Basic molecular biology and overview of major bioinformatics web resources

Yanbin Yin

Outline

- Basic molecular biology
- Web Databases
- Web Servers

References

- NAR database and web server annual special issues
- http://pbil.univ-lyon1.fr/bookmarks.html
- http://www.ebi.ac.uk/2can/resources/index.html
- http://www.ebi.ac.uk/training/online/
- http://www.ncbi.nlm.nih.gov/Class/FieldGuide/
- http://www.csd.hku.hk/bruhk/resources.html
- http://anil.cchmc.org/BioInfoRes.html
- http://www.ncbi.nlm.nih.gov/Class/NAWBIS/

How to run bioinformatics applications

- **Most naive**: use web-based tools [e.g. NCBI blast server]
  – Primarily need biology

- **More professional**: Use stand-alone tools
  1. tools with GUI (graphical user interface: click buttons)
  2. tools without GUI (terminal-based: type commands)
  – Need biology plus ability to use Unix, write wrappers in Perl/Python, write shell scripts
  – Sometimes need an understanding of data storage and scalability

- **Most advanced**: Algorithm development
  – Computer-science focus, usually partner with a biologist

- Making data/methods public
  – Creating databases/web pages

http://chagall.med.cornell.edu/BioinfoCourse/presentations2012/Lecture1_Intro_2012.pdf

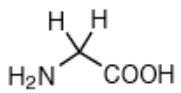
Proteins: the building blocks of life

- Proteins are the main building blocks and functional molecules of the cell, taking up almost 20% of a eukaryotic cell's weight, the largest contribution after water (70%).

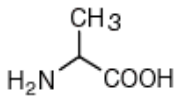
- Proteins are made from small molecules called amino acids. There are many types of proteins some of which are:
  - Structural proteins which can be thought of as the organism's basic building blocks.
  - Enzymes which perform (catalyse) a multitude of biochemical reactions, such as altering, joining together or chopping up other molecules. Together these reactions and the pathways they make up is called metabolism.
  - Regulatory proteins: transcription factors
  - Transmembrane proteins are key in the maintenance of the cellular environment, regulating cell volume, extraction and concentration of small molceules from the extracellular environment and generation of ionic gradients essential for muscle and nerve cell function

Proteins are chains of 20 different types of amino acids. This sequence of amino acids is known as the primary structure, and it can be represented as a string of 20 different symbols (i.e., a word over the common alphabet of 20 letters).

Small

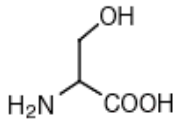


Glycine (Gly, G)
MW: 57.05

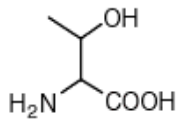


Alanine (Ala, A)
MW: 71.09

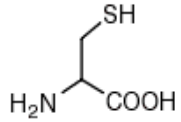
Nucleophilic



Serine (Ser, S)
MW: 87.08, pK_a ~ 16

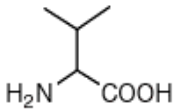


Threonine (Thr, T)
MW: 101.11, pK_a ~ 16

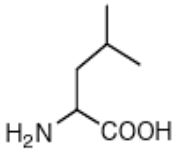


Cysteine (Cys, C)
MW: 103.15, pK_a = 8.35

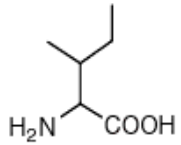
Hydrophobic



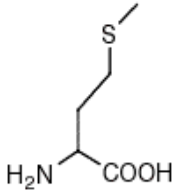
Valine (Val, V)
MW: 99.14



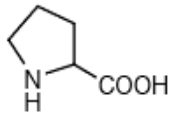
Leucine (Leu, L)
MW: 113.16



Isoleucine (Ile, I)
MW: 113.16

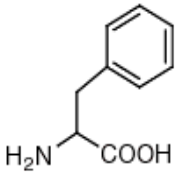


Methionine (Met, M)
MW: 131.19

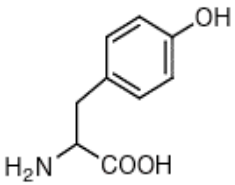


Proline (Pro, P)
MW: 97.12

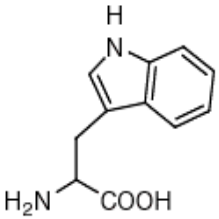
Aromatic



Phenylalanine (Phe, F)
MW: 147.18

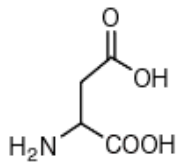


Tyrosine (Tyr, Y)
MW: 163.18

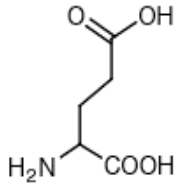


Tryptophan (Trp, W)
MW: 186.21

Acidic

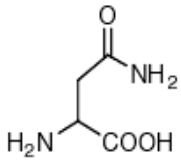


Aspartic Acid (Asp, D)
MW: 115.09, pK_a = 3.9

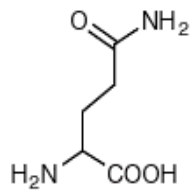


Glutamic Acid (Glu, E)
MW: 129.12, pK_a = 4.07

Amide

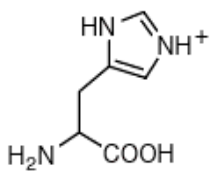


Asparagine (Asn, N)
MW: 114.11

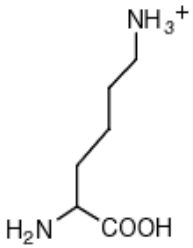


Glutamine (Gln, Q)
MW: 128.14

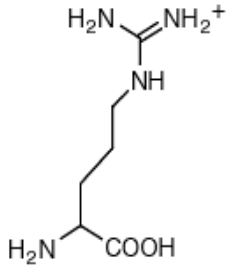
Basic



Histidine (His, H)
MW: 137.14, pK_a = 6.04



Lysine (Lys, K)
MW: 128.17, pK_a = 10.79



Arginine (Arg, R)
MW: 156.19, pK_a = 12.48

Protein primary sequence

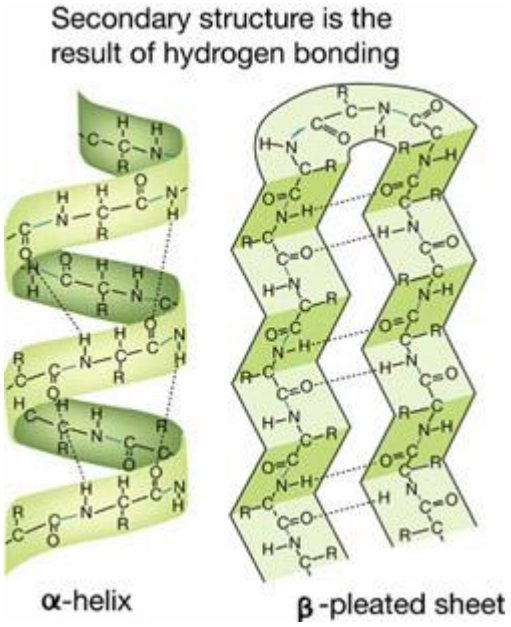
>gi|170083040|ref|YP_001732360.1| cellulose synthase subunit BcsC [Escherichia coli str. K-12 substr. DH10B]
MRKFTLNIFTLSLGLAVMPMVEAAPTAQQQLLEQVRLGEATHREDLVQQSLYRLELIDPNNPDVVAARFR
SLLRQGDIDGAQKQLDRLSQLAPSSNAYKSSRTTMLLSTPDGRQALQQARLQATTGHAEEAVASYNKLFN
GAPPEGDIAVEYWSTVAKIPARRGEAINQLKRINADAPGNTGLQNNLALLLFSSDRRDEGFAVLEQMAKS
NAGREGASKIWYGQIKDMPVSDASVSALKKYLSIFSDGDSVAAAQSQLAEQQKQLADPAFRARAQGLAAV
DSGMAGKAIPELQQAVRANPKDSEALGALGQAYSQKGDRANAVANLEKALALDPHSSNNDKWNSLLKVNR
YWLAIQQGDAALKANNPDRAERLFQQARNVDNTDSYAVLGLGDVAMARKDYPAAERYYQQTLRMDSGNTN
AVRGLANIYRQQSPEKAEAFIASLSASQRRSIDDIERSLQNDRLAQQAEALENQGKWAQAAALQRQRLAL
DPGSVWITYRLSQDLWQAGQRSQADTLMRNLAQQKSNDPEQVYAYGLYLSGHDQDRAALAHINSLPRAQW
NSNIQELVNRLQSDQVLETANRLRESGKEAEAEAMLRQQPPSTRIDLTLADWAQQRRDYTAARAAYQNVL
TREPANADAILGLTEVDIAAGDKAAARSQLAKLPATDNASLNTQRRVALAQAQLGDTAAAQRTFNKLIPQ
AKSQPPSMESAMVLRDGAKFEAQAGDPTQALETYKDAMVASGVTTTRPQDNDTFTRLTRNDEKDDWLKRG
VRSDAADLYRQQDLNVTLEHDYWGSSGTGGYSDLKAHTTMLQVDAPYSDGRMFFRSDFVNMNVGSFSTNA
DGKWDDNWGTCTLQDCSGNRSQSDSGASVAVGWRNDVWSWDIGTTPMGFNVVDVVGGISYSDDIGPLGYT
VNAHRRPISSSLLAFGGQKDSPSNTGKKWGGVRADGVGLSLSYDKGEANGVWASLSGDQLTGKNVEDNWR
VRWMTGYYYKVINQNNRRVTIGLNNMIWHYDKDLSGYSLGQGGYYSPQEYLSFAIPVMWRERTENWSWEL
GASGSWSHSRTKTMPRYPLMNLIPTDWQEEAARQSNDGGSSQGFGYTARALLERRVTSNWFVGTAIDIQQ
AKDYAPSHFLLYVRYSAAGWQGDMDLPPQPLIPYADW

Protein secondary structure

Although the primary structure of a protein is linear, the molecule is not straight, and the sequence of the amino acids affects the folding

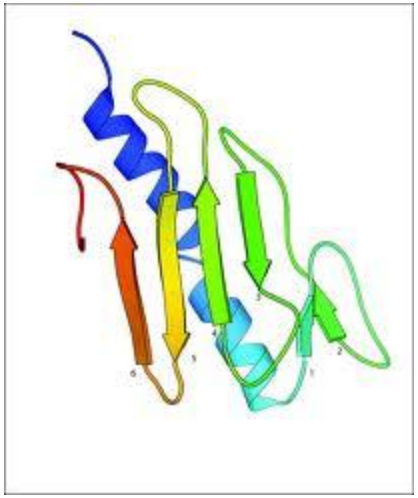
These three are called secondary structure elements:

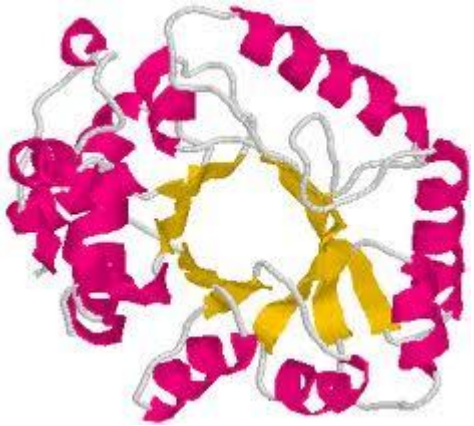
alpha-helices, beta-strands, loops



3D structure

As the result of the folding, parts of a protein molecule chain come into contact with each other and various attractive or repulsive forces (hydrogen bonds, disulfide bridges, attractions between positive and negative charges, and hydrophobic and hydrophilic forces) between such parts cause the molecule to adopt a fixed relatively stable 3D structure. This is called tertiary structure. In many cases the 3D structure is quite compact





DNA: genetic information carrier

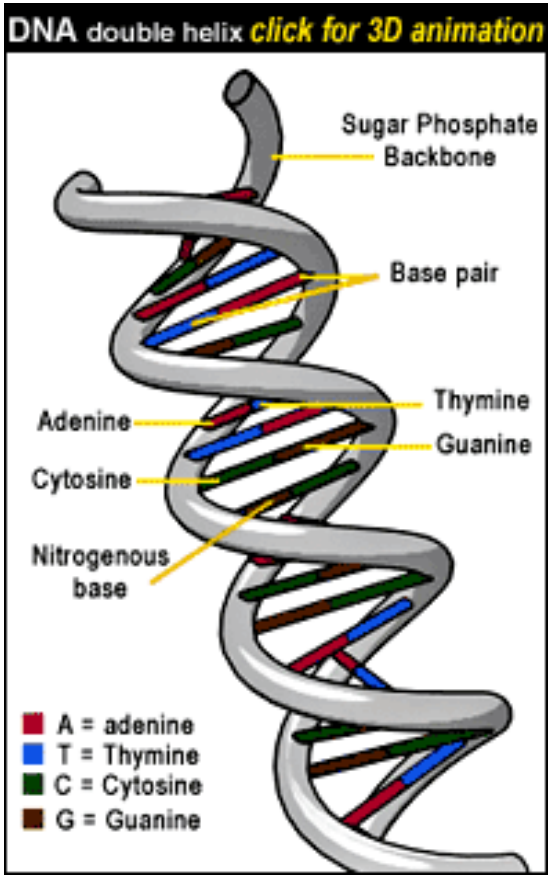
- DNA may be single or double stranded. A single stranded DNA molecule, also called a polynucleotide, is a chain of small molecules, called nucleotides. There are four different nucleotides grouped into two types, purines: adenine and guanine and pyrimidines: cytosine and thymine. They are usually referred to as bases and denoted by their initial letters, A, C, G and T (not to be confused with amino acids!)

```
5' C-G-A-T-T-G-C-A-A-C-G-A-T-G-C 3'
   | | | | | | | | | | | | | | |
3' G-C-T-A-A-C-G-T-T-G-C-T-A-C-G 5'
```

DNA Double Helix

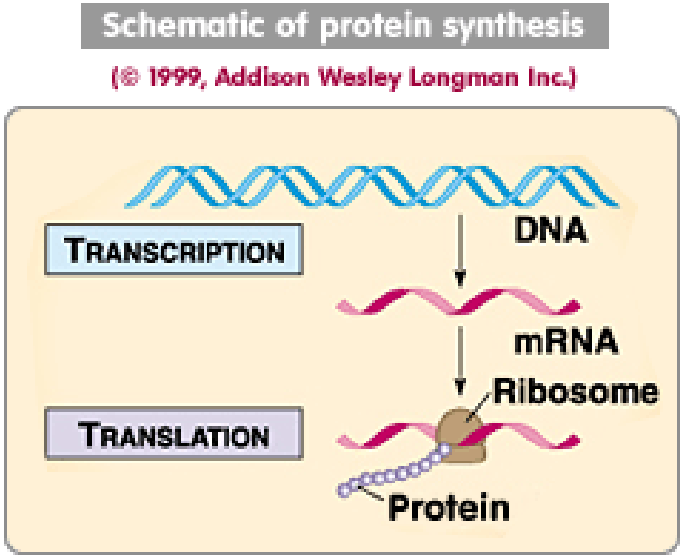
Two complementary polynucleotide chains form a stable structure, which resembles a helix and is known as a the DNA double helix. About 10 bp in this structure takes a full turn, which is about 3.4 nm long.

This structure was first figured out in 1953 in Cambridge by Watson and Crick (with the help of others). Later they got the Nobel Prize for this discovery, The DNA Double Helix.



RNA: genetic information messenger and much more unknown roles

RNA has various functions in a cell, e.g., mRNA and tRNA are functionally different types of RNA which are required for the two main steps in protein synthesis, transcription and translation.




```
C-G-A-T-T-G-C-A-A-C-G-A-T-G-C DNA
| | | | | | | | | | | | | |
G-C-U-A-A-C-G-U-U-G-C-U-A-C-G RNA
```

DNA sequence

>gi|49175990|ref|NC_000913.2| Escherichia coli str. K-12 substr. MG1655 chromosome, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAG
CCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAA
GTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTGGAAAGCAATGCC
AGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGGTGGCGATGATTG
AAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTGCCGAACTTTT
GACGGGACTCGCCGCCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTCGATCAGGAATTT
GCCCAAATAAAACATGTCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGC
TGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTCACAACGT
TACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCT
GAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGTTTCACCG
CCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGACTACTCTGCTGCGGTGCTGGC
TGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGT

...

4,639,675 bp circular DNA

Working with text files

- Always use plain text editors: notepad etc.

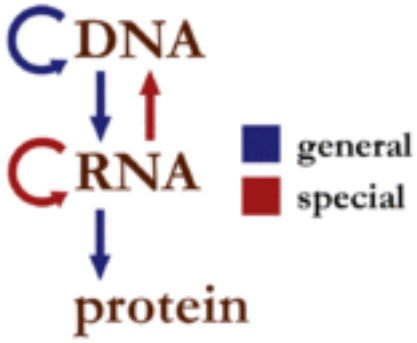
- ALL bioinformatics applications do NOT take sequence files saved in word or excel formats

Try notepad++

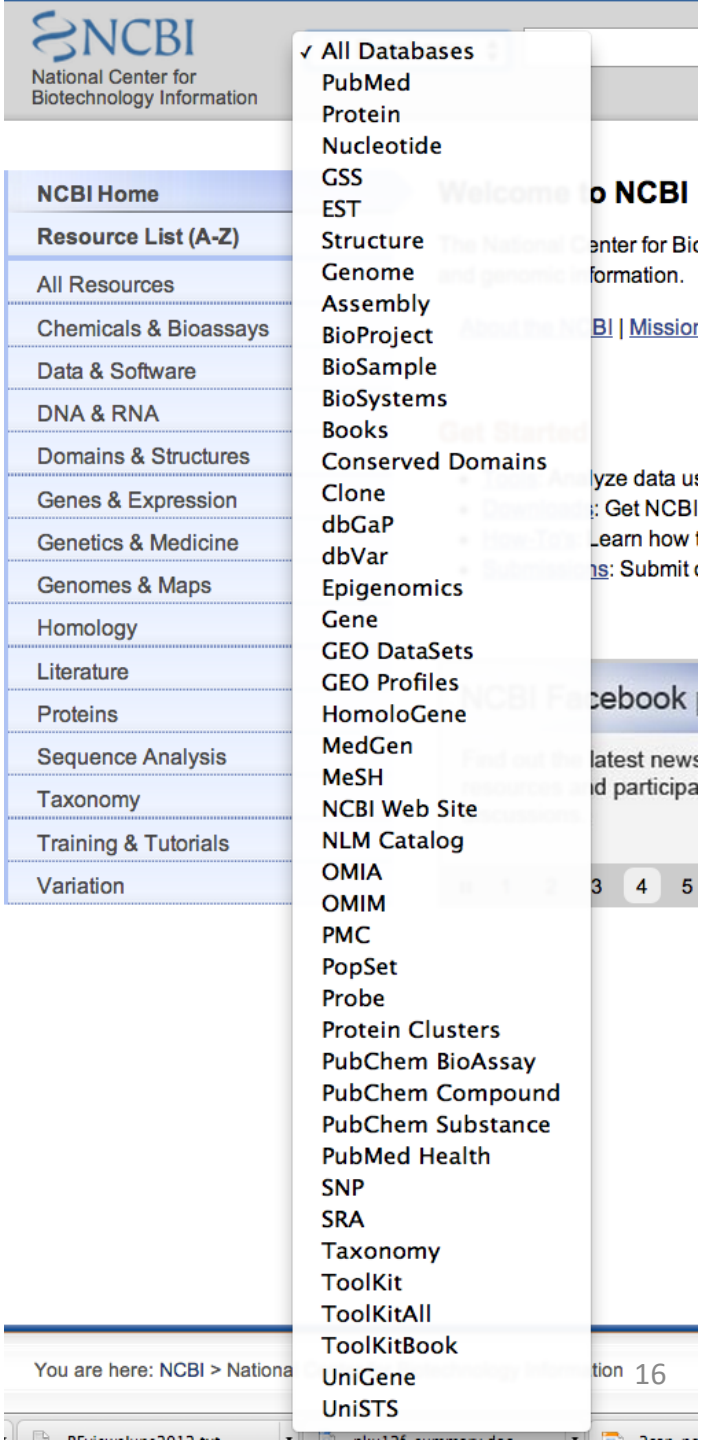
Web resource types

- Databases
- Web servers
- Genome browsers

Databases



- Sequence databases
  - Nucleotide Databases
  - Protein Databases
- Function databases
  - Protein domain/family
  - Function classification
  - Pathway
- Protein structure database PDB
- Gene expression databases
- Literature databases
- Taxonomic databases
- Other databases



Nucleotide Databases

- Genomic DNAs
- mRNAs or full length cDNAs
- EST (expressed sequence tags)
- SRA (short read archive)

GenBank
EMBL
DDBJ

Where the data come from:

- Large genome sequencing centers
- Individual research labs

The first genome sequenced ever

- **Haemophilus influenzae Rd KW20**

Frequency and distribution of DNA uptake signal sequences in the Haemophilus influenzae Rd genome. Smith HO, et al. Science 1995 Jul 28

http://www.ncbi.nlm.nih.gov/genome/165/?project_id=57771

All published sequences will go to GenBank eventually

- Most journals require an accession number of sequence data for submitted papers

http://genome.cshlp.org/site/misc/ifora_weblinks.xhtml

- After published, the paper is linked to data

http://www.ncbi.nlm.nih.gov/pubmed?term=21303537

Protein Databases

- Experimentally characterized proteins
- Computational predicted proteins (e.g. automatically predicted by gene finding programs)

Swiss-Prot
TrEMBL
PIR
PDB
GenPept
RefSeq
PRF

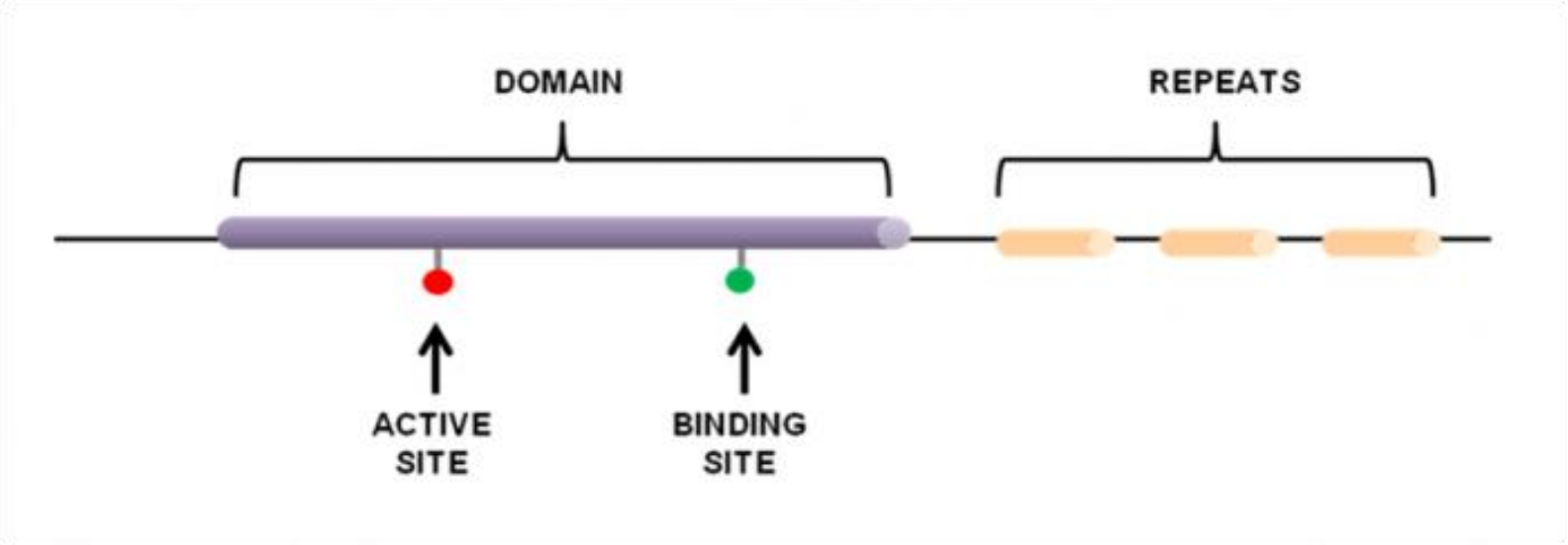
Where the data come from:

- Genome annotation efforts from large genome sequencing centers
- Individual research labs

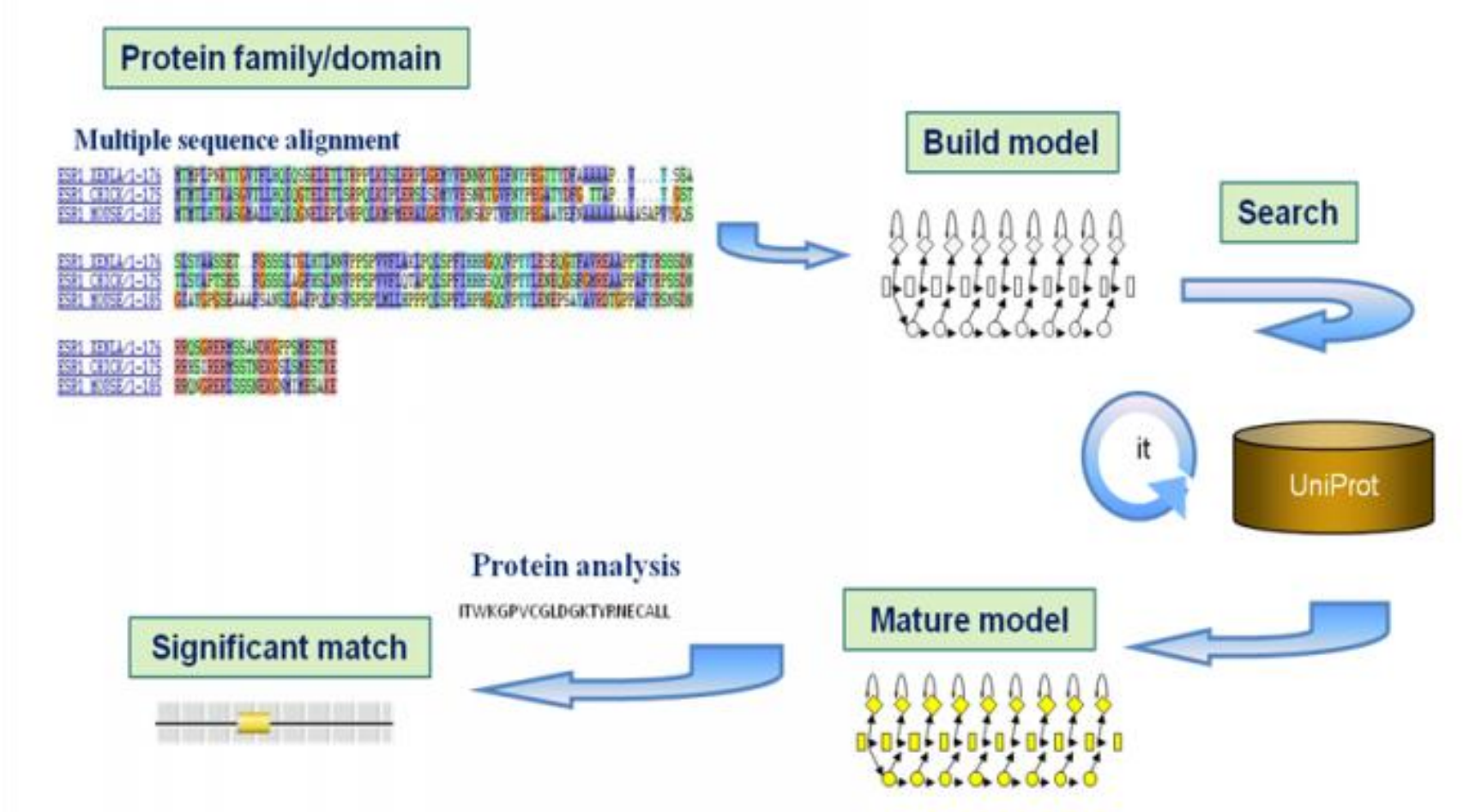
Protein function databases

All databases are built upon existing knowledge

- Start from function known proteins and literatures
- Retrieve functional domains/motifs/sites
- Retrieve homologs to form family
- Use statistical models to represent family
- Assign functions to family and link family to literatures

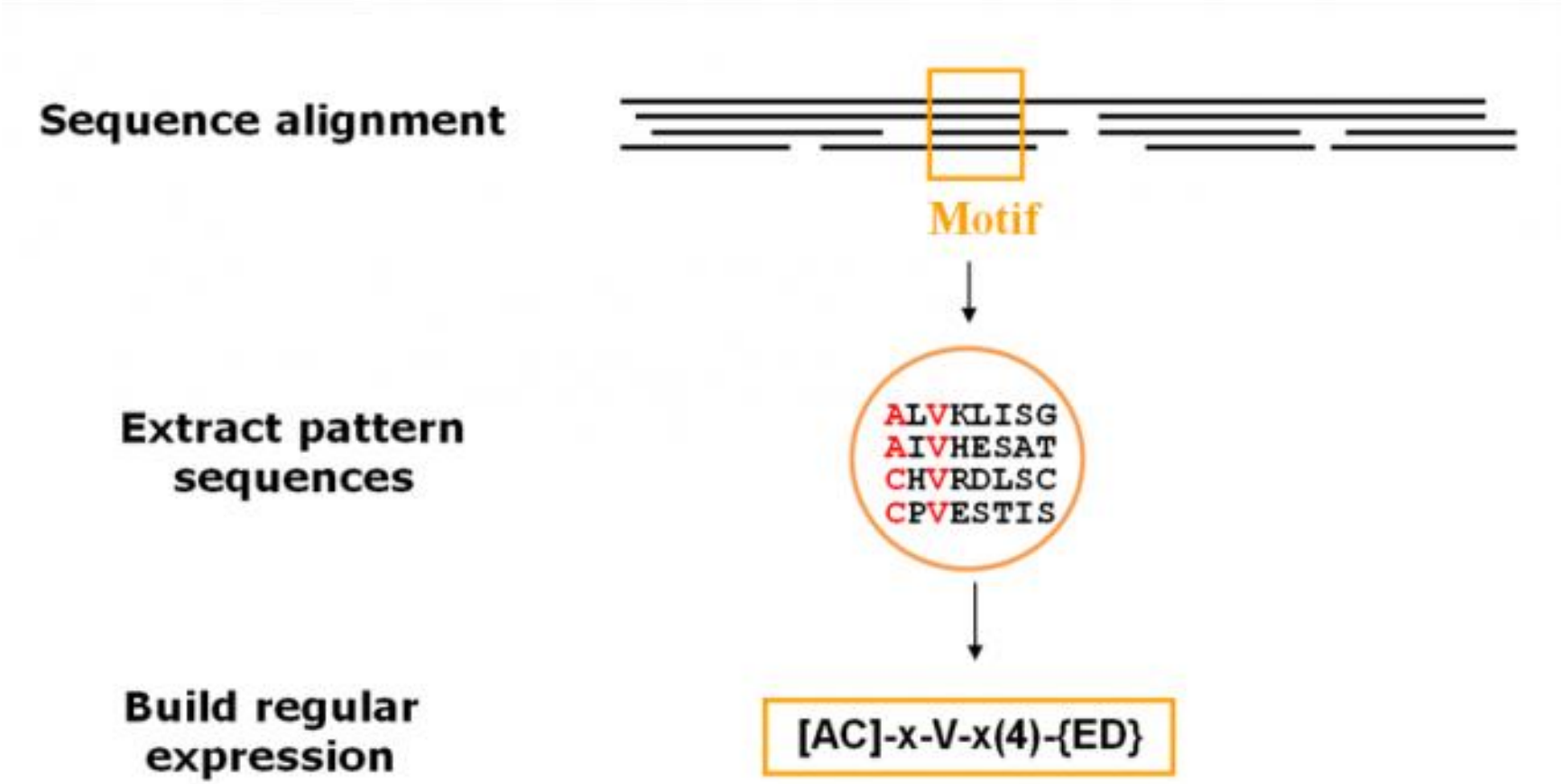


How protein family databases are built



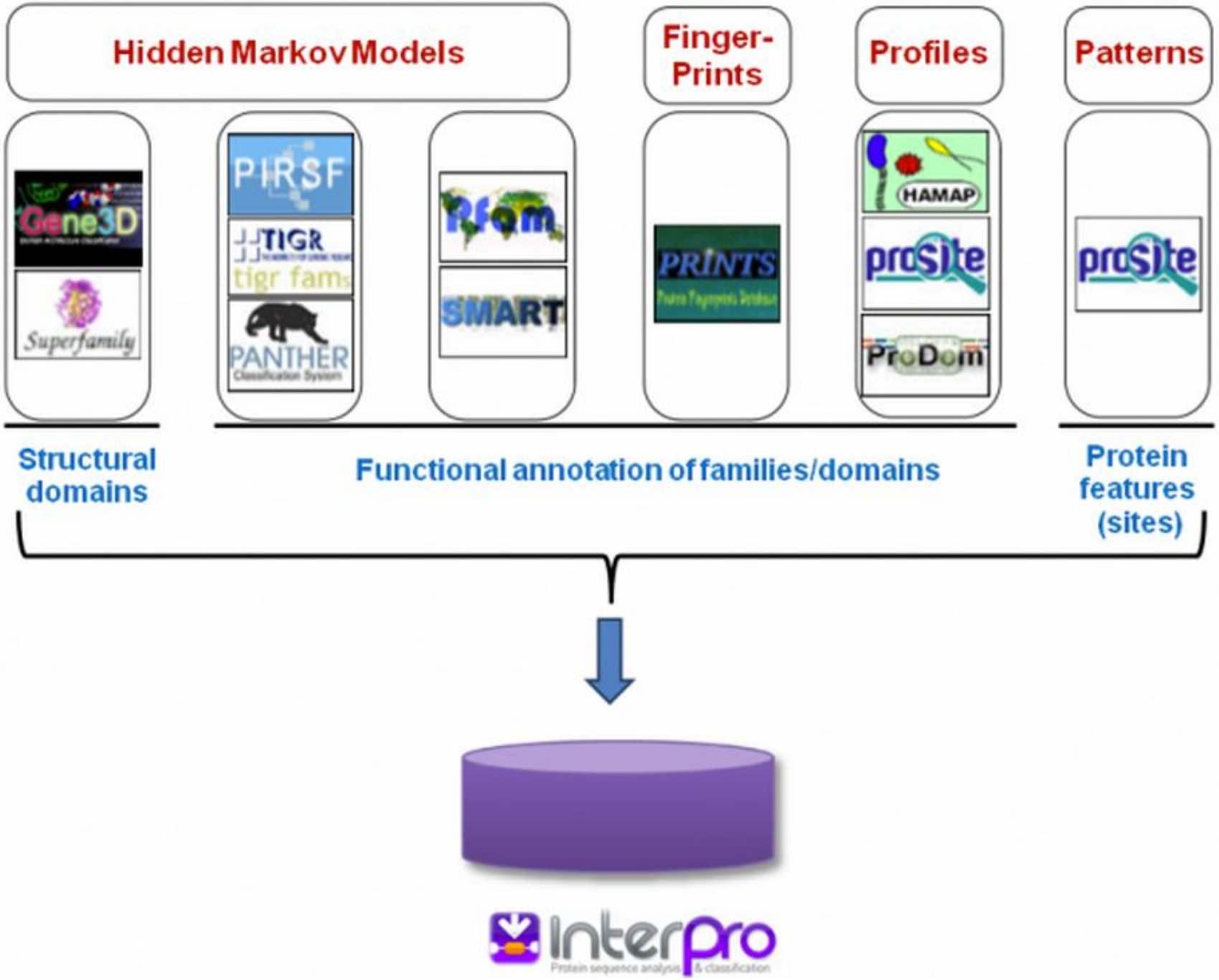
http://www.ebi.ac.uk/training/online/course/introduction-protein-classification-ebi

One example: PROSITE



[Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}

http://prosite.expasy.org/scanprosite/scanprosite_doc.html


Hidden Markov Models
Finger-Prints
Profiles
Patterns
Gene3D
Superfamily
PIRSF
TIGR
tigr fams
PANTHER
Classification System
Pfam
SMART
PRINTS
HAMAP
proSite
ProDom
proSite
Structural domains
Functional annotation of families/domains
Protein features (sites)
InterPro
Protein sequence analysis & classification

Over 1000 bioinformatics databases

Nucleic Acids Research's annual database special issue (20 years already)

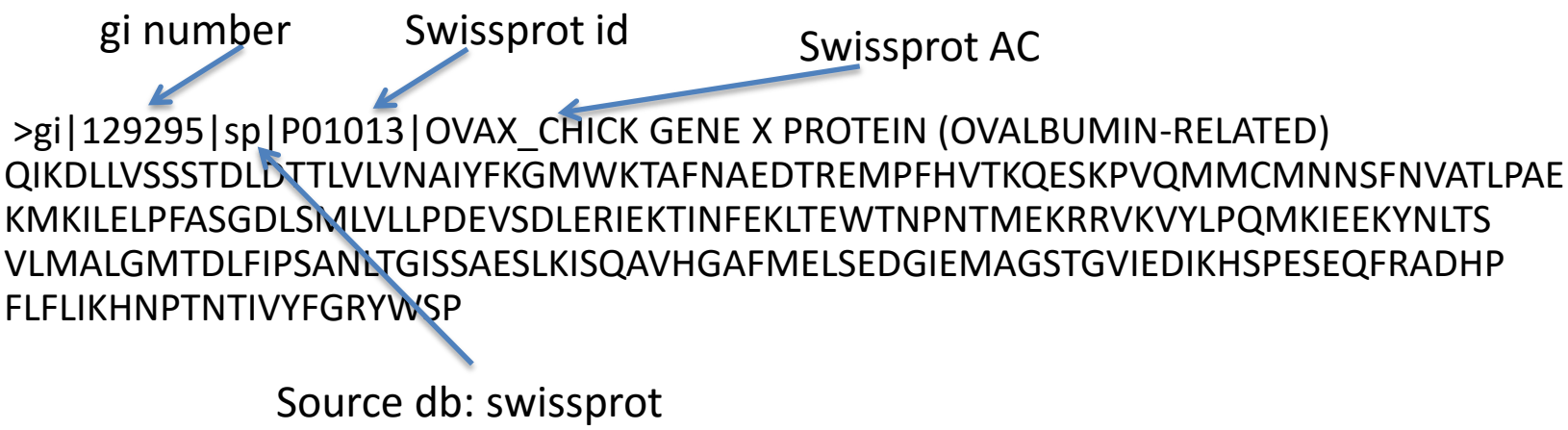
http://www.oxfordjournals.org/nar/database/c

Sequence formats (all plain text)

- FASTA
- GenBank
- EMBL
- PDB
- Alignment format
- http://emboss.sourceforge.net/docs/themes/SequenceFormats.html

FASTA format

A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line (defline) is distinguished from the sequence data by a greater-than (">") symbol at the beginning. An example sequence in FASTA format is:



FastQ format

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Four lines:

1. Sequence id
2. Sequence
3. +
4. Quality score as ASCII characters

Nucleic Acids Res. 2010 April; 38(6): 1767–1771

GenBank and EMBL formats

Much richer info: many different fields for annotation of the protein/DNA

http://www.ncbi.nlm.nih.gov/protein/129295

http://www.uniprot.org/uniprot/P01013

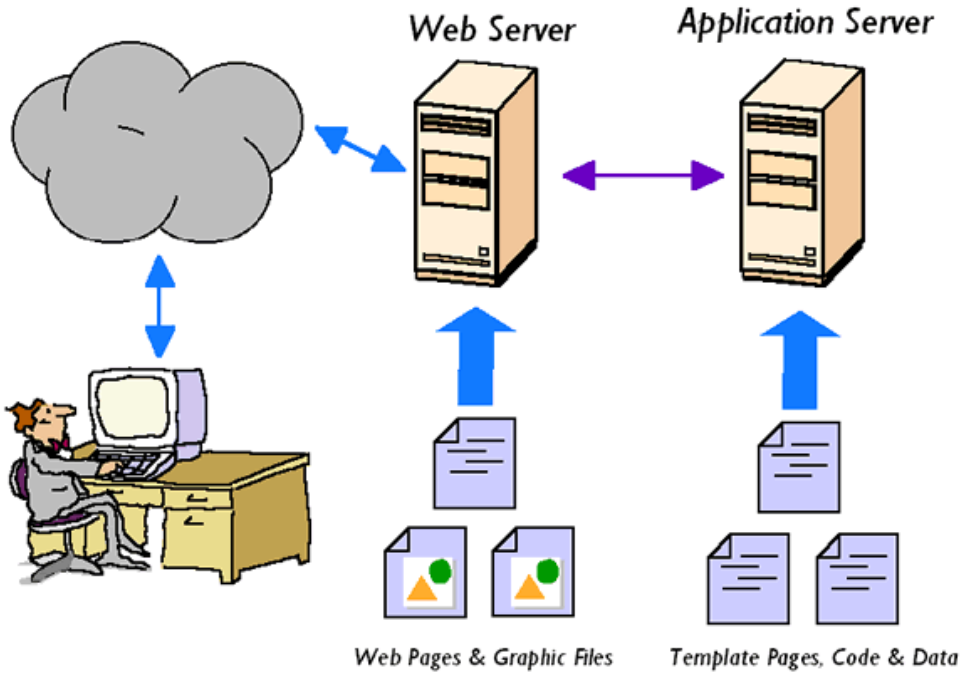
PDB format

- PDB: Protein Data Bank, a protein 3D structure database

http://www.rcsb.org/pdb/explore/explore.do?structureId=1GOT

Web servers

- Remote servers that allow users to upload data, invoke bioinfo softwares running on the remote server and return the results to the user with a graphical web interface



Over 1000 bioinformatics web servers

Nucleic Acids Research's annual Web Server special issue (10 years already)

http://nar.oxfordjournals.org/content/40/W1/W3

This one might be better:

http://www.hsls.pitt.edu/obrc/

Genome browsers

- Integrated graphical presentation of genomes to allow visualize and browse entire genomes with annotated data including gene prediction and structure, proteins, expression, regulation, variation, comparative analysis, etc

Examples: NCBI MapViewer, UCSC Genome Browser, ENSEMBL Genome Browser

http://en.wikipedia.org/wiki/Genome_browser

Genomic databases

- Microbial genomes:
  - NCBI, JGI IMG
- Plant genomes: JGI phytozome
- Animal genomes: EBI ensembl, UCSC genome browser
- Fungal genomes: JGI and Broad Institute

Next lecture: NCBI resources I