To begin — some terminology.

What is bioinformatics, genomics, proteomics, sequence analysis, computational molecular biology . . . ?
My Definitions, lots of overlap —

**Biocomputing** and computational biology are synonyms and describe the use of computers and computational techniques to analyze any type of a biological system, from individual molecules to organisms to overall ecology.

**Bioinformatics** describes using computational techniques to access, analyze, and interpret the biological information in any type of biological database.

**Sequence analysis** is the study of molecular sequence data for the purpose of inferring the function, interactions, evolution, and perhaps structure of biological molecules.

**Genomics** analyzes the context of genes or complete genomes (the total DNA content of an organism) within the same and/or across different genomes.

**Proteomics** is the subdivision of genomics concerned with analyzing the complete protein complement, i.e. the proteome, of organisms, both within and between different organisms.

One way to think about it —

The Reverse Biochemistry Analogy.

Biochemists no longer have to begin a research project by isolating and purifying massive amounts of a protein from its native organism in order to characterize a particular gene product. Rather, now scientists can amplify a section of some genome based on its similarity to other genomes, sequence that piece of DNA and, using sequence analysis tools, infer all sorts of functional, evolutionary, and, perhaps, structural insight into that stretch of DNA!

The computer and molecular databases are a necessary, integral part of this entire process.

---

<table>
<thead>
<tr>
<th>Year</th>
<th>BasePars</th>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>680,338</td>
<td>606</td>
</tr>
<tr>
<td>1983</td>
<td>2,274,029</td>
<td>2,427</td>
</tr>
<tr>
<td>1984</td>
<td>3,368,765</td>
<td>4,175</td>
</tr>
<tr>
<td>1985</td>
<td>5,294,420</td>
<td>6,700</td>
</tr>
<tr>
<td>1986</td>
<td>9,131,271</td>
<td>9,070</td>
</tr>
<tr>
<td>1987</td>
<td>13,511,776</td>
<td>14,308</td>
</tr>
<tr>
<td>1988</td>
<td>20,464,000</td>
<td>20,579</td>
</tr>
<tr>
<td>1989</td>
<td>34,762,955</td>
<td>28,791</td>
</tr>
<tr>
<td>1990</td>
<td>60,170,000</td>
<td>36,533</td>
</tr>
<tr>
<td>1991</td>
<td>71,947,426</td>
<td>55,627</td>
</tr>
<tr>
<td>1992</td>
<td>10,106,868</td>
<td>79,859</td>
</tr>
<tr>
<td>1993</td>
<td>15,162,442</td>
<td>13,462</td>
</tr>
<tr>
<td>1994</td>
<td>21,719,042</td>
<td>21,187</td>
</tr>
<tr>
<td>1995</td>
<td>38,610,465</td>
<td>53,568</td>
</tr>
<tr>
<td>1996</td>
<td>60,917,044</td>
<td>72,127</td>
</tr>
<tr>
<td>1997</td>
<td>116,230,847</td>
<td>170,087</td>
</tr>
<tr>
<td>1998</td>
<td>20,464,174</td>
<td>20,376</td>
</tr>
<tr>
<td>1999</td>
<td>29,740,320</td>
<td>29,740</td>
</tr>
<tr>
<td>2000</td>
<td>38,987,420</td>
<td>38,987</td>
</tr>
<tr>
<td>2001</td>
<td>109,004,238</td>
<td>109,004</td>
</tr>
<tr>
<td>2002</td>
<td>248,506,114</td>
<td>248,506</td>
</tr>
<tr>
<td>2003</td>
<td>282,569,895</td>
<td>282,569</td>
</tr>
<tr>
<td>2004</td>
<td>445,714,174</td>
<td>445,714</td>
</tr>
<tr>
<td>2005</td>
<td>642,773,662</td>
<td>642,773</td>
</tr>
</tbody>
</table>

The exponential growth of molecular sequence databases & cpu power

Doubling time just over a year!

Database Growth (cont.) —

The International Human Genome Sequencing Consortium announced the completion of the “Working Draft” of the human genome in June 2000; independently that same month, the private company Celera Genomics announced that it had completed the first “Assembly” of the human genome. Both articles were published mid-February 2001 in the journals *Science* and *Nature*.

Genome projects have kept the data coming at alarming rates. As of August 2008, 49 Archaea, 574 Bacteria, and 22 Eukaryote complete genomes, and over 200 Eukaryote assemblies were represented, not counting the almost 3,000 virus and viroid genomes available.
Some neat stuff from the human genome papers —
Homo sapiens, aren’t nearly as special as we once thought. Of the 3.2 billion base pairs in our DNA:
Traditional gene number estimates were often in the 100,000 range; turns out we’ve only got about twice
as many as a fruit fly, between 25’ and 30,000!
The protein coding region of the genome is only about 1% or so, a bunch of the remainder is ‘jumping,’
‘junk,’ ‘selfish DNA,’ much of which may be involved in regulation and control.
Over 100-200 genes were transferred from an ancestral bacterial genome to an ancestral vertebrate
genome! (Later shown to be false by more extensive analyses, and to be due to gene loss not transfer.)

What are sequence databases?
Sequence databases are an organized way to store exponentially accumulating sequence data. Most have a specific format. An
‘alphabet soup’ of three major organizations maintain most of this data. They largely ‘mirror’ one another and share accession codes, but NOT proper identifier names:
North America: the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), at the
National Institute of Health (NIH), maintains the GenBank
nucleotide, GenPept amino acid, and RefSeq genome, transcriptome, and proteome databases.
Europe: the European Molecular Biology Laboratory (EMBL), the
European Bioinformatics Institute (EBI), and the Swiss Institute of
Bioinformatics (SIB) all help maintain the EMBL nucleotide sequence database, and the UNIPROT (SWISS-PROT + TrEMBL)
amino acid sequence database (with PIR/NBRF support also).
Asia: The National Institute of Genetics (NIG) supports the Center for Information Biology’s (CIG) DNA Data Bank of Japan (DDBJ).

A little history —
Developments that affect software and the end user —
The first well recognized sequence database was Dr. Margaret Dayhoff’s
hardbound Atlas of Protein Sequence and Structure begun in the mid-sixties. DDBJ began in 1984, GenBank in 1982, and EMBL in
1980. They are all attempts at establishing an organized, reliable, comprehensive, and openly available library of genetic sequences. 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. Each program needs to be able to recognize particular aspects of the sequence files; whenever they change it screws everything up. Database format standards are constantly argued over — relational vs. object-oriented vs. XML vs. ASN.1, etc. Unfortunately, until all biologists and computer scientists worldwide agree on one standard and all software is (re)written to that standard, neither of which is likely to happen very quickly, if ever, format issues will remain one of the most confusing and troubling aspects of working with primary sequence data.

What are sequence databases like?
Just what are primary sequences?
(Central Dogma: DNA → RNA → protein)
Primary refers to one dimension — all of the ‘symbol’ information written in sequential order necessary to specify a particular biological molecular entity, be it polypeptide or nucleotide.
The symbols are the one letter codes for all of the biological nitrogenous bases and amino acid residues and their ambiguity codes. Biological carbohydrates, lipids, and structural and functional information are not sequence data. Not even DNA CDS translations in a DNA database are sequence data!
However, much of this feature and bibliographic type information is available in the reference documentation sections associated with primary sequences in the databases.
Sequence database installations are commonly a complex ASCII/Binary mix, though usually not relational or Object Oriented (but proprietary and Web-based ones often are). They’ll contain several very long text files each containing different types of related information, such as all of the sequences themselves, versus all of the title lines, or all of the reference sections. Binary files often help ‘glue together’ all of these other files by providing indexing functions.

Software is usually required to successfully interact with these databases and access is most easily handled through various software packages and interfaces, either on the World Wide Web or otherwise.

Parts and problems —
Sequence databases contain several elements associated with each sequence:

- **Name**: LOCUS, ENTRY, ID, all are unique identifiers.
- **Definition**: a.k.a. title, a brief textual sequence description.
- **Accession Number**: a constant data identifier.
- **Source and taxonomy information**.
- **Complete literature references**.
- **Comments and keywords**.
- The all important **FEATURE table**!
- A summary or checksum line.
- The sequence itself.

But:
- Each major database as well as each major suite of software tools that you are likely to use has its own distinct format requirements.
- This can be a huge problem and an enormous time sink, even with helpful tools such as Don Gilbert’s ReadSeq. Therefore, becoming familiar with some of the common formats is a big help. Look for key features of each type of entry, as seen here.

More organization stuff —
Nucleic acid sequence databases (and TRESMBL) are split into subdivisions based on taxonomy (historical rankings — the Fungi and Archeae warning!). TRESMBL sequences are merged into SWISS-PROT as they receive increased levels of annotation. Both together comprise Uniprot. GenPept has minimal annotation.

### Nucleic Acid DB’s
- GenBank/EMBL/DBJ
  - all Taxonomic categories + HTG’s & STS’s
  - “Tags” EST’s, GSS’s, & HTC’s

### Amino Acid DB’s
- Uniprot = SWISS-PROT + TrEMBL
- GenPept

### NCBi GenBank and GenPept format
Look for “LOCUS,” “FEATURES,” “ORIGIN,” the sequence itself, and then “//.”

### Parts and problems
- Sequence databases contain several elements associated with each sequence:
  - **Name**: LOCUS, ENTRY, ID, all are unique identifiers.
  - **Definition**: a.k.a. title, a brief textual sequence description.
  - **Accession Number**: a constant data identifier.
  - **Source and taxonomy information**.
  - **Complete literature references**.
  - **Comments and keywords**.
  - The all important **FEATURE table**!
  - A summary or checksum line.
  - The sequence itself.

But:
- Each major database as well as each major suite of software tools that you are likely to use has its own distinct format requirements.
- This can be a huge problem and an enormous time sink, even with helpful tools such as Don Gilbert’s ReadSeq. Therefore, becoming familiar with some of the common formats is a big help. Look for key features of each type of entry, as seen here.
Steve Thompson

GCG MSF & RSF format

—

EMBL and UniProt format

—

GCG single sequence

—

Specialized ‘sequence’-type DB’s —

Databases that contain special types of sequence information, such as patterns, motifs, and profiles. These include: REBASE, EPD, PROSITE, BLOCKS, ProDom, Pfam . . . .

Databases that contain multiple sequence entries aligned, e.g. ProSet, RDP, and ALN.

Databases that contain families of sequences ordered functionally, structurally, or phylogenetically, e.g., PiProClass and HOVERGEN.

Databases of species specific sequences, e.g. the HIV Database and the Giardia lamblia Genome Project. And on and on . . . . See Amos Bairoch’s excellent links page: http://us.expasy.org/links.html.
What about other types of biological databases?

Three-dimensional structure databases —
the Protein Data Bank and Rutgers Nucleic Acid Database.

These databases contain all of the 3D atomic coordinate data necessary to define the tertiary shape of a particular biological molecule. The data is usually experimentally derived, either by X-ray crystallography or by NMR, sometimes it’s hypothetical. The source of the structure and its resolution is always given. Secondary structure boundaries, sequence data, and reference information are often associated with the coordinate data, but it is the 3D data that really matters, not the annotation.

Molecular visualization or modeling software is required to interact with the data. It has little meaning on its own.

And still other types of Genomics DB’s —
These can be considered ‘non-molecular’:

Reference Databases (also w/ pointers to sequences): e.g.
LocusLink/Gene — integrated knowledge base
OMIM — Online Mendelian Inheritance in Man
PubMed/MedLine — over 11 million citations from more than 4 thousand bio/medical scientific journals.
Phylogenetic Tree Databases: e.g. the Tree of Life.
Metabolic Pathway Databases: e.g. WIT (What Is There), Japan’s GenomeNet KEGG (the Kyoto Encyclopedia of Genes and Genomes), and the human Reactome.
Population studies data — which strains, where, etc.

And then databases that many biocomputing people don’t even usually consider: e.g. GIS/GPS/remote sensing data, medical records, census counts, mortality and birth rates . . . .

Tying it all together: map browsers —

Genetic linkage mapping databases for most large genome projects— H. sapiens, Mus, Drosophila, C. elegans, Saccharomyces, Arabidopsis, E. coli . . . .
. . . often tie it all together with links to all the other databases within the context of a genome browser or map viewer. Examples include:
Some of my favorite WWW genomics analyses access sites —

<table>
<thead>
<tr>
<th>Site</th>
<th>URL (Uniform Resource Locator)</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMBL</td>
<td><a href="http://www.ebi.ac.uk">http://www.ebi.ac.uk</a></td>
<td>database/analysis/software</td>
</tr>
<tr>
<td>The Institute for Genomic Research</td>
<td><a href="http://www.tigr.org">http://www.tigr.org</a></td>
<td>microbial genome projects</td>
</tr>
<tr>
<td>Human Genome Project</td>
<td><a href="http://www.gdb.org">http://www.gdb.org</a></td>
<td>Human Genome DataBase</td>
</tr>
<tr>
<td>The Sanger Institute</td>
<td><a href="http://www.sanger.ac.uk">http://www.sanger.ac.uk</a></td>
<td>databases/analysis/software</td>
</tr>
<tr>
<td>European Bioinformatics Institute</td>
<td><a href="http://www.ebi.ac.uk">http://www.ebi.ac.uk</a></td>
<td>databases/analysis/software</td>
</tr>
<tr>
<td>Japan’s GenomeNet Server</td>
<td><a href="http://www.genome.ad.jp">http://www.genome.ad.jp</a></td>
<td>database/software archive</td>
</tr>
<tr>
<td>Univ. of Montreal MegaSun</td>
<td><a href="http://megasun.bch.umontreal.ca">http://megasun.bch.umontreal.ca</a></td>
<td>database/software archive</td>
</tr>
<tr>
<td>IUBIO Biology Archive</td>
<td><a href="http://iubio.bio.indiana.edu">http://iubio.bio.indiana.edu</a></td>
<td>databases/analysis/software</td>
</tr>
<tr>
<td>Univ. of California, Santa Cruz Genome Browser</td>
<td><a href="http://genome.ucsc.edu">http://genome.ucsc.edu</a></td>
<td>University of California, Santa Cruz Genome Browser (UCSC Genome Browser)</td>
</tr>
<tr>
<td>Lawrence Livermore National Laboratory ECR Browser</td>
<td><a href="http://ecrbrowser.dcode.org">http://ecrbrowser.dcode.org</a></td>
<td>Lawrence Livermore National Laboratory ECR Browser (Lawrence Livermore National Laboratory ECR Browser)</td>
</tr>
</tbody>
</table>

With tools like NCBI’s Entrez, EMBL’s EMBOSS, and various genome browsers and map viewers.

Web-based biological molecular database access tools, pros and cons —

Advantages: Accesses the very latest updates. It’s fun and very fast. It can be very powerful and efficient, if you know what you’re doing. In most cases relational links between different databases ease navigation, and in some cases neighboring concepts link similar entries. Genome-scale analysis is possible.

Disadvantages: Can be very inefficient, if you don’t know what you’re doing. Reformatting is usually essential, if the sequence is to be used in any other software. And, it’s very easy to get lost and distracted in cyberspace!
Also, problems sometimes arise with the Web, like dropped or slow connections . . . . So what are the alternatives?

Personal computer software solutions — public domain programs are available, but . . . a bit complicated to install, configure, and maintain. User must be pretty computer savvy. So, good commercial software packages are also available, e.g. Sequencher, MacVector, DNAsis, etc., but . . . license hassles, especially big expense per machine, and Internet and/or CD database access all complicate matters!

Therefore, server-based, non-Web solutions— we’re talking UNIX server computers here (OS issues). Public domain solutions available, and cooperative systems managers can install, configure, and maintain everything for users. Centralized products, e.g. the ‘retired’ Accelrys GCG Wisconsin Package and the SeqLab Graphical User Interface (GUI), simplify matters for administrators and users. Connections from any networked terminal or workstation anywhere, anytime provide fast, convenient database access on local server disks!

Within the GCG suite, LookUp is an SRS derivative used to find a sequence of interest from local GCG server databases.

Advantage: Search output is a legitimate GCG list file, appropriate input to other GCG programs; no need to reformat — all GCG.

Disadvantage: DB’s only as new as administrator maintains them.

The Genetics Computer Group — The Accelrys Wisconsin Package for Sequence Analysis

GCG began in 1982 in Oliver Smithies’ Genetics lab at the University of Wisconsin; in 1990 it became a private company; it was acquired by the Oxford Molecular Group, U.K., in 1997; by Pharmacopeia, U.S.A., in 2000; in 2004 Accelrys, U.S.A., left Pharmacopeia to become an independent entity; in 2008 Accelrys ‘retired’ the product.

The suite contains around 150 programs designed to work in a “toolbox” fashion. Several simple programs used in succession can lead to very sophisticated results.

‘Internal compatibility,’ i.e. once you learn to use one program, all programs can be run similarly, and, the output from many programs can be used as input for other programs.

Still used all over the world, so learning it will often be useful at other research institutions as well.

To answer the always perplexing GCG question — “What sequence(s)? . . . .” Specifying sequences, GCG style; in order of increasing power and complexity:

The sequence is in a local GCG format single sequence file in your UNIX account. (GCG Reformat and SeqConv+ programs)

The sequence is in a local GCG database in which case you ‘point’ to it by using any of the GCG database logical names. A colon, “:,” always sets the logical name apart from either an accession number or a proper identifier name or a wildcard expression and they are case insensitive.

The sequence is in a GCG format multiple sequence file, either an MSF (multiple sequence format) file or an RSF (rich sequence format) file. To specify sequences contained in a GCG multiple sequence file, supply the file name followed by a pair of braces, “{},” containing the sequence specification, e.g. a wildcard — “{}.”

Finally, the most powerful method of specifying sequences is in a GCG “list” file. This is merely a list of other sequence specifications and can even contain other list files within it. The convention to use a GCG list file in a program is to precede it with an at sign, “@”. Furthermore, attribute information within list files can specify particular sequence aspects.
Logical terms for the Wisconsin Package —

Sequence databases, nucleic acids:

- **GENBANKPLUS**: All of GenBank plus EST, HTC, and GSS
- **SYNTHETIC**: GenBank synthetic
- **GBP**: All of GenBank plus EST, HTC, and GSS
- **SY**: GenBank synthetic
- **GENBANK**: All of GenBank except EST, HTC, and GSS
- **UNANNOTATED**: GenBank unannotated
- **GB**: All of GenBank except EST, HTC, and GSS
- **UN**: GenBank unannotated
- **BACTERIAL**: GenBank bacteria and archaea
- **REFSEQNUC**: NCBI RefSeq transcriptomes
- **BA**: GenBank bacteria and archaea
- **RS_RNA**: NCBI RefSeq transcriptomes
- **INVERTEBRATE**: GenBank invertebrate
- **IN**: GenBank invertebrate
- **OTHERMAMMAL**: GenBank other mammals
- **OM**: GenBank other mammals
- **OTHERVERTEBRATE**: GenBank other vertebrates
- **OV**: GenBank other vertebrates
- **PHAGE**: GenBank phage
- **PH**: GenBank phage
- **PLANT**: GenBank plant and fungi
- **PRIMATE**: GenBank primate
- **UNIPROT**: All of Swiss-Prot and all of SPTREMBL
- **PR**: GenBank primate
- **UNI**: All of Swiss-Prot and all of SPTREMBL
- **RODENT**: GenBank rodent
- **RO**: GenBank rodent
- **SWISSPROTPLUS**: All of Swiss-Prot and all of SPTREMBL
- **SWP**: All of Swiss-Prot and all of SPTREMBL
- **VI**: GenBank viral
- **SWISS**: All of Swiss-Prot (fully annotated)
- **VIRAL**: GenBank viral
- **SW**: All of Swiss-Prot (fully annotated)
- **TAGS**: GenBank EST, HTC, and GSS

Sequence databases, amino acids:

- **PLANT**: GenBank plant and fungi
- **PRIMATE**: GenBank primate
- **UNIPROT**: All of Swiss-Prot and all of SPTREMBL
- **PR**: GenBank primate
- **UNI**: All of Swiss-Prot and all of SPTREMBL
- **RODENT**: GenBank rodent
- **RO**: GenBank rodent
- **SWISSPROTPLUS**: All of Swiss-Prot and all of SPTREMBL
- **SWP**: All of Swiss-Prot and all of SPTREMBL
- **VI**: GenBank viral
- **SWISS**: All of Swiss-Prot (fully annotated)
- **SW**: All of Swiss-Prot (fully annotated)

The List File Format —

```
!!SEQUENCE_LIST 1.0
An example GCG list file of many elongation 1a and Tu factors follows. As with all GCG data files, two periods separate documentation from data.
...my-special.pep begin:24 end:134
SwissProt:EfTu_Ecoli
Ef1a-Tu.msf(*)
/usr/accounts/test/another.rsf(ef1a_*)
@another.list
```

SeqLab — GCG’s X-based GUI!

*SeqLab* is the merger of Steve Smith’s Genetic Data Environment and GCG’s Wisconsin Package Interface:

GDE + WPI = *SeqLab*

Requires an X-Windowing environment — either native on UNIX computers (including LINUX, but not installed by default on Mac OS X v.10+) but see Apple’s free X11 package or XDarwin), or emulated with X-Server Software on personal computers.

SeqLab — GCG’s Graphical User Interface
There's a bewildering assortment of different biological molecular databases and ways to access and manipulate the information within them. The key is to learn how to use the data and the methods in the most efficient manner — knowing which to use when and how to combine their inferences will go a long way toward success!

A comprehensive sequence analysis software suite, such as the GCG Package, expedites the chore, putting a large assortment of tools all under one organizational model with one user interface.

Conclusions —