't know what inteins are? Check out that link.

Recently there have been more projects to establish databases for the deposition of microarray gene expression data. The NCBI version of this data is housed in Gene Expression Omnibus (GEO). GEO is a gene expression and hybridization array data repository, as well as an online resource for the retrieval of gene expression data from any organism or artificial source. The EBI microarray ArrayExpress Database is a similar database to store and permit the query of microarray experiments.

There are many more databases. I have only given you a taste of some of the major databases. In addition to each of these major databases, there are databases on each of the organisms that have major genomics projects

- E. coli,
- Yeast,
- Arabidopsis,
- Mouse,
- Cattle,
- Drosophila,
- Caenorhabditis elegans,
- Fungi,
- Maize,
- Rice,
- Grasses - Gramene,
- HIV,

and so on. So pick your favorite organism and do a search for it – there will be a web site devoted to it's genome (so long as it is not too unusual).

There are many other databases on diverse aspects such as the

- CEPH-Genethon human physical map data and Genethon database provide a connection between the physical map of the human genome and the genetic/sequence data;
- a human cDNA database;
- genome sequencing centers such as Baylor College of Medicine and the Sanger Center maintain their own databases on projects they are working on;
- an immunogenetics database Immuno Polymorphism Database (IPD);
- the Database of Expressed Sequence Tags dbEST;
- the Database of Sequence Tagged Sites dbSTS;
- the Eukaryotic Promoter Database EPD;
- a database of 3D-diagrams of proteins;
- BMRB (BioMagResBank) a database of NMR Spectroscopy data;
- CCDC (Cambridge Crystallographic Data Centre);
- HIC-Up (Hetero-compound Information Centre Uppsala); a database of small molecules commonly found associated with larger molecules when their 3D-structure is determined
- HIV Drug Database;
- HIV Structural Database;
- Library of Protein Family Cores;
- NDB (Nucleic Acid Database);
- Prolysis: A Protease and Protease Inhibitor Web Server;
- Protein Kinase Resource;
- Protein Motions Database (see Figure 4.4);
- RELIBase;

Figure 4.4: Proteins are not static sequences, they move and there is a database devoted to this subject. The movement of the actin protein is shown here. From the Protein Motions Database.

- SCOP: Structural Classification of Proteins;
- CATH Protein Structure Classification;
- Enzyme Structures Database;
- PDBsum, a database of the known 3D structures of proteins and nucleic acids; etc.

This list goes on and on (and increases each month). Choose what you are interested in ... chances are others are interested as well and have built a database.

4.8 Remote Database Entry retrieval

4.8.1 Entrez

The premier method that should be mentioned is probably the one method that you will use more than any other. This is a NCBI project termed ENTREZ. This program can search across databases or natively through

- PubMed: biomedical literature citations and abstracts
- PubMed Central: free, full text journal articles
- Books: online books
- OMIM: online Mendelian Inheritance in Man
- Site Search: NCBI web and FTP sites
- Nucleotide: sequence database (GenBank)
- Protein: sequence database
- Genome: whole genome sequences

- Structure: three-dimensional macromolecular structures
- Taxonomy: organisms in GenBank
- SNP: single nucleotide polymorphism
- Gene: gene-centered information
- HomoloGene: eukaryotic homology groups
- PubChem Compound: small molecule chemical structures
- PubChem Substance: chemical substances screened for bioactivity
- Genome Project: genome project information
- UniGene: gene-oriented clusters of transcript sequences
- CDD: conserved protein domain database
- 3D Domains: domains from Entrez Structure
- UniSTS: markers and mapping data
- PopSet: population study data sets
- GEO Profiles: expression and molecular abundance profiles
- GEO DataSets: experimental sets of GEO data
- Cancer Chromosomes: cytogenetic databases
- PubChem BioAssay: bioactivity screens of chemical substances
- GENSAT: gene expression atlas of mouse central nervous system
- Journals: detailed information about the journals indexed in PubMed and other Entrez databases
- NLM Catalog: catalog of books, journals, and audiovisuals in the NLM collections
- MeSH: detailed information about NLM's controlled vocabulary

One of the unique features of ENTREZ is that it was the first molecular biology database to incorporate links from one type of data (e.g. nucleotide) to the others (e.g. to proteins via translations, to MEDLINE entries via their MeSH numbers (NLM's Medical Subject Headings)). In addition, it incorporates an algorithm that identifies *related* entries in the databases. By *related*, we might mean genes in the same multigene family, or articles written about genes that have the same function, other proteins that function in the same biochemical pathway. Hence besides requesting the sequence for something, you can also find "everything else *like* this one". Thus, you can request MEDLINE abstracts of papers that are on similar or related topics (without any prior knowledge of their existence). Besides these "soft-links" via MeSH numbers there are also hard links encoded in the database that relate the abstract of the paper that reported the sequence or the protein entry of the nucleotide sequence.

When using the ENTREZ program, as with any web search engnine you must learn how to limit your search and to perform (as far as possible) formated queries. In the web based form of the ENTREZ program this is done through the "Limits" tab and the "Preview/Index" tab located just below the query entry box. These tabs permit the search to be restricted to individual organisms, to particular features and so on. They permit previous queries to be combined with logical operators such as AND, OR, NOT. The "neighbor" tab will be found among one of the many menu items that make this program a very powerful search engine. These are best explored through actual use.

The search fields that are available are shown in Table 4.3. These fields can be entered directly into the query search as "(adh OR mdh) AND Drosophila [ORGN] AND 1000:5000 [SLEN]" for example. The square brackets limit the previous term to the designated field – in this case search for the word Drosophila only in the organism field (ORGN). But the adh and mdh terms are searched in all fields by default. The range operator ':' is permissible with the ACCN, MOLWT, and SLEN fields. The boolean terms are AND, OR, NOT — they must be in upper case and can be combined with brackets ('(','(') to clarify meaning.

Table 4.3: The ENTREZ search fields

		Available for Database ...				
Field	Short term	Nucleotide	Protein	Genome	Structure	PopSet
Accession	ACCN	Yes	Yes	Yes	Yes	Yes
All Fields	ALL	Yes	Yes	Yes	Yes	Yes
Author Name	AUTH	Yes	Yes	Yes	Yes	Yes
EC/RN Number	ECNO	Yes	Yes	Yes	Yes	Yes
Feature Key	FKEY	Yes	No	Yes	No	Yes
Filter	FILT	Yes	Yes	Yes	Yes	Yes
Gene Name	GENE	Yes	Yes	Yes	No	Yes
Issue	ISS	Yes	Yes	Yes	Yes	Yes
Journal Name	JOUR	Yes	Yes	Yes	Yes	Yes
Keyword	KYWD	Yes	Yes	Yes	No	Yes
Modification Date	MDAT	Yes	Yes	Yes	Yes	Yes
Molecular Weight	MOLWT	No	Yes	No	No	No
Organism	ORGN	Yes	Yes	Yes	Yes	Yes
Page Number	PAGE	Yes	Yes	Yes	Yes	Yes
Primary Accession	PACC	Yes	Yes	Yes	No	Yes
Properties	PROP	Yes	Yes	Yes	No	Yes
Protein Name	PROT	Yes	Yes	Yes	No	Yes
Publication Date	PDAT	Yes	Yes	Yes	Yes	Yes
SeqID String	SQID	Yes	Yes	Yes	No	Yes
Sequence Length	SLEN	Yes	Yes	Yes	No	No
Substance Name	SUBS	Yes	Yes	No	Yes	No
Text Word	WORD	Yes	Yes	Yes	Yes	Yes
Title Word	TITL	Yes	Yes	Yes	No	No
Volume	VOL	Yes	Yes	Yes	Yes	Yes

Table 4.4: Some ENTREZ PubMed search fields

Field	Short term	Field	Short term
Affiliation	AD	All Fields	ALL
Author	AU	Corporate Author	CN
EC/RN Number	RN	Entrez Date	EDAT
Filter	FILTER	First Author	1AU
Full Author Name	FAU	Grant Name	GR
Issue	IP	Investigator	IR
Journal Title	TA	Language	LA
MeSH Date	MHDA	MeSH Major Topic	MAJR
MeSH Subheadings	SH	MeSH Terms	MH
NLM Unique ID	JID	Other Term	OT
Pagination	PG	Personal Name as Subject	PS
Pharmacological Action	PA	Place of Publication	PL
Publication Date	DP	Publication Type	PT
Publisher Identifier	AID	Secondary Source ID	SI
Subset	SB	Substance Name	NM
Text Words	TW	Title	TI
Title/Abstract	TIAB	Unique Identifiers	UID
Volume	VI		

The search fields for Entrez PubMed are slightly different from these and are shown in Table 4.4. The boolean operators ("AND", "OR", etc.) are the same and all of these can be combined to yield a highly structured query. The documentation for PubMed can be found at http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html.

The ENTREZ program is also released as a standalone program free of charge and you can download them to your computer from the NCBI ftp site.

4.8.2 NCBI retrieve

Discontinued Apr 2002: a great loss of convenience!

There is also an e-mail retrieval service at NCBI that is at times more convenient when you want to retreive large numbers of entries and when you are not interested in an interactive display (e.g. for remote access via a comptuer program). Mail your inquiry off to query@ncbi.nlm.nih.gov with the following example format for your query.

```
DATALIB gb
TITLES
MAXDOCS 30
BEGIN
BOVPRL
J02459 [ACC]
```

The DATALIB must be gb or genbank (GenBank), gbu or gbupdate (only updates since the last release), gbonly (official release only), emb or embl (EMBL), emblu or emblupdate (only updates since the last release), emblonly (full release only),sp or swiss or swissprot (Swiss-Prot), spu or swissprotupdate (updates only), pir (PIR database), omim (OMIM), vector (vector sequences), gp or genpept (translated GenBank), gpu or gpuupdate (updates only), kabatnuc (immunological nucleotide sequences), kabatpro (immunological protein sequences), and, though not stated in the official documentation, MEDLINE also works.

TITLES will display only the title of the matching record. MAXDOCS/MAXLINES restrict the volume of returns. Only DATALIB and BEGIN are mandatory. The above will search for records with LOCUS titles "BOVPRL" or accession number J02459. (NOTE: to put an underscore in the search, enclose the locus name in double quotes).

This retrieval service permits boolean searches. A logical **OR** is the implied default - as above, BOVPRL **or** J02459. But a logical **AND** and a logical **NOT** can be added to the query. Hence, "BOVPRL AND J02459" will retrieve records with both BOVPRL and J02459 in the record. The queries can be constructed with parenthesis to group items and with asterisks to match anything. For example, "(lysine OR glutamine) NOT vitelloge*". The field restrictor [ACC] restricts J02459 to be located in the accession number field. The field restrictors (the three letter codes) for the major databases are:

```
#        GENBANK and GBUPDATE Field Descriptions
DEFINITION [DEF]LOCUS [LOC] ACCESSION NO. [ACC]
KEYWORDS [KEY] SEGMENT [SEG] SOURCE [SRC]
REFERENCE [REF] COMMENT [COM] FEATURES [FEA]
ORIGIN [ORG]

#        EMBL and EMBLUPDATE Field Descriptions
DEFINITION [DEF] ID [LOC] ACCESSION [ACC]
KEYWORDS [KEY] DATES [DAT] SOURCE [SRC]
CROSS-REF [DXR] REFERENCE [REF] COMMENT [COM]
FEATURES [FEA]

#        SWISS-PROT Field Descriptions
DEFINITION [DEF] ID [LOC]  ACCESSION [ACC]
KEYWORDS [KEY]  DATES [DAT]  GENE NAME[GEN]
SOURCE [SRC]  ORGANISM CLASSIFICATION [CLS]
ORGANELLE [ORG]  REFERENCE [REF]  COMMENT [COM]
FEATURES [FEA]  CROSS REFERENCE [DCR]  SEQUENCE DATA [BAS]

#        PIR Protein Data Base (NBRF)
DEFINITION [DEF] ALT-NAME [ALT]  SUMMARY [SUM]
DATE [DAT]  SUPERFAMILY [SUP]  ACCESSION NO [ACC]
HOST [HST]  KEYWORDS [KEY]
SOURCE [SRC]  GENETICS [GEN]  INCLUDES [INC]
REFERENCE [REF]  COMMENT [COM]  FEATURES [FEA]

#        Online Mendelian Inheritance in Man (OMIM)
```

```
MIM NUMBER [NO] TITLE [TI] MINI-MIN [MN]
TEXT [TX] ALLELIC VARIANTS [AV] SEE ALSO REFERENCES [SA]
REFERENCES [RF] CLINICAL SYNOPSES [CS] CREATION DATE [CD]
EDIT HISTORY [ED]

#        Brookhaven Protein Data Bank (PDB)
DEFINITION [DEF] HEADER [HDR] ACCESSION [ACC]
DATE [DAT] SOURCE [SRC] AUTHOR [AUT]
SUPERSEDE [SPR] REFERENCE [REF] COMMENTS [COM]
FOOTNOTE [FTN] HETEROGENS [HET]
```

4.8.3 EMBL get

EMBL sequences can also be obtained either via the emblfetch program. In addition to this simple search form, there is also a more extensive and powerful internet form that permits many databases to be searched at once. This is termed the SRS (sequence retrieval system) which was developed by LION Bioscience (now defunct) and released for public use. The particular feature of SRS is its ability to link seamlessly between multiple life science databases and to integrate this data.

Data can also be obtained via an e-mail message send to the databases at netserv@ebi.ac.uk. The subject line of an e-mail message should be blank. In the interior of the message put the following:

```
get nuc:pip03xx
get nuc:x03392
get prot:kap_yeast
```

This will get the sequence with accession numbers pip03xx and x03392 from the nucleotide databases and the protein sequence with locus name kap_yeast from the protein database.

4.8.4 Others

There are, again, many other database access tools. For example, there is the DBGET system. This is run out of Japan (the Supercomputer Laboratory (SCL) in Kyoto and the Human Genome Center (HGC) in Tokyo). Once again, this search engine can find relevant data from several databases, including:

```
DNA: GenBank and EMBL
       GenBank: nucleic acid sequence database
       EMBL: nucleic acid sequence database
Protein: SWISS-PROT, PIR, PRF and PDBSTR
       SWISS-PROT: protein sequence database
       PIR: protein sequence database
       PRF: protein sequence database
       PDBSTR: Re-organized Protein Data Bank
KEGG Pathway Database
       PATHWAY: KEGG Pathway Database
       GENES: KEGG Genes Database
       BRITE: Biomolecular Relations in Information Transmission and Expression
       LIGAND: Ligand chemical database for enzyme reactions
PMD: Protein Mutant Database
PDB: Protein Structure Database
AAindex: Amino Acid index database
LITDB: PRF protein/peptide literature database
OMIM: Online Mendelian Inheritance in Man
Medline: Literature database
EPD: Eukaryotic promoter database
TRANSFAC: Transcription factor database
MotifDic: Dictionary of protein sites and patterns
```

Each of the databases can have individual access tools that can provide more specialized access. For example, the PDB database supports viewing of protein structures via VRML (virtual reality modelling language), Rasmol (a freely available program for displaying molecules in three dimensions), FirstGlance and Protein Explorer (two other programs that require a commercial product), and via still photographs. There is also a special browser for the SWISS 3DIMAGE database and so on.

For each database, look for a specialized browser to access the data making use of the peculiarities of the data stored.

4.9 Reliability

The data within the databases may not always be what it pretends to be. This venture is a human one and humans make mistakes. Indeed, the venture relies on the contributions of many people and they all have different standards of accuracy. One of the most common errors in the early days was the presence of vector sequence in the midst of some other sequence. Today this is not such a large problem since most entries are now automatically screened against known vectors and the error can be caught before the sequence makes it into the databases.

Smaller errors in sequences are also common. The human APRT gene sequence was determined and entered into the data base by one laboratory. A few months later, another laboratory published a paper with a sequence that differed from the previous work by 13 nucleotides and 60 insertions/deletions over 3 kb. It is impossible to tell how much of this may be due to polymorphism and how much may be due to actual sequencing error. Because this kind of duplication is not done for every sequence it is impossible to say that this is typical or atypical of the sequencing done. However, as a counter example, a check of the yeast genome revealed only a couple of differences over many megabases (H. Bussey - personal communication).

Unfortunately, many errors are not easily corrected. Current policy for most of the databases is that the people running them are responsible for the database en masse while the actual data is the business of the researchers. Hence the databases are meant to act as an archive and unless the original author corrects their data, it will remain archived in the database. On a more positive note, you will also find many entries that were created long ago and yet, last modified very recently to incorporate the latest information. Further more, many of the databases devoted to one organism will gather and carefully curate the data.

Still another problem that has shown up is trivial data entries. The following entry was noticed by Reinhard Doelz.

Database Silliness

```
LOCUS       A00674                     6 bp    DNA     linear   PAT 29-JAN-1993
DEFINITION  Nucleotide sequence 3 from patent number WO8601533.
ACCESSION   A00674
VERSION     A00674.1  GI:14588
KEYWORDS    .
SOURCE      unidentified
  ORGANISM  unidentified
            unclassified sequences.
REFERENCE   1  (bases 1 to 6)
  AUTHORS   Neuberger,M.S.
  TITLE     PRODUCTION OF CHIMERIC ANTIBODIES
  JOURNAL   Patent: WO 8601533-A1 3 13-MAR-1986;
            CellTech Ltd
FEATURES             Location/Qualifiers
     source          1..6
                     /organism="unidentified"
                     /mol_type="unassigned DNA"
                     /db_xref="taxon:32644"
ORIGIN
        1 cactaa
//
```

This is truly an amazing entry. It is fully six nucleotides long, it comes from an unknown source, it comes from an unknown organism and from unclassified sequences. But it is patented !! What could possibly be the purpose of entering this sequence in the database and even more incredulously, why would one ever patent it? By random chance your DNA must contain this sequence. Since it does and the sequence was patented, be forewarned that you should obtain correct written permission from the patent holders before you replicate it again. More seriously, if you construct oligos for PCR or sequencing, you are probably guilty of patent infringement. Reinhard calculated that this silly hexanucleotide occurs 28340 times within the database and in over 70000 sequences (circa 1993). This entry is perhaps extreme but there are other, less extreme entries of equally doubtful quality.

The take home message from all of this is to look at the data with a critical eye. The actual quantity and type of errors within the databases are not known - some researchers are very careful and can check their sequence data, for others a double check of the sequence data may not be possible. When doing your own research, assume that the sequence may contain some errors and take measures to prevent this from destroying the validity of your conclusions.

Caveat Emptor