Introduction to Bioinformatics
BINF 630

Dr. Andrew Carr

September 6, 2006

Lecture 2: Sequencing, information sharing and databases

The beginnings of bioinformatics data.

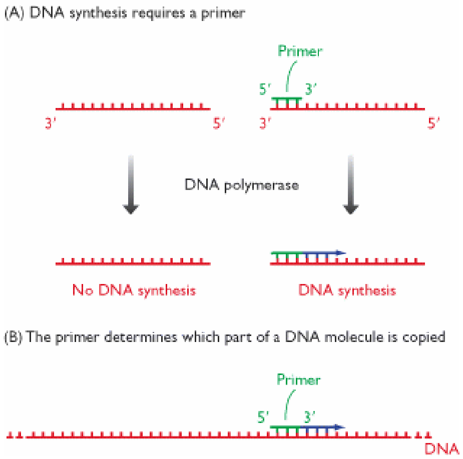
- The raw signals in bioinformatics
  - In the form of sequences and structures
    - DNA
    - RNA
    - Proteins
    - Other …
      - Metabolic rates (Cellular modeling)
      - Phylogenitic information (Genetic history / evolution)
      - Phenotypic information (Gene expression)
      - Pathway participation
      - ….
- Where do we get this information?
    - Lab
    - Shared resources (data warehouse)
- What do we do with the information once we have it?
    - Computational analysis…

In the lab…

- DNA and RNA sequencing
  - PCR (Polymerase Chain Reaction)
  - Contig and Genome cloning
- Protein Sequencing

The Basics of DNA Sequencing

- Primers
  - Target specific regions
  - 20 to 30 bases in length
  - Indicate the portion of the DNA to be copied


(A) DNA synthesis requires a primer

(B) The primer determines which part of a DNA molecule is copied

Image from Genomes2 © 2002 Garland Science

Amplification of DNA sequences

- PCR (Polymerase Chain Reaction)
  - Fast and fairly error free
  - PCR used in forensic work
    - Needs very little DNA to start
  - PCR requires choice of primers
  - Rapid amplification of the target.
    - 30 cycles yields more than 250 million targets
- Cloning
  - Slower
  - More error prone
  - Does not produce as many copies

PCR

- One Cycle
  - Target strand heated to ~ 94C
  - Primers are added for targeted region.
  - Taq Polymerase copies sequence

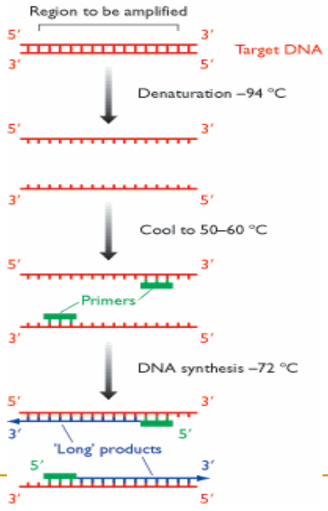


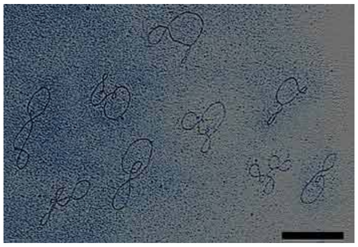
Image from Genomes2 © 2002 Garland Science

Why not use PCR for Genomic Sequencing?

- Genomic target to large
  - PCR works well for short seqeunces
    - aroung 5kb is standard
    - Up to 40 with improved technique
  - For sequences > 100 kb not possible

Cloning

- Uses small organism to grow multiple copies of the DNA
- Restriction enzymes used to cleave DNA so that the desired sequence can be inserted.
- Organisms natural DNA replication process produces copies.
- YAC and BAC cloning vectors are used.



Scott Camazine, PLASMID DNA (2000).

Comparison between YAC and BAC Cloning Systems

Features	YAC	BAC
Configuration	Linear	Circular
Host	Yeast	Bacteria
Cloning Capacity	Unlimited	up-to 350
Insert Stability	Unstable	Stable
Copy per Cell	1	1-2
Chimerism	up to 40%	None to low

The Cloning Process

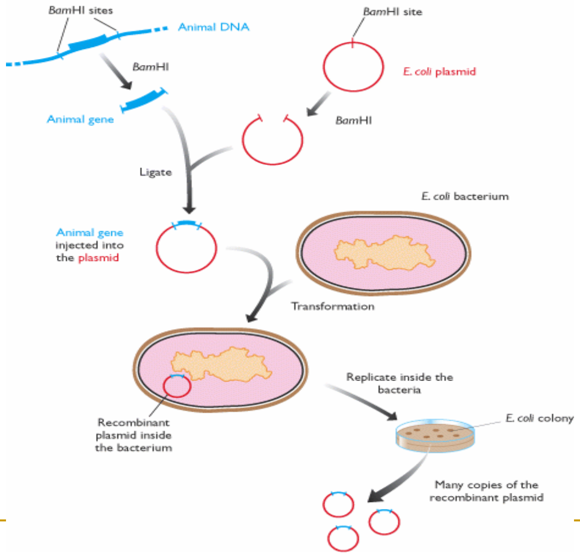

BamHI sites
Animal DNA
BamHI site
E. coli plasmid
BamHI
Animal gene
BamHI
Ligate
E. coli bacterium
Animal gene injected into the plasmid
Transformation
Replicate inside the bacteria
E. coli colony
Recombinant plasmid inside the bacterium
Many copies of the recombinant plasmid

Image from Genomes2 © 2002 Garland Science

Basic Sequencing

- Tagging of nucleotides
  - dideoxynucleotide terminal markers
- DNA sequences grown from primer
  - Sequences terminate when a dideoxynucleotide gets placed in the sequence
- Electrophoresis
  - Porous agar gel
  - Ion charge and pores control movement
    - Distance a chain travel correlates to sequence length
  - Color of tag indicates type of base called
- Good gel tutorial web site:
  - http://learn.genetics.utah.edu/units/biotech/gel/

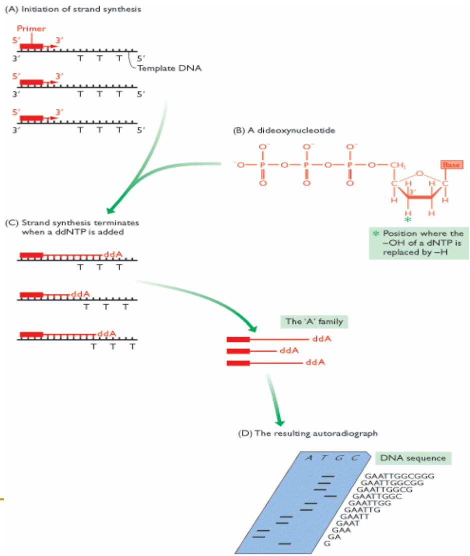
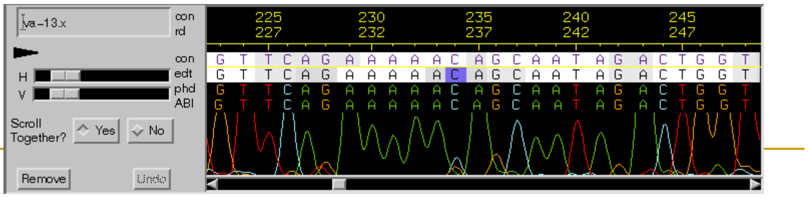


Image from Genomes2 © 2002 Garland Science

Getting the sequence

- Laser reads the signals from fluorescent dyes.
  - ABI sequencer
- Produces a signal file
- Errors in sequencing
  - End of gel run less precise
  - Lane shift
- Phred/ Phrap/ Consed
  - Basic industry standard tools for base calling
    - Expert makes the final decision


```
va–13.x        con      225       230       235       240       245
               rd       227       232       237       242       247
               con   G T T C A G A A A A A C A G C A A T A G A C T G G T
               edt   G T T C A G A A A A A C A G C A A T A G A C T G G T
               phd   G T T C A G A A A A A C A G C A A T A G A C T G G T
               ABI   G T T C A G A A A A A C A G C A A T A G A C T G G T
```

Constructing Larger Sequences

- Shotgun Approach to sequencing

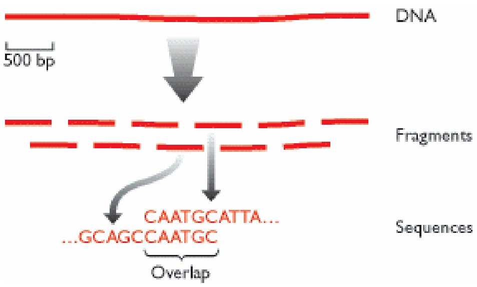


Image from Genomes2 © 2002 Garland Science

Genomic Sequencing

- Contig
  - Defined segment
  - Shotgun segment
- Whole Genome Shotgun
  - Makes use of markers to keep location
- Challenges
  - repeated DNA section
  - SNP
  - Markerless regions

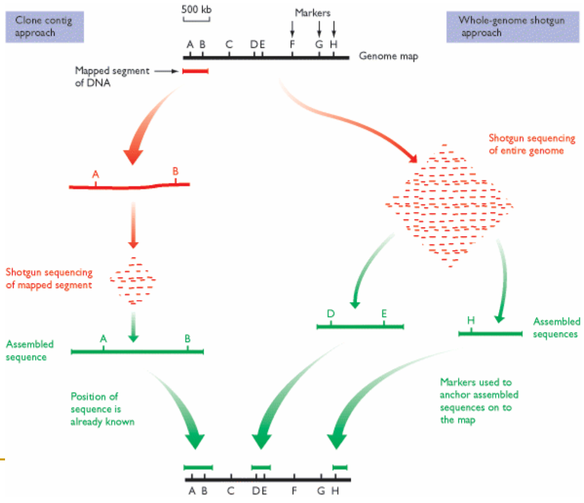
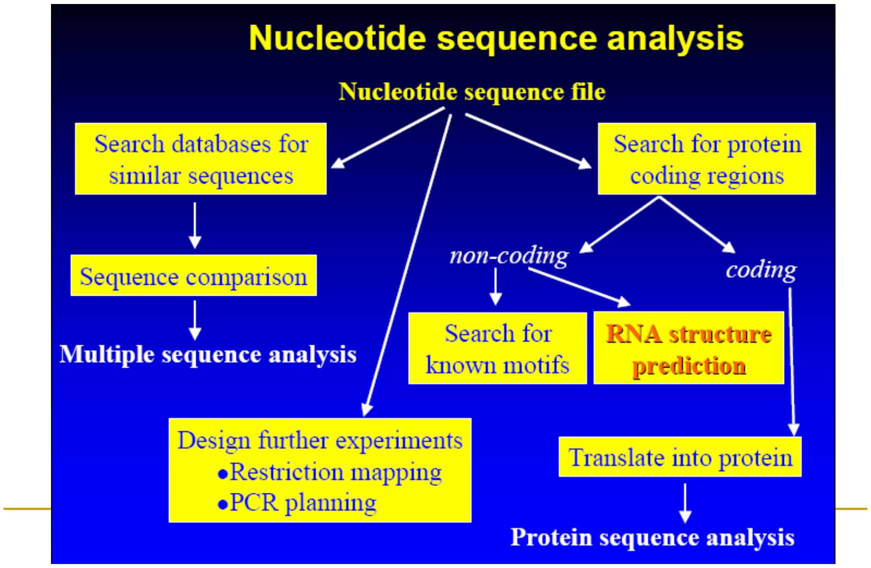

Clone contig approach
500 kb
Markers
Whole-genome shotgun approach
A B C DE F G H
Genome map
Mapped segment of DNA
A
B
Shotgun sequencing of entire genome
Shotgun sequencing of mapped segment
D
E
H
Assembled sequences
Assembled sequence
A
B
Position of sequence is already known
Markers used to anchor assembled sequences on to the map
A B C DE F G H

Image from Genomes2 © 2002 Garland Science

What to do with our “knowledge”

- Share it…
  - Information is knowledge or intelligence communicated, received or gained
- Learn more about it…

Further research directions



Nucleotide sequence analysis

Nucleotide sequence file

Search databases for similar sequences

Sequence comparison

Multiple sequence analysis

Search for protein coding regions

non-coding

coding

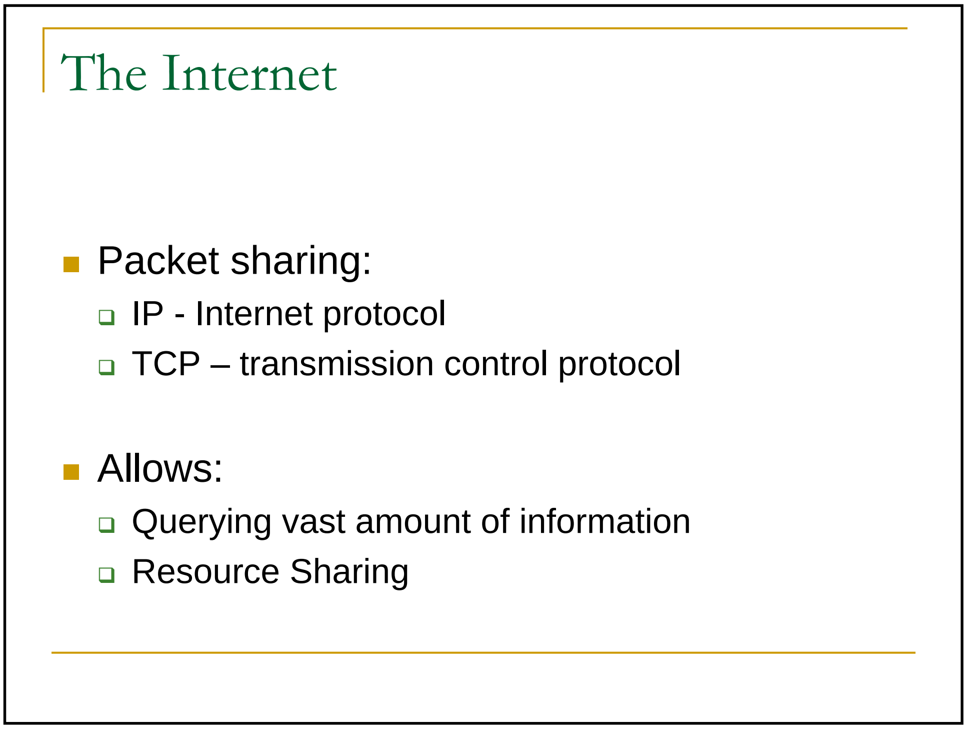
Search for known motifs

RNA structure prediction

Design further experiments
- Restriction mapping
- PCR planning

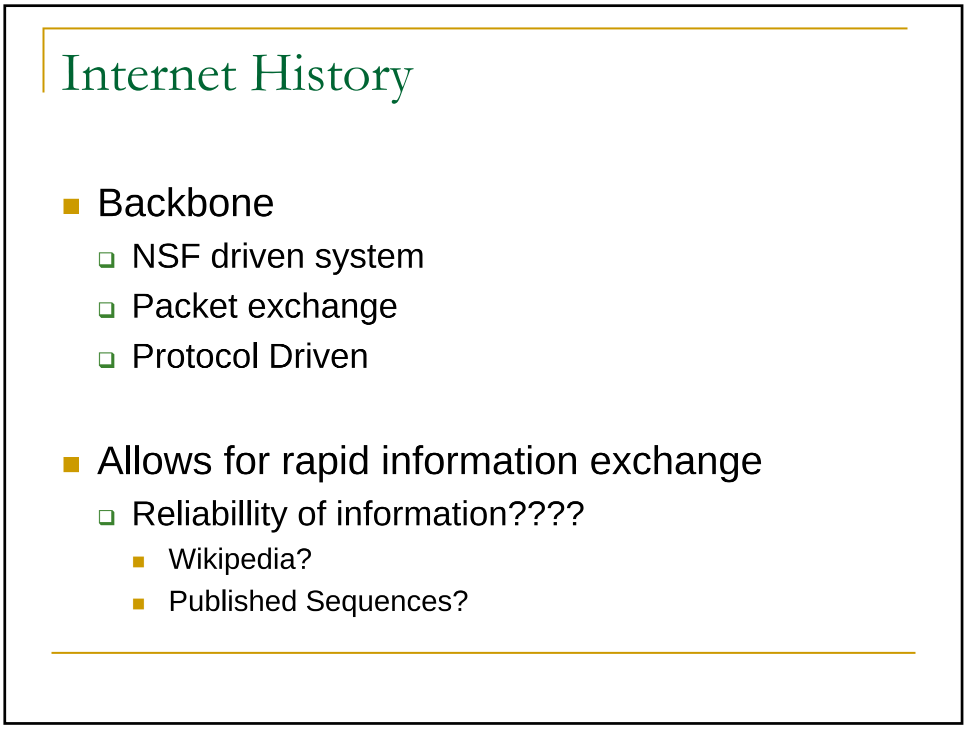
Translate into protein

Protein sequence analysis



The Internet

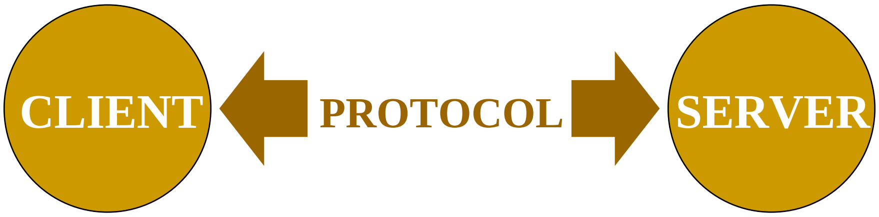
- Packet sharing:
  - IP - Internet protocol
  - TCP – transmission control protocol
- Allows:
  - Querying vast amount of information
  - Resource Sharing



Internet History

- Backbone
  - NSF driven system
  - Packet exchange
  - Protocol Driven
- Allows for rapid information exchange
  - Reliabillity of information????
    - Wikipedia?
    - Published Sequences?

Basic information exchange format



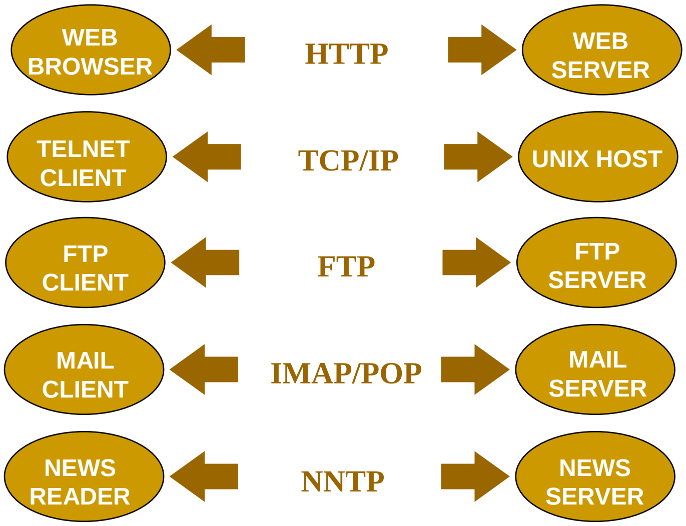
CLIENT

PROTOCOL

SERVER

The protocol or “middle layer” is a translator that is essential to communication.

Common Protocols



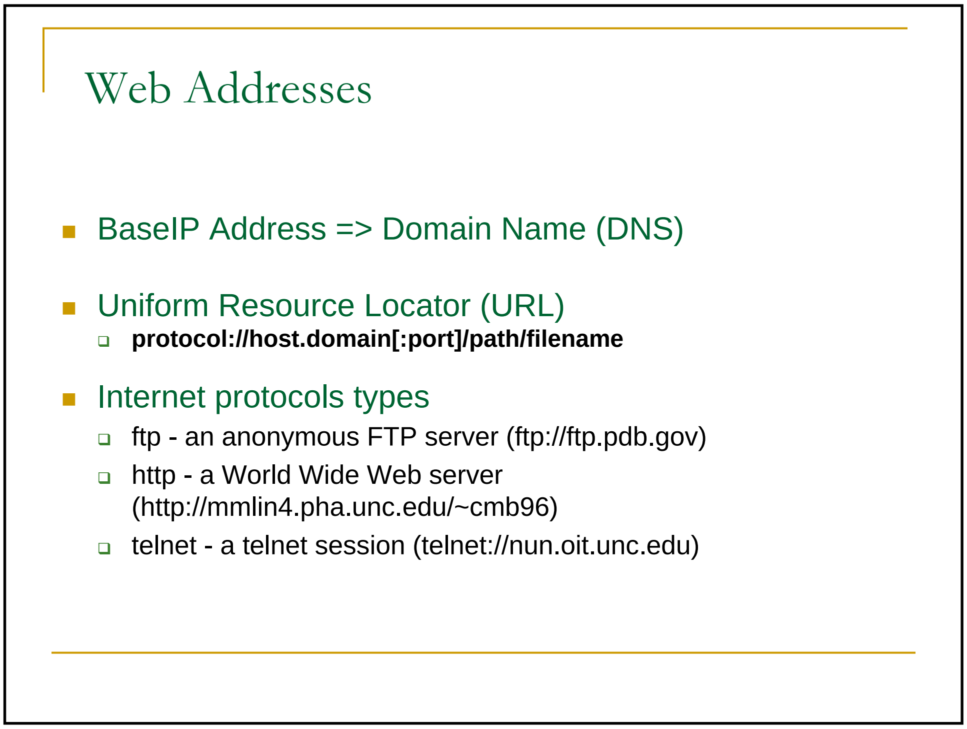
WEB BROWSER ← HTTP → WEB SERVER

TELNET CLIENT ← TCP/IP → UNIX HOST

FTP CLIENT ← FTP → FTP SERVER

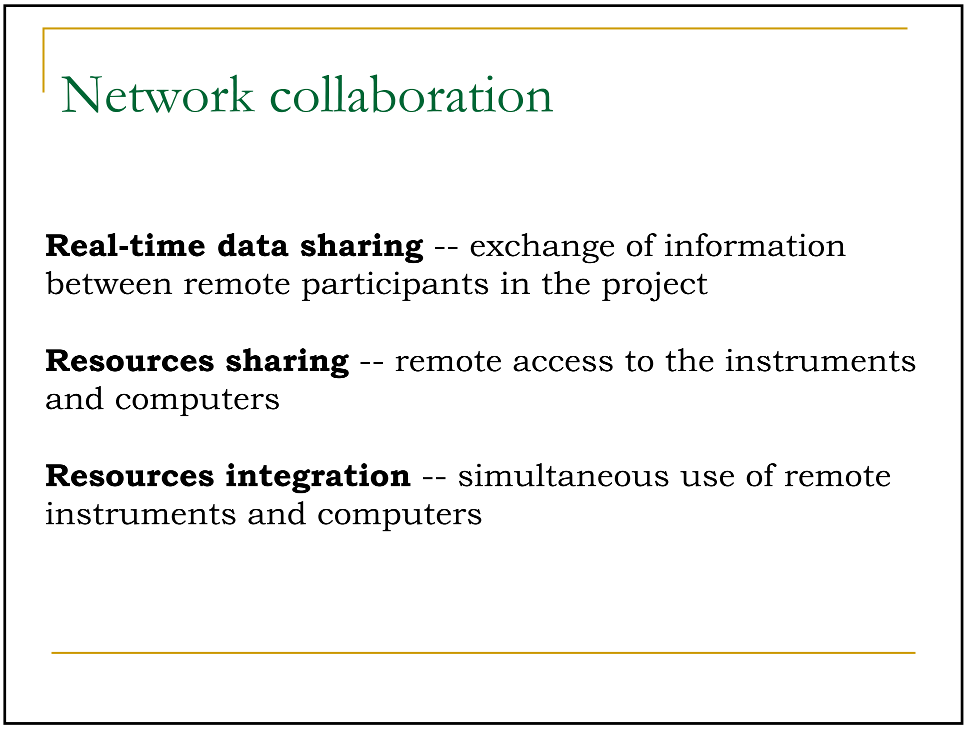
MAIL CLIENT ← IMAP/POP → MAIL SERVER

NEWS READER ← NNTP → NEWS SERVER



Web Addresses

- BaseIP Address => Domain Name (DNS)
- Uniform Resource Locator (URL)
  - **protocol://host.domain[:port]/path/filename**
- Internet protocols types
  - ftp - an anonymous FTP server (ftp://ftp.pdb.gov)
  - http - a World Wide Web server (http://mmlin4.pha.unc.edu/~cmb96)
  - telnet - a telnet session (telnet://nun.oit.unc.edu)

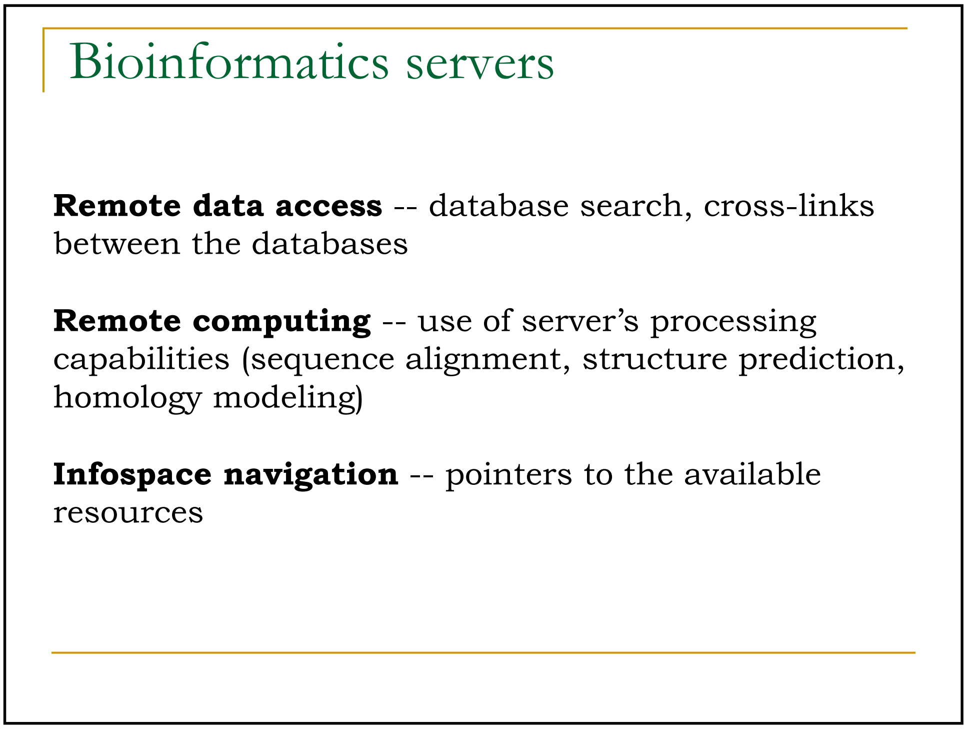


Network collaboration

Real-time data sharing -- exchange of information between remote participants in the project

Resources sharing -- remote access to the instruments and computers

Resources integration -- simultaneous use of remote instruments and computers

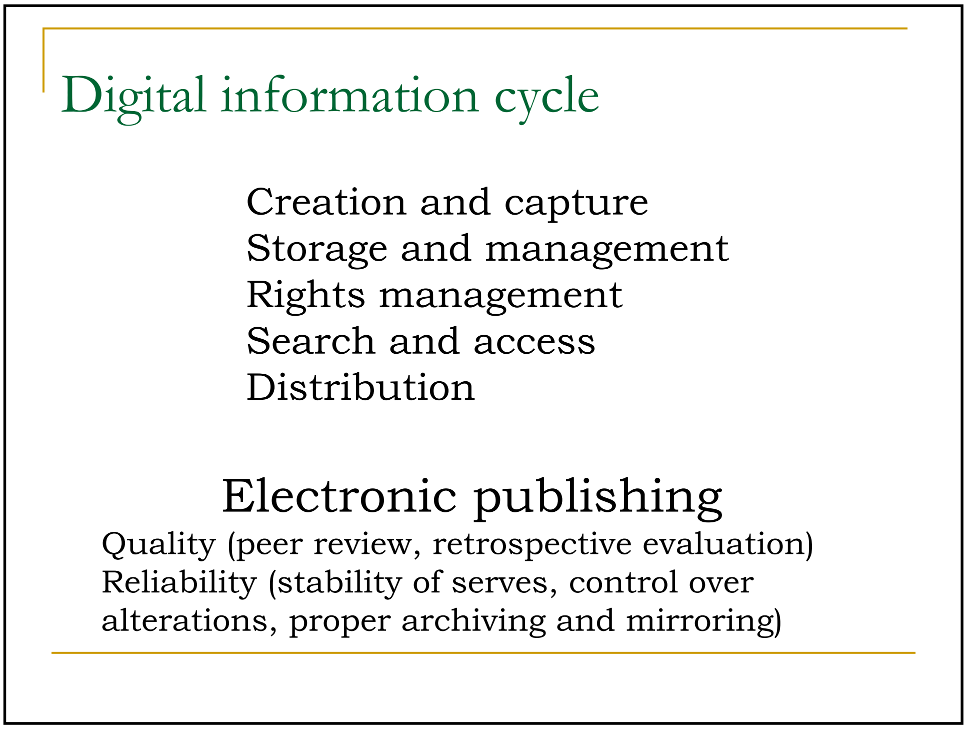


Bioinformatics servers

Remote data access -- database search, cross-links between the databases

Remote computing -- use of server’s processing capabilities (sequence alignment, structure prediction, homology modeling)

Infospace navigation -- pointers to the available resources

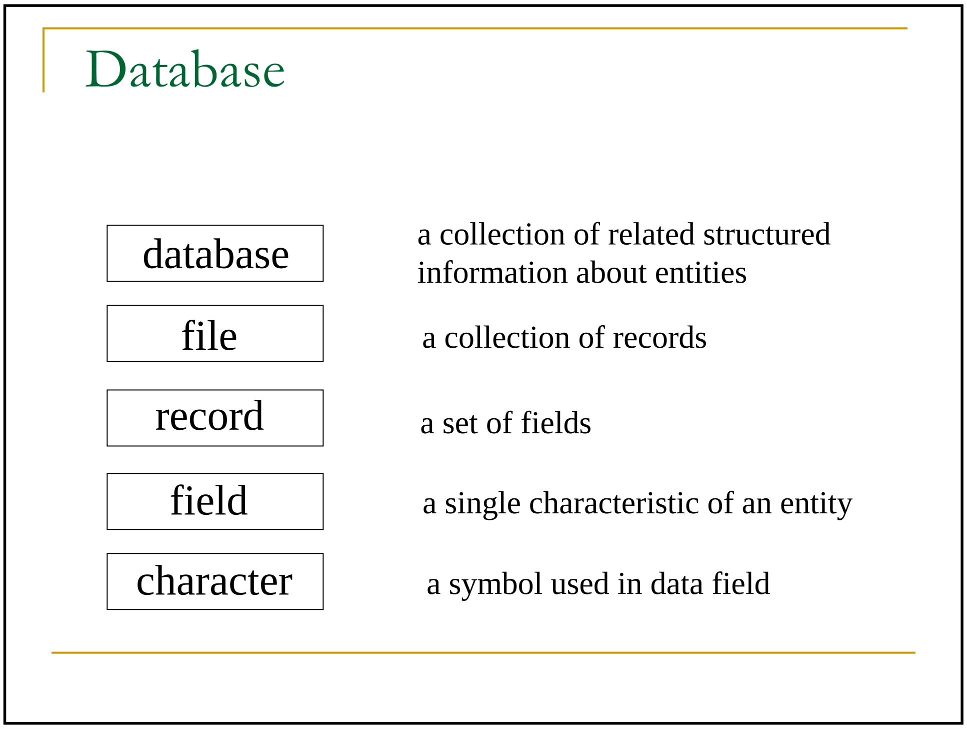


Digital information cycle

Creation and capture
Storage and management
Rights management
Search and access
Distribution

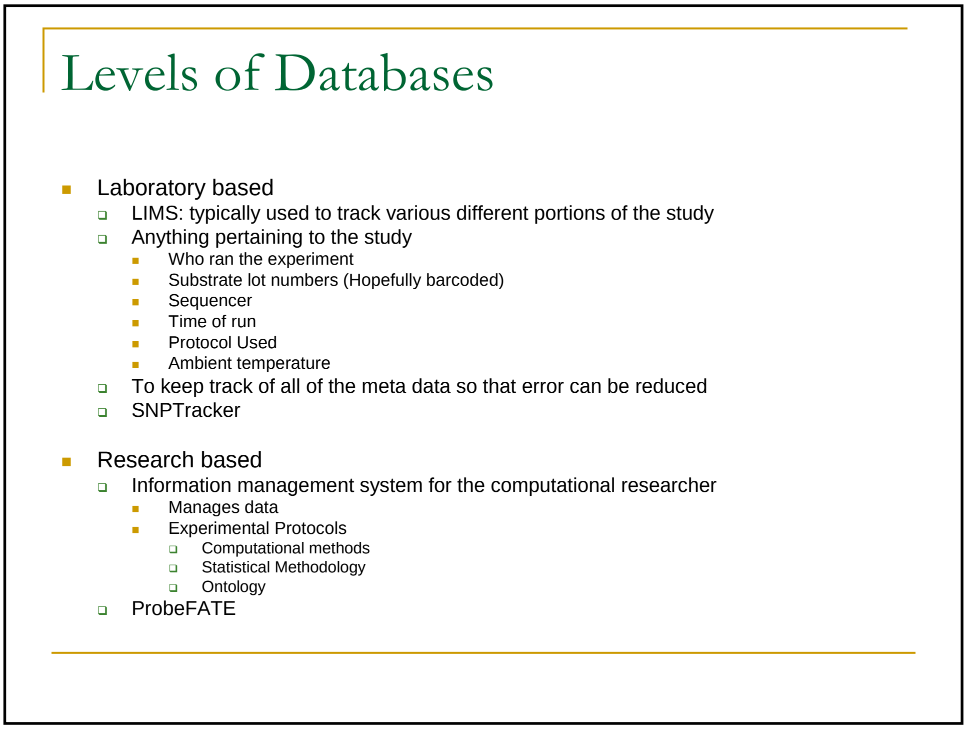
Electronic publishing

Quality (peer review, retrospective evaluation)
Reliability (stability of serves, control over alterations, proper archiving and mirroring)



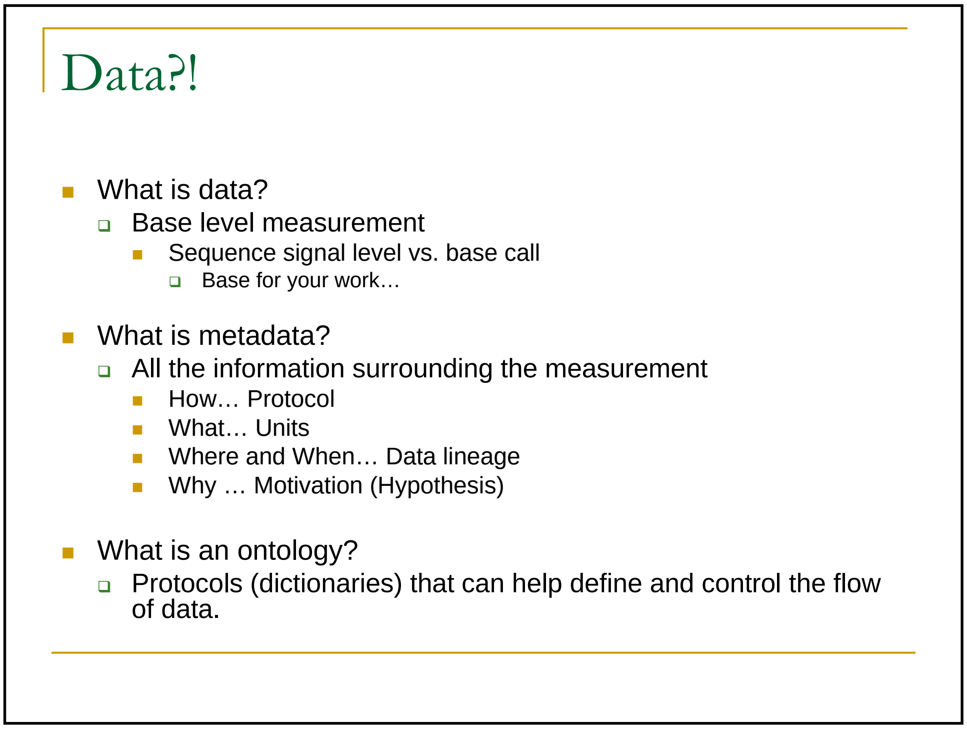
Database

database	a collection of related structured information about entities
file	a collection of records
record	a set of fields
field	a single characteristic of an entity
character	a symbol used in data field



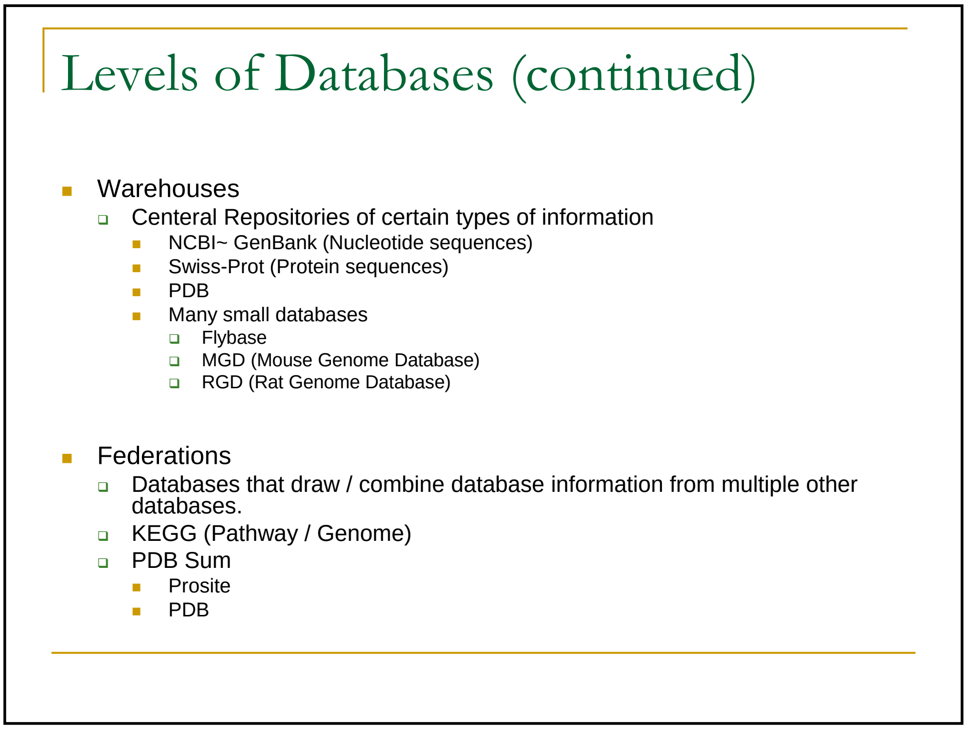
Levels of Databases

- Laboratory based
  - LIMS: typically used to track various different portions of the study
  - Anything pertaining to the study
    - Who ran the experiment
    - Substrate lot numbers (Hopefully barcoded)
    - Sequencer
    - Time of run
    - Protocol Used
    - Ambient temperature
  - To keep track of all of the meta data so that error can be reduced
  - SNPTracker
- Research based
  - Information management system for the computational researcher
    - Manages data
    - Experimental Protocols
      - Computational methods
      - Statistical Methodology
      - Ontology
  - ProbeFATE



Data?!

- What is data?
  - Base level measurement
    - Sequence signal level vs. base call
      - Base for your work…
- What is metadata?
  - All the information surrounding the measurement
    - How… Protocol
    - What… Units
    - Where and When… Data lineage
    - Why … Motivation (Hypothesis)
- What is an ontology?
  - Protocols (dictionaries) that can help define and control the flow of data.



Levels of Databases (continued)

- Warehouses
  - Centeral Repositories of certain types of information
    - NCBI~ GenBank (Nucleotide sequences)
    - Swiss-Prot (Protein sequences)
    - PDB
    - Many small databases
      - Flybase
      - MGD (Mouse Genome Database)
      - RGD (Rat Genome Database)
- Federations
  - Databases that draw / combine database information from multiple other databases.
  - KEGG (Pathway / Genome)
  - PDB Sum
    - Prosite
    - PDB

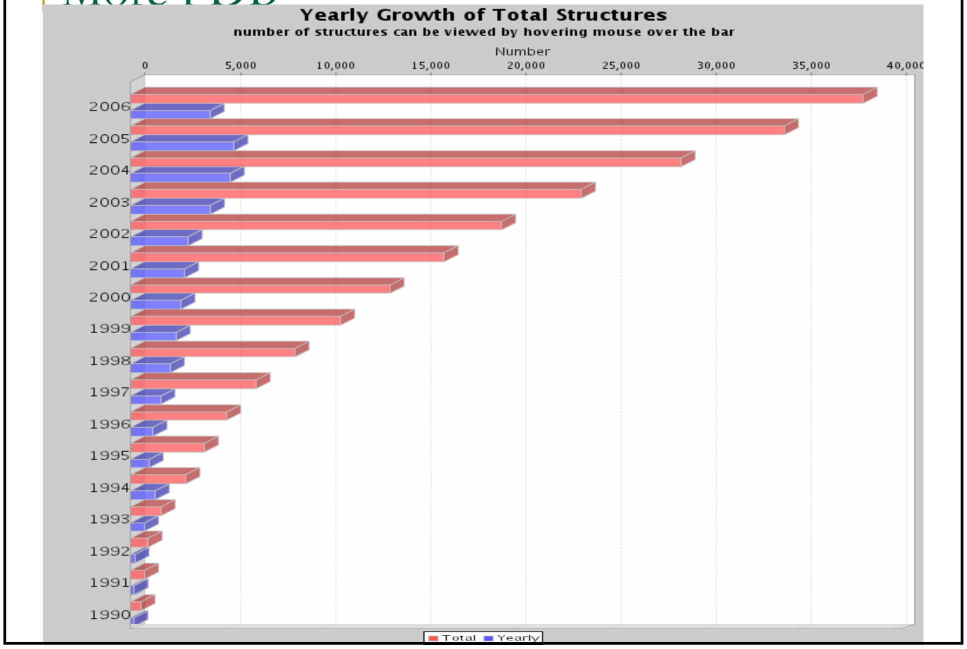
Important Central Databases

- Genome
  - NCBI
    - GenBank
    - PubMed
  - European Molecular Biology Laboratory (EMBL)
  - DNA Database of Japan (DDBJ).
  - GO (Gene Ontology)
    - Consortium of databases
      - Flybase, RGD, MGD,….
- Protein
  - RCSB ~ PDB
  - EMBL ~ EBI (European Bioinformatics Institute)
    - UniProt ~ ExPasy ~ Swiss-Prot
- KEGG: Kyoto Encyclopedia of Genes and Genomes

PDB at RCSB

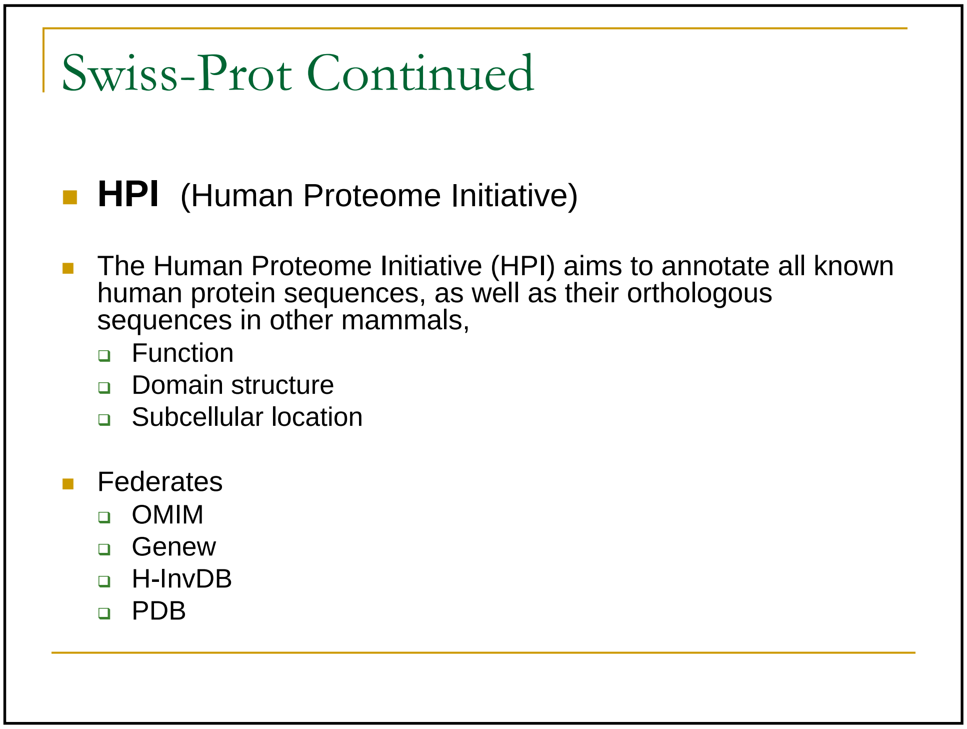
- The Protein Data Bank (PDB) is the single worldwide depository of information about the three-dimensional structures of large biological molecules, including proteins and nucleic acids. These are the molecules of life that are found in all organisms including bacteria, yeast, plants, flies, and mice, and in healthy as well as diseased humans. Understanding the shape of a molecule helps to understand how it works.
- In 1998, the Research Collaboratory for Structural Bioinformatics (RCSB) became responsible for the management of the PDB.
- The PDB was established in 1971 at Brookhaven National Laboratory and originally contained 7 structures.
  - New structures released every Wednesday
  - As of September 5, 2006 there were 38620 Structures
- PDB provides
  - Sequence
  - Atomic Coordinates
  - Derived geometric data
  - Secondary Structure Content
  - Annotations about protein literature references
- http://www.pdb.org/.
- http://www.rcsb.org/pdb/explore.do?structureId=1GBG

More PDB



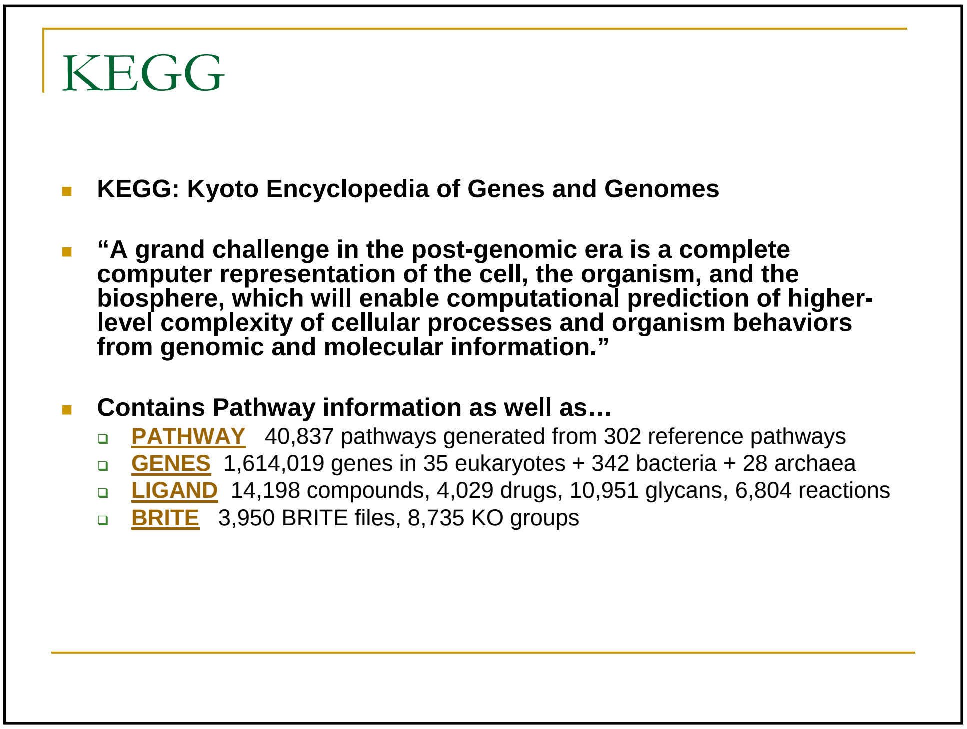
SWISS-PROT

- The **ExPASy** (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures
- **UniProtKB/Swiss-Prot**; a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases
- **UniProtKB/TrEMBL**; a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.
- **UniProtKB/Swiss-Prot Release 50.6 of 05-Sep-2006: 231,434 entries**
- **UniProtKB/TrEMBL Release 33.6 of 05-Sep-2006: 3,182,016 entries**
- http://www.expasy.org/sprot/



Swiss-Prot Continued

- **HPI** (Human Proteome Initiative)
- The Human Proteome Initiative (HPI) aims to annotate all known human protein sequences, as well as their orthologous sequences in other mammals,
  - Function
  - Domain structure
  - Subcellular location
- Federates
  - OMIM
  - Genew
  - H-InvDB
  - PDB



KEGG

- **KEGG: Kyoto Encyclopedia of Genes and Genomes**
- **“A grand challenge in the post-genomic era is a complete computer representation of the cell, the organism, and the biosphere, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and molecular information.”**
- **Contains Pathway information as well as…**
  - **PATHWAY** 40,837 pathways generated from 302 reference pathways
  - **GENES** 1,614,019 genes in 35 eukaryotes + 342 bacteria + 28 archaea
  - **LIGAND** 14,198 compounds, 4,029 drugs, 10,951 glycans, 6,804 reactions
  - **BRITE** 3,950 BRITE files, 8,735 KO groups

PDBSUM

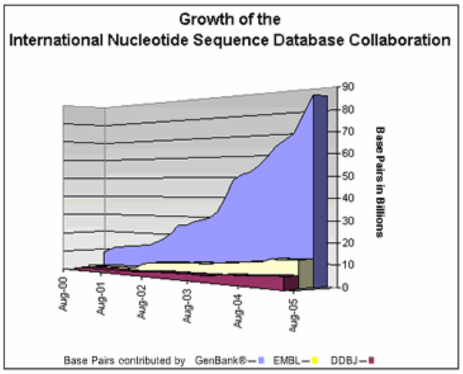
- The **PDBsum** is a pictorial database that provides an at-a-glance overview of the contents of each 3D structure deposited in the Protein Data Bank (PDB).
- Federates tools and data
  - PDB
  - Prosite
  - Rasmol/Chime
- http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/

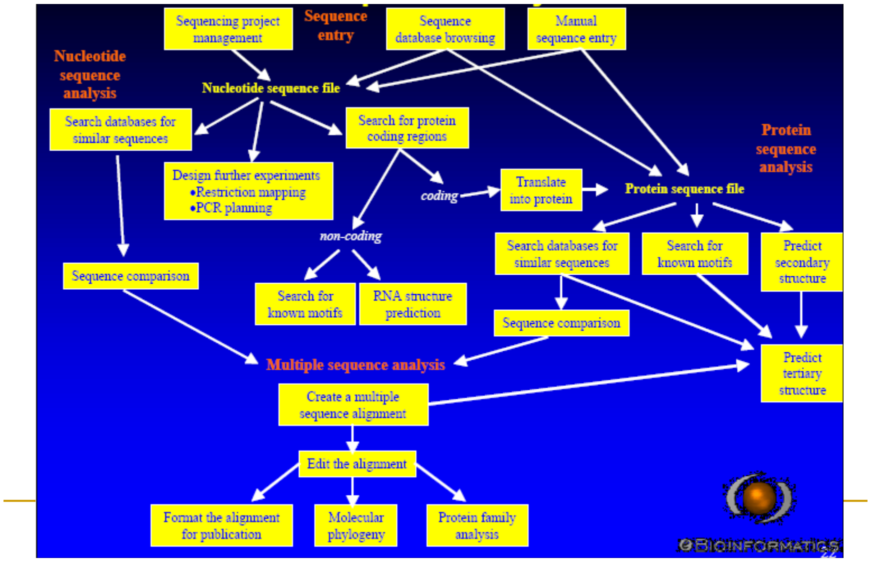
NCBI (National Center for Biotechnology Information)

- “Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease.”
- Databases
  - Sequence
    - GeneBank
    - SNP
    - GEO
    - MMDB
  - Literature
    - PubMed
    - Online Mendelian Inheritance in Man (OMIM)
    - Molecular Modeling Database (MMDB)
    - Unique Human Gene Sequence Collection (UniGene)
  - Gene Map of the Human Genome
  - the Taxonomy Browser
  - the Cancer Genome Anatomy Project (CGAP), in collaboration with the National Cancer Institute.
- Tools
  - Entrez is NCBI's search and retrieval system that provides users with integrated access to sequence, mapping, taxonomy, and structural data
    - http://www.ncbi.nlm.nih.gov/Database/

GenBank

- **GenBank®** is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences
- Repository of nucleotide sequences
  - From (EMBL) and (DDBJ)
  - As of April 2006, there are over 130 billion bases in GenBank and RefSeq alone
- Many journals require submission of sequence information to a database prior to publication so that an accession number may appear in the paper.
- Basic Local Alignment Search Tool (**BLAST**) finds regions of local similarity between sequences
- Sample Record





Sequencing project management

Sequence entry

Sequence database browsing

Manual sequence entry

Nucleotide sequence analysis

Nucleotide sequence file

Search databases for similar sequences

Search for protein coding regions

Design further experiments
- Restriction mapping
- PCR planning

coding

Translate into protein

Protein sequence file

Protein sequence analysis

non-coding

Sequence comparison

Search for known motifs

RNA structure prediction

Search databases for similar sequences

Search for known motifs

Predict secondary structure

Sequence comparison

Multiple sequence analysis

Predict tertiary structure

Create a multiple sequence alignment

Edit the alignment

Format the alignment for publication

Molecular phylogeny

Protein family analysis

Why informatics is needed…

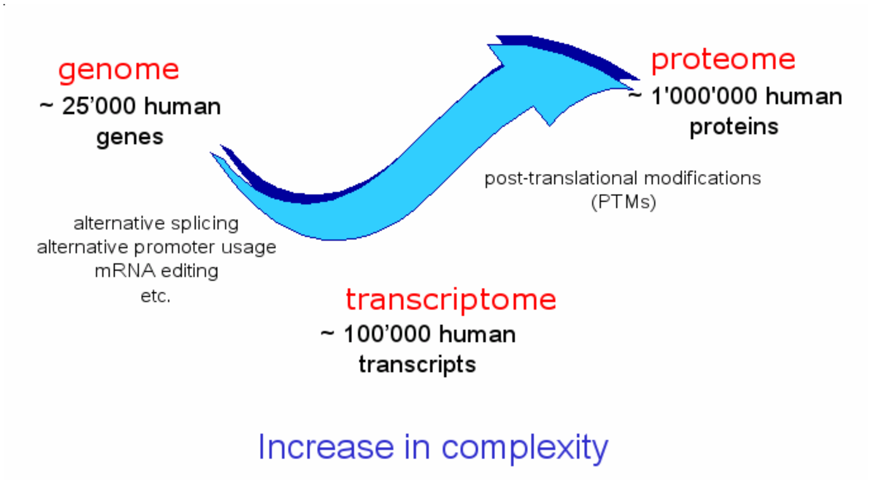
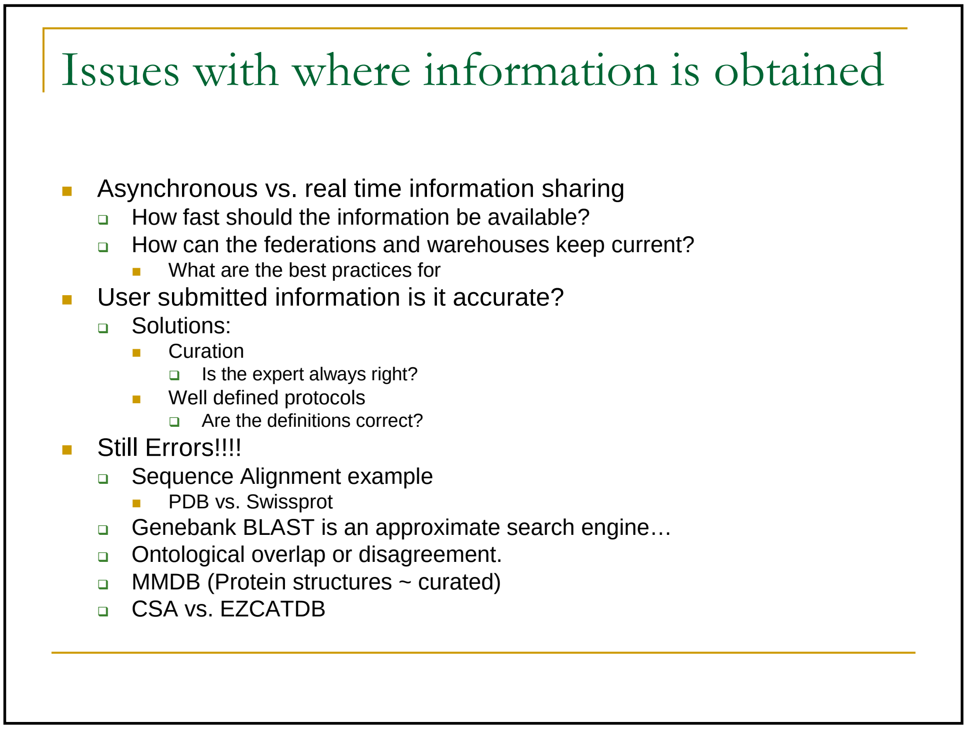

genome
~ 25'000 human genes
alternative splicing
alternative promoter usage
mRNA editing
etc.
transcriptome
~ 100'000 human transcripts
post-translational modifications (PTMs)
proteome
~ 1'000'000 human proteins
Increase in complexity

Image from: http://www.expasy.org/sprot/hpi/hpi_desc.html

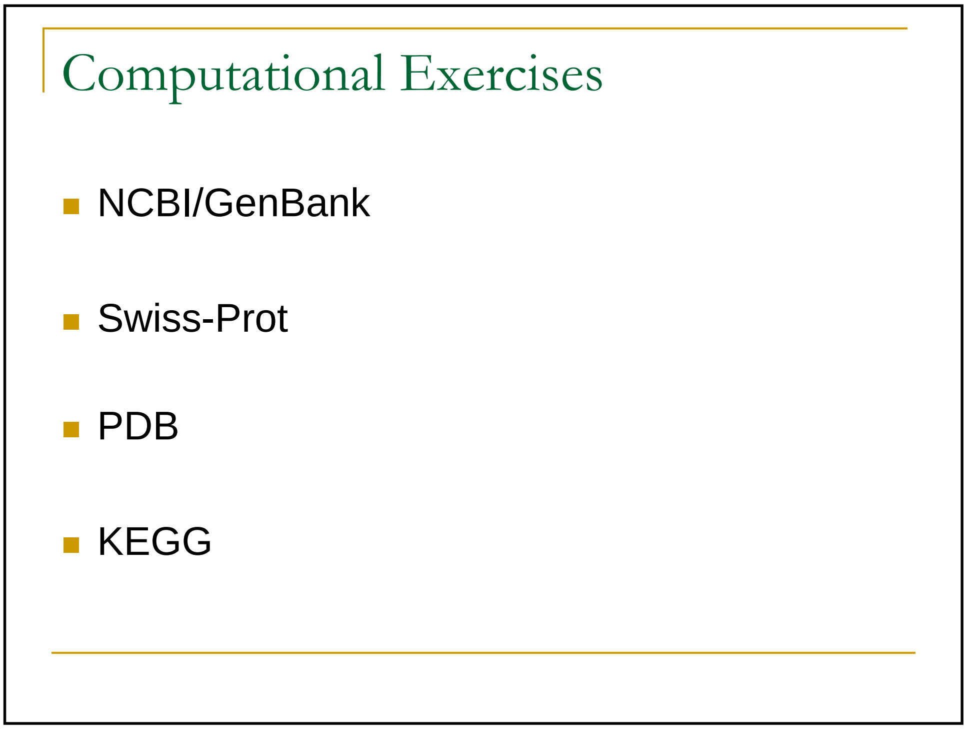
PTM

- Post Translational Modification
- Large-scale studies on chromosomes 21 and 22 indicate that over 80% of the genes could undergo alternative splicing.
- Genomic information does not suffice to predict all the PTMs of which the majority of proteins are the target. Once synthesized on the ribosomes, proteins are subject to a multitude of PTMs. They are cleaved (thus eliminating signal sequences, transit or pro-peptides and initiator methionines); many simple chemical groups can be attached to them (acetyl, methyl, phosphoryl, etc.), as well as a number of more complex molecules, such as sugars and lipids; and finally, proteins can be internally or externally cross-linked (e.g. disulfide bonds). More than two hundred different types of PTM are currently known and many more are yet to be discovered.



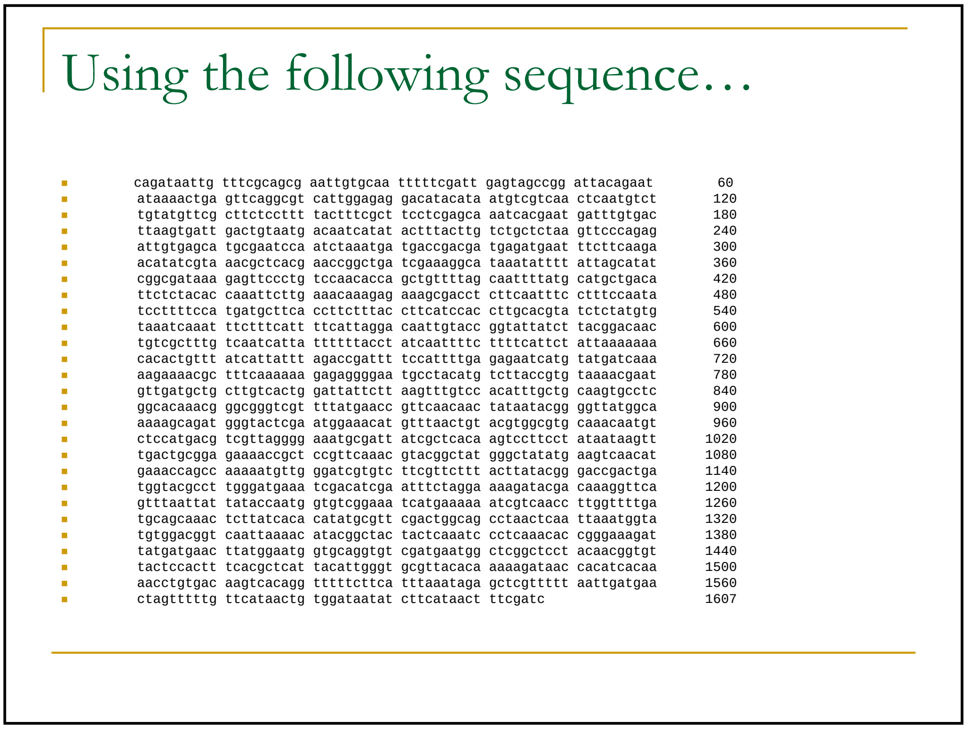
Issues with where information is obtained

- Asynchronous vs. real time information sharing
  - How fast should the information be available?
  - How can the federations and warehouses keep current?
    - What are the best practices for
- User submitted information is it accurate?
  - Solutions:
    - Curation
      - Is the expert always right?
    - Well defined protocols
      - Are the definitions correct?
- Still Errors!!!!
  - Sequence Alignment example
    - PDB vs. Swissprot
  - Genebank BLAST is an approximate search engine…
  - Ontological overlap or disagreement.
  - MMDB (Protein structures ~ curated)
  - CSA vs. EZCATDB



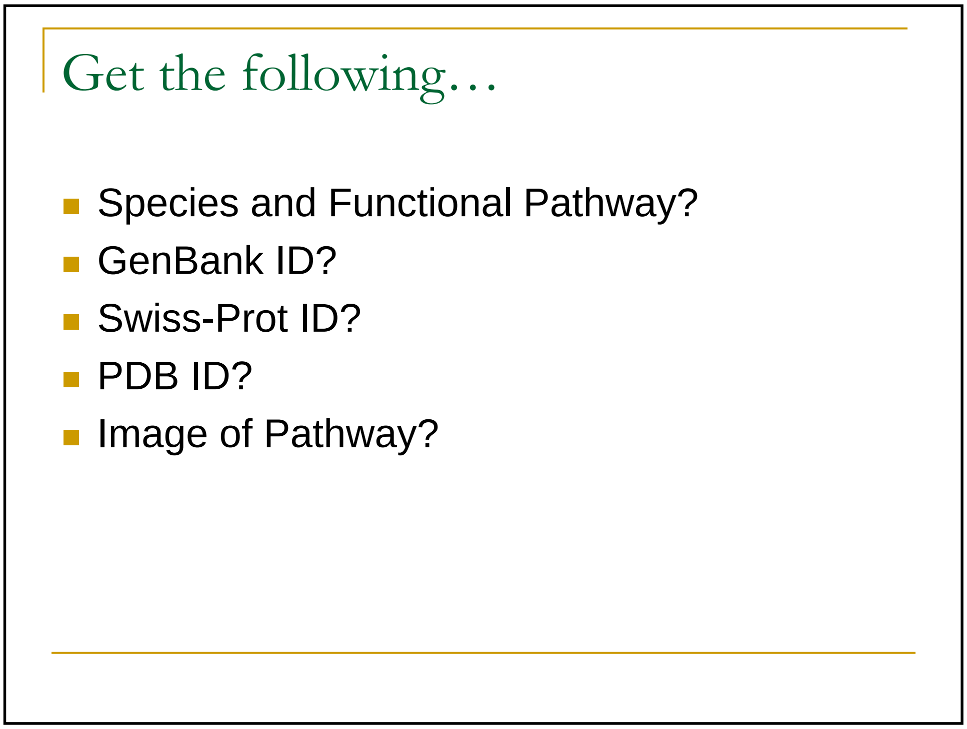
Computational Exercises

- NCBI/GenBank
- Swiss-Prot
- PDB
- KEGG



Using the following sequence…

```
cagataattg tttcgcagcg aattgtgcaa tttttcgatt gagtagccgg attacagaat     60
ataaaactga gttcaggcgt cattggagag gacatacata atgtcgtcaa ctcaatgtct    120
tgtatgttcg cttctccttt tactttcgct tcctcgagca aatcacgaat gatttgtgac    180
ttaagtgatt gactgtaatg acaatcatat actttacttg tctgctctaa gttcccagag    240
attgtgagca tgcgaatcca atctaaatga tgaccgacga tgagatgaat ttcttcaaga    300
acatatcgta aacgctcacg aaccggctga tcgaaaggca taaatatttt attagcatat    360
cggcgataaa gagttccctg tccaacacca gctgttttag caattttatg catgctgaca    420
ttctctacac caaattcttg aaacaaagag aaagcgacct cttcaatttc ctttccaata    480
tccttttcca tgatgcttca ccttctttac cttcatccac cttgcacgta tctctatgtg    540
taaatcaaat ttctttcatt ttcattagga caattgtacc ggtattatct tacggacaac    600
tgtcgctttg tcaatcatta ttttttacct atcaattttc ttttcattct attaaaaaaa    660
cacactgttt atcattattt agaccgattt tccattttga gagaatcatg tatgatcaaa    720
aagaaaacgc tttcaaaaaa gagaggggaa tgcctacatg tcttaccgtg taaaacgaat    780
gttgatgctg cttgtcactg gattattctt aagtttgtcc acatttgctg caagtgcctc    840
ggcacaaacg ggcgggtcgt tttatgaacc gttcaacaac tataatacgg ggttatggca    900
aaaagcagat gggtactcga atggaaacat gtttaactgt acgtggcgtg caaacaatgt    960
ctccatgacg tcgttagggg aaatgcgatt atcgctcaca agtccttcct ataataagtt   1020
tgactgcgga gaaaaccgct ccgttcaaac gtacggctat gggctatatg aagtcaacat   1080
gaaaccagcc aaaaatgttg ggatcgtgtc ttcgttcttt acttatacgg gaccgactga   1140
tggtacgcct tgggatgaaa tcgacatcga atttctagga aaagatacga caaaggttca   1200
gtttaattat tataccaatg gtgtcggaaa tcatgaaaaa atcgtcaacc ttggttttga   1260
tgcagcaaac tcttatcaca catatgcgtt cgactggcag cctaactcaa ttaaatggta   1320
tgtggacggt caattaaaac atacggctac tactcaaatc cctcaaacac cgggaaagat   1380
tatgatgaac ttatggaatg gtgcaggtgt cgatgaatgg ctcggctcct acaacggtgt   1440
tactccactt tcacgctcat tacattgggt gcgttacaca aaaagataac cacatcacaa   1500
aacctgtgac aagtcacagg tttttcttca tttaaataga gctcgttttt aattgatgaa   1560
ctagttttg ttcataactg tggataatat cttcataact ttcgatc                  1607
```



Get the following…

- Species and Functional Pathway?
- GenBank ID?
- Swiss-Prot ID?
- PDB ID?
- Image of Pathway?