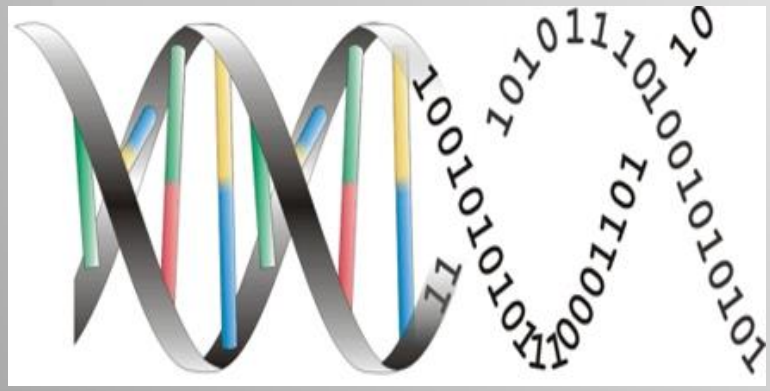
The University of Georgia

The University of Georgia

Programming for Computational & Systems Biology

Instructor: Paul Xie

Tue. & Thr. 9:35~10:50



Course Format

The University of Georgia

- 1 hour lecture + 2 hour computer lab per week
- 2 topics/week; 1 computational biology + 1 coding concept
- 1 assignment/week, please check eLC
- 1 term paper (50%)

The University of Georgia

Databases & Data Retrieval

Instructor: Paul Xie (6)

Last Weeks

The University of Georgia

- Protein alignment
  - Translation
  - PAM
  - BLOSUM
- Coding
  - List
  - Dictionary
  - Function

This Week

The University of Georgia

- Biological Databases
- Data retrieval
- How to obtain the data you need in your research project

Introduction

The University of Georgia

- Fast increase in biological information
- Biological science has now turned into a data rich science
- Gene sequences
- Amino acid sequences in proteins
- Motifs and domains in proteins
- Structural data from XRD & NMR
- Metabolic pathways
- Protein-protein interactions
- Gene expression data DNA microarrays

WHY

The University of Georgia

- What is a database and what are the features of an ideal database?
- What are the relationships/differences between primary and secondary (derived) sequence databases?
- What are the benefits of RefSeq?
- Why is data integration useful?

ADVANTAGES OF DATA INTEGRATION

The University of Georgia

- More relevant *inter-related* information in one place
- Makes it easier to find additional relevant information related to your initial query
- Potentially find information *indirectly* linked, but *relevant* to your subject of interest
  - uncover *non-obvious* genetic features that explain phenotype or disease
- Easier to build a ‘story’ based on *multiple* pieces of biological evidence

WHAT

The University of Georgia

- **Structured** collection of information.
- Consists of basic units called records or entries.
- Each record consists of fields, which hold **pre-defined** data related to the record.
- For example, a protein database would have protein entries as records and protein properties as fields (e.g., name of protein, length, amino-acid sequence)

Biological databases

The University of Georgia

- Biological database is a collection of data which is structured, searchable, updated periodically and also cross-referenced.
- Some databases are multi functional
- Major purposes of databases is as follows:

Availability of biological data

Systemization of data

Analysis of computed biological data

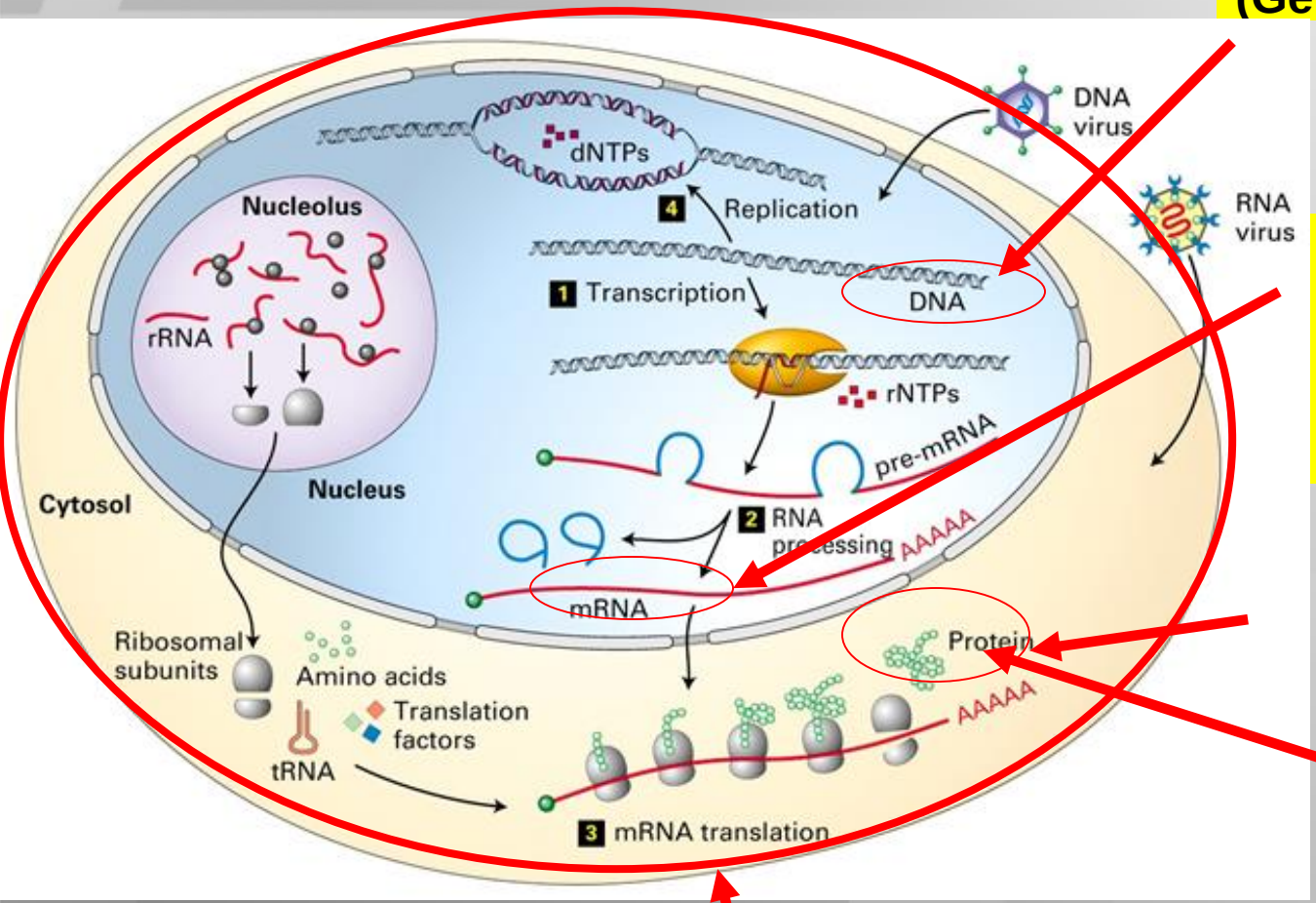
Data type

The University of Georgia

- Genome database
- Sequence database
- Structure database
- Microarray database
- Chemical database
- Pathway database
- Enzyme database
- Disease database
- Literature database

The Central Dogma & Biological Data

The University of Georgia



Original DNA Sequences (Genomes)

Expressed DNA sequenc
(= mRNA Sequences
= cDNA sequences)
Expressed Sequence Tag
(ESTs)

Protein Sequences
-Inferred
-Direct sequencing

Protein structures
-Experiments
-Models (homologues)

Literature information

History

The University of Georgia

- 1956; first sequence database when insulin was sequenced
- 51 amino acids
- Atlas of protein sequences and structures in 1965 by Margaret Day Hoff et al was a printed book.
- Became base for PIR protein information resource
- First nucleotide sequence: yeast tRNA
- 77 bases
- During this time 3D structure of proteins was being studied and renowned PDB was made.

History (2)

The University of Georgia

- First genome published was of free living virus *haemophilus influenzae* in 1995
- Genome?
- All genes ? Or all DNA?
- Why are complete genome interesting?

Biological Databases

The University of Georgia

- Primary Databases
- Secondary Databases

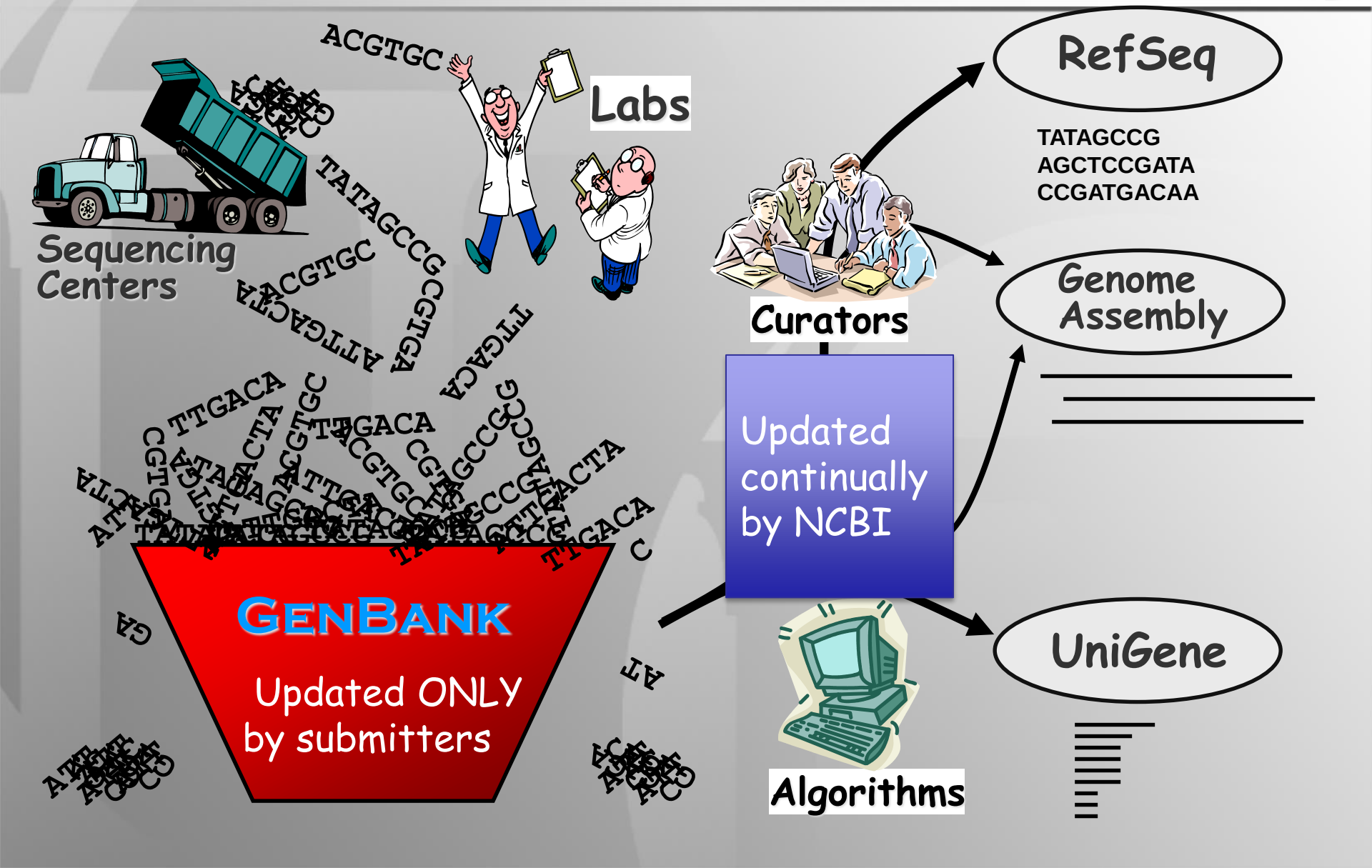
Primary and Secondary Databases

The University of Georgia

- Primary (archival)
  - GenBank/EMBL/DDBJ
  - UniProt (EBI:European Bioinformatics Institute SIB:Swiss Institute of Bioinformatics, PIR:Protein Information Resource)
  - PDB
  - Medline (PubMed)
  - BIND
- Secondary (curated)
  - RefSeq
  - Taxon
  - UniProtKB
  - OMIM (online Mendelian Inheritance in Man)

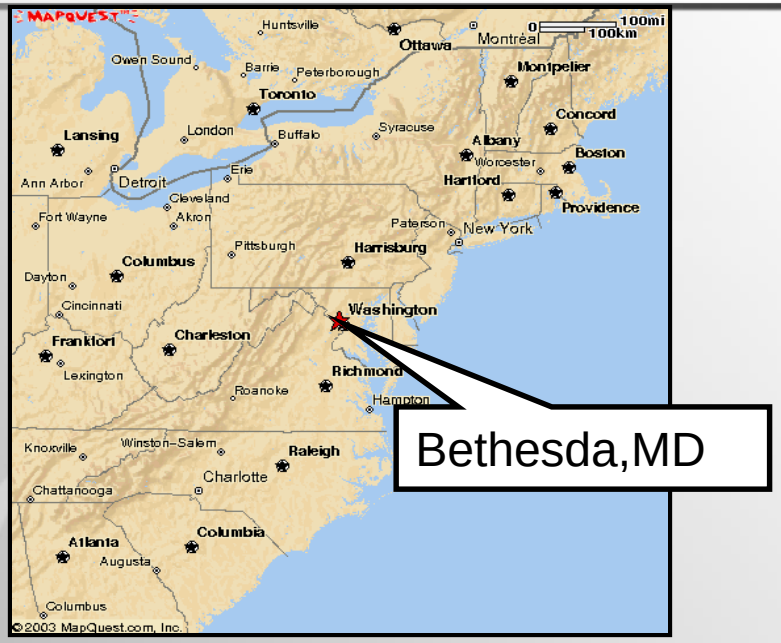
PRIMARY VS. DERIVATIVE SEQUENCE DATABASES

The University of Georgia



The National Center for Biotechnology Information

The University of Georgia



Created in 1988 as a part of the National Library of Medicine at NIH

- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information

Web Access: www.ncbi.nlm.nih.gov

The University of Georgia

NCBI Databases and Services

The University of Georgia

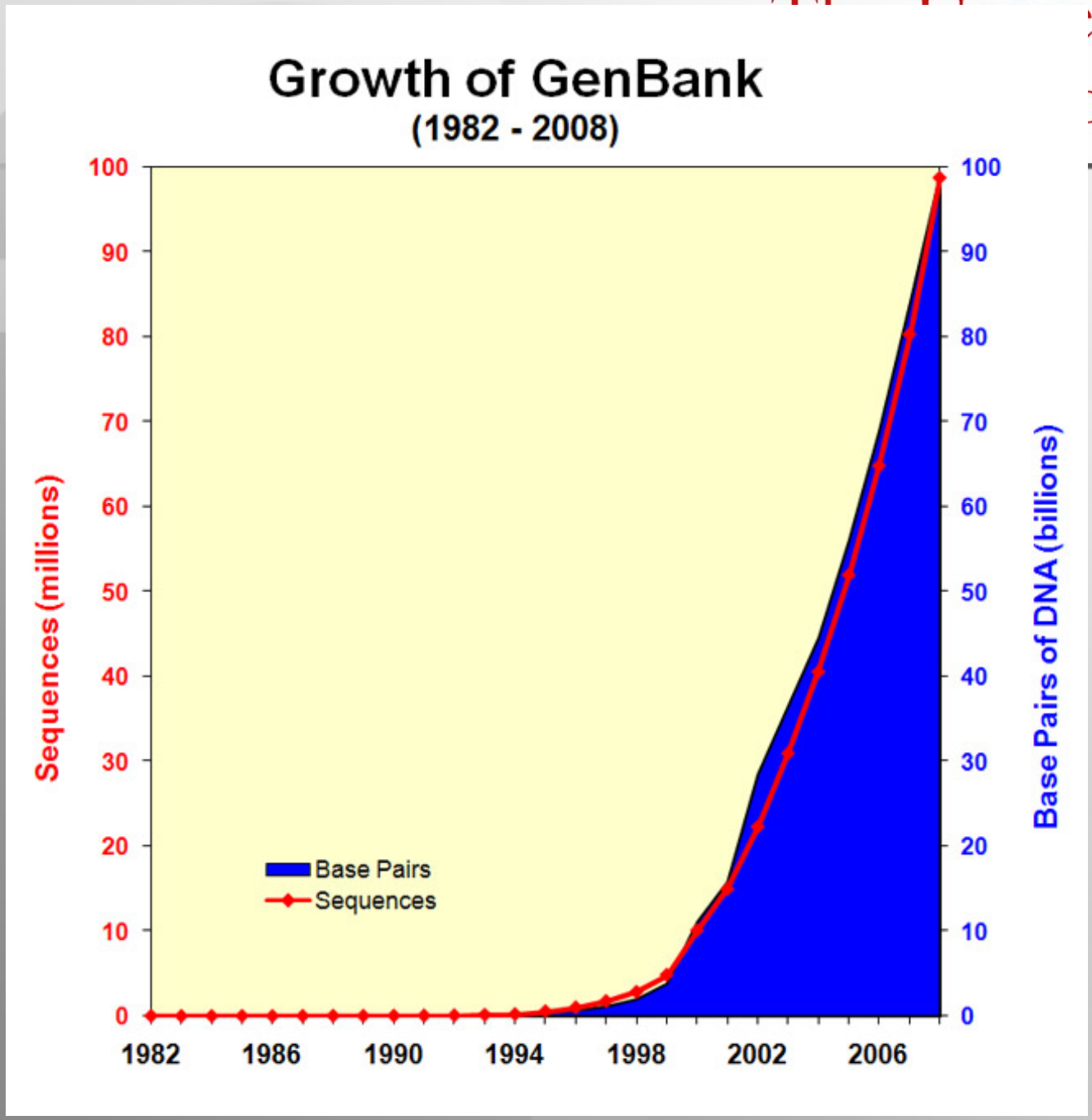
- GenBank primary sequence database
- Free public access to biomedical literature
  - PubMed free Medline (3 million searches per day)
  - PubMed Central full text online access
- Entrez integrated molecular and literature databases

GENBANK - PRIMARY SEQUENCE DB

- **Nucleotide only** sequence database
- **Archival** in nature
  - Historical
  - Reflective of submitter point of view (subjective)
  - Redundant
- Data
  - Direct submissions (traditional records)
  - Batch submissions
  - FTP accounts (genome data)

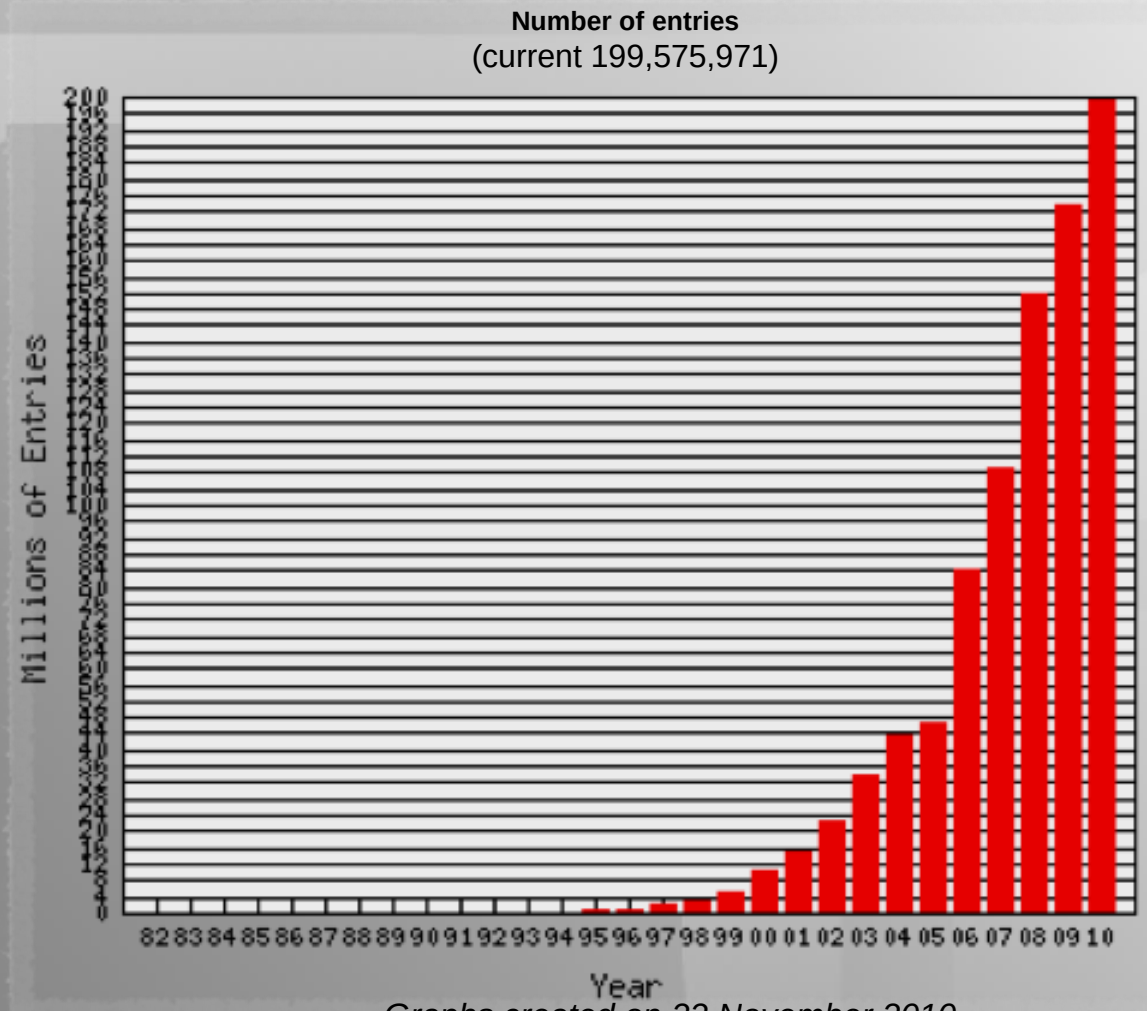
GENBANK - PRIMARY SEQUENCE DB (2)

- Three collaborating databases
  1. GenBank
  2. DNA Database of Japan (DDBJ)
  3. European Molecular Biology Laboratory (EMBL) Database


Growth of GenBank
(1982 - 2008)
Sequences (millions)
Base Pairs of DNA (billions)
0
10
20
30
40
50
60
70
80
90
100
1982
1986
1990
1994
1998
2002
2006
Base Pairs
Sequences

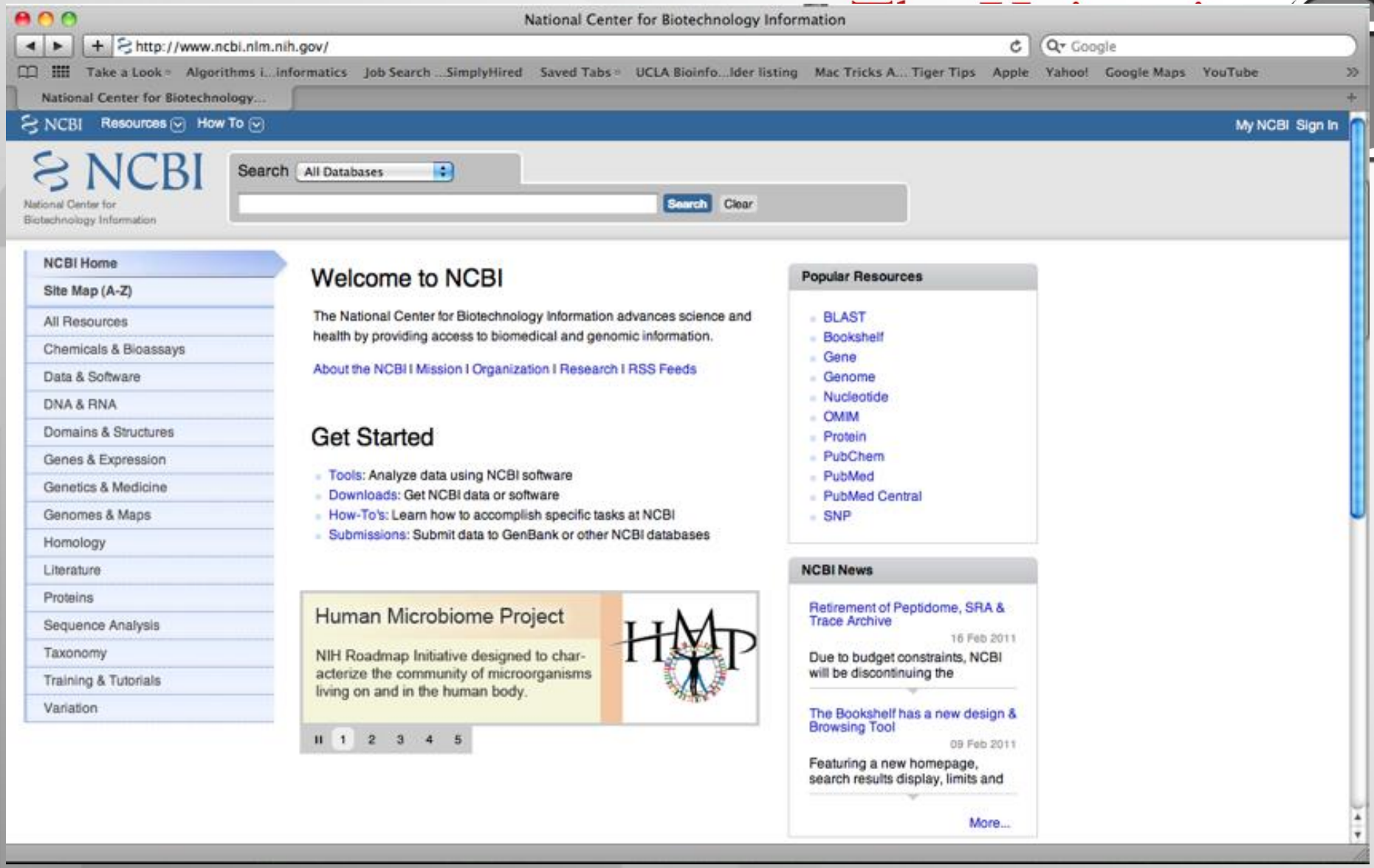
EMBL Database

The University of Georgia

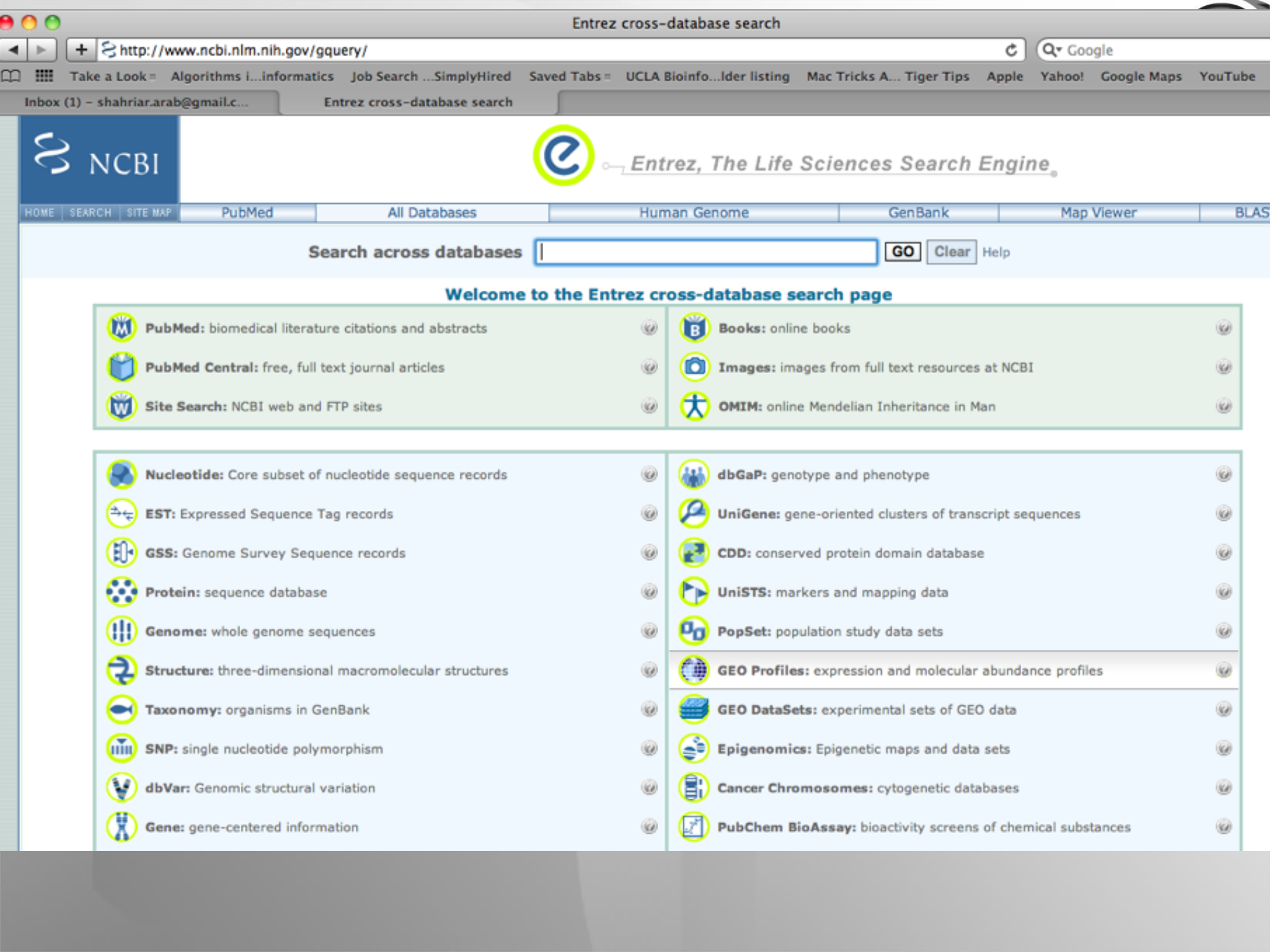


Graphs created on 22 November 2010

http://www.ebi.ac.uk/embl/Services/DBStats/



www.ncbi.nlm.nih.gov


Entrez cross-database search
http://www.ncbi.nlm.nih.gov/gquery/
Google
Take a Look
Algorithms i...informatics
Job Search ...SimplyHired
Saved Tabs
UCLA Bioinfo...lder listing
Mac Tricks A... Tiger Tips
Apple
Yahoo!
Google Maps
YouTube
Inbox (1) - shahriar.arab@gmail.c...
Entrez cross-database search
NCBI
Entrez, The Life Sciences Search Engine
HOME
SEARCH
SITE MAP
PubMed
All Databases
Human Genome
GenBank
Map Viewer
BLAS
Search across databases
GO
Clear
Help
Welcome to the Entrez cross-database search page
PubMed: biomedical literature citations and abstracts
PubMed Central: free, full text journal articles
Site Search: NCBI web and FTP sites
Books: online books
Images: images from full text resources at NCBI
OMIM: online Mendelian Inheritance in Man
Nucleotide: Core subset of nucleotide sequence records
EST: Expressed Sequence Tag records
GSS: Genome Survey Sequence records
Protein: sequence database
Genome: whole genome sequences
Structure: three-dimensional macromolecular structures
Taxonomy: organisms in GenBank
SNP: single nucleotide polymorphism
dbVar: Genomic structural variation
Gene: gene-centered information
dbGaP: genotype and phenotype
UniGene: gene-oriented clusters of transcript sequences
CDD: conserved protein domain database
UniSTS: markers and mapping data
PopSet: population study data sets
GEO Profiles: expression and molecular abundance profiles
GEO DataSets: experimental sets of GEO data
Epigenomics: Epigenetic maps and data sets
Cancer Chromosomes: cytogenetic databases
PubChem BioAssay: bioactivity screens of chemical substances

Traditional GenBank Record

The University of Georgia

```
LOCUS        HSHMLHI                    2503 bp    mRNA    linear   PRI 31-MAR-1994
DEFINITION   Human DNA misma
ACCESSION    U07418
VERSION      U07418.1  GI:46
KEYWORDS     .
SOURCE       Homo sapiens (h
  ORGANISM   Homo sapiens
             Eukaryota; Meta
             Mammalia; Euthe
REFERENCE    1  (bases 1 to
  A
  T
  J
  M
    PUBMED   8128251
R
                20850, USA
```

ACCESSION

VERSION

Version
Tracks change

```
FEATURES             Location/Qualifiers
     source          1..2503
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chrom
                     /map="
                     /tissu
                     /dev_s
     gene            1..250
                     /gene=
     CDS             42..23
                     /gene=
                     /funct
                     /note=
                     Swiss-
                     /codon
                     /prote
                     /db_xr
                     /trans
                     TSIQVI
                     ALASIS
                     TRRKAI
                     VFGNAV
                     ETVYAA
                     LGSNSS
                     LDAFLQ
                     TTKGTS
                     LQEEIN
                     DFANFG
                     YFSLEI
                     FYSIRR
                     LQLANI
```

```
BASE COUNT      723 a    539 c    599 g    642 t
ORIGIN
        1 gttgaacatc tagacgtttc cttggctctt ctggcgccaa aatgtcgttc gtggcagggg
       61 ttattcggcg gctggacgag acagtggtga accgcatcgc ggcgggggaa gttatccagc
      121 ggccagctaa tgctatcaaa gagatgattg agaactgttt agatgcaaaa tccacaagta
      181 ttcaagtgat tgttaaagag ggaggcctga agttgattca gatccaagac aatggcaccg
      241 ggatcaggaa agaagatctg gatattgtat gtgaaaggtt cactactagt aaactgcagt
      301 cctttgagga tttagccagt atttctacct atggctttcg aggtgaggct ttggccagca
      361 taagccatgt ggctcatgtt actattacaa cgaaaacagc tgatggaaag tgtgcataca
      421 gagcaagtta ctcagatgga aaactgaaag cccctcctaa accatgtgct ggcaatcaag
      481 ggaccccagat cacggtggag gacctttttt acaacatagc cacgaggaga aaagctttaa
      541 aaaatccaag tgaagaatat gggaaaattt tggaagttgt tggcaggtat tcagtacaca
      601 atgcaggcat tagtttctca gttaaaaaac aaggagagac agtagctgat gttaggacac
      661 tacccaatgc ctcaaccgtg gacaatattc gctccgtctt tggaaatgct gttagtcgag
      721 aactgataga aattggatgt gaggataaaa ccctagcctt caaaatgaat ggttacatat
      781 ccaatgcaaa ctactcagtg aagaagtgca tcttcttact cttcatcaac catcgtctgg
      841 tagaatcaac ttccttgaga aaagccatag aaacagtgta tgcagcctat ttgcccaaaa
      901 acacacaccc attcctgtac ctcagtttag aaatcagtcc ccagaatgtg gatgttaatg
      961 tgcaccccac aaagcatgaa gttcacttcc tgcacgagga gagcatcctg gagcgggtgc
     1021 agcagcacat cgagagcaag ctcctgggct ccaattcctc caggatgtac ttcacccaga
     1081 ctttgctacc aggacttgct ggcccctctg gggagatggt taaatccaca acaagtctga
     1141 cctcgtcttc tacttctgga agtagtgata aggtctatgc ccaccagatg gttcgtacag
     1201 attcccggga acagaagctt gatgcatttc tgcagcctct gagcaaaccc ctgtccagtc
     1261 agccccaggc cattgtcaca gaggataaga cagatatttc tagtggcagg gctaggcagc
     1321 aagatgagga gatgcttgaa ctcccagccc ctgctgaagt ggctgccaaa aatcagagct
     1381 tggaggggga tacaacaaag gggacttcag aaatgtcaga gaagagagga cctacttcca
     1441 gcaaccccag aaagagacat cgggaagatt ctgatgtgga aatggtggaa gatgattccc
     1501 gaaaggaaat gactgcagct tgtacccccc ggagaaggat cattaacctc actagtgttt
     1561 tgagtctcca ggaagaaatt aatgagcagg gacatgaggt tctccgggag atgttgcata
     1621 accactcctt cgtgggctgt gtgaatcctc agtgggcctt ggcacagcat caaaccaagt
     1681 tataccttct caacaccacc aagcttagtg aagaactgtt ctaccagata ctcatttatg
     1741 attttgccaa ttttggtgtt ctcaggttat cggagccagc accgctcttt gaccttgcca
     1801 tgcttgcctt agatagtcca gagagtggct ggacagagga agatggtccc aaagaaggac
     1861 ttgctgaata cattgttgag tttctgaaga agaaggctga gatgcttgca gactatttct
     1921 ctttggaaat tgatgaggaa gggaacctga ttggattacc ccttctgatt gacaactatg
     1981 tgcccccttt ggagggactg cctatcttca ttcttcgact agccactgag gtgaattggg
     2041 acgaagaaaa ggaatgtttt gaaagcctca gtaaagaatg cgctatgttc tattccatcc
     2101 ggaagcagta catatctgag gagtcgaccc tctcaggcca gcagagtgaa gtgcctggct
     2161 ccattccaaa ctcctggaag tggactgtgg aacacattgt ctataaagcc ttgcgctcac
     2221 acattctgcc tcctaaacat ttcacagaag atggaaatat cctgcagctt gctaacctgc
     2281 ctgatctata caaagtcttt gagaggtgtt aaatatggtt atttatgcac tgtgggatgt
     2341 gttcttcttt ctctgtattc cgatacaaag tgttgtatca aagtgtgata tacaaagtgt
     2401 accaacataa gtgttggtag cacttaagac ttatacttgc cttctgatag tattccttta
     2461 tacacagtgg attgattata aataaataga tgtgtcttaa cat
//
```

well annotated

the sequence is the data

GenPept: GenBank CDS translations


```
FEATURES             Location/Qualifiers
     source          1..2484
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="3"
                     /map="3p22-p23"
     gene            1..2484
                     /gene="M
     CDS             22..2292
                     /gene="M
                     /note="homolog of S. cerevisiae PMS1 (Swiss-Prot Accession
                     Number P14242), S. cerevisiae MLH1 (GenBank Accession
                     Number U07187), E. coli MUTL (Swiss-Prot Accession Number
                     P23367), Salmonella typhimurium MUTL (Swiss-Prot Accession
                     Number P14161) and Streptococcus pneumoniae (Swiss-Prot
                     Accession Number P14160)"
                     /codon_start=1
                     /product="DNA mismatch repair protein homolog"
                     /protein_id="AAC50285.1"
                     /db_xref="GI:463989"
                     /translation="MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKS
                     TSIQVIVKEGGLKLIQIQDNGTGIRKEDLDIVCERFTTSKLQSFEDLASISTYGFRGE
                     ALASISHVAHVTITTKTADGKCAYRASYSDGKLKAPPKPCAGNQGTQITVEDLFYNIA
                     TRRKALKNPSEEYGKILEVVGRYSVHNAGISFSVKKQGETVADVRTLPNASTVDNIRS
```

```
>gi|463989|gb|AAC50285.1| DNA mismatch repair prote...
MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIQVIV...
EDLDIVCERFTTSKLQSFEDLASISTYGFRGEALASISHVAHVTITTKTAD...
```

REFSEQ: *DERIVATIVE* SEQUENCE DATABASE

The University of Georgia

- **Curated transcripts and proteins**
- **Model transcripts and proteins**
- **Assembled Genomic Regions**
- **C**
  - Human genome
  - microbial
  - organelle

ftp://ftp.ncbi.nih.gov/refseq/release/

Selected RefSeq Accession Numbers

The University of Georgia

mRNAs and Proteins	
NM_123456	Curated mRNA
NP_123456	Curated Protein
NR_123456	Curated non-coding RNA
XM_123456	Predicted mRNA
XP_123456	Predicted Protein
XR_123456	Predicted non-coding RNA
Gene Records	
NG_123456	Reference Genomic Sequence
Chromosome	
NC_123455	Microbial replicons, organelle
AC_123455	Alternate assemblies
Assemblies	
NT_123456	Contig
NW_123456	WGS Supercontig

GenBank to RefSeq

The University of Georgia

1. Human apolipoprotein E (epsilon-4 allele) gene, complete cds
   5,515 bp linear DNA
   M10065.1 GI:178852

2. Human mRNA fragment for apolipoprotein E (apo E)
   528 bp linear mRNA
   X00199.1 GI:28808

3. H.sapiens mRN
   275 bp linear m
   Z70760.1 GI:1263

4. Homo sapiens
   1,023 bp linear
   AK314898.1 GI:16

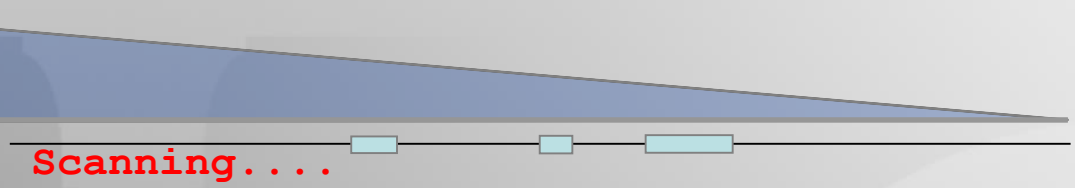
5. Human apolipoprotein E mRNA, complete cds
   1,157 bp linear mRNA
   M12529.1 GI:178848

6. Homo sapiens preapolipoprotein E (APOE) mRNA, complete cds
   1,156 bp linear mRNA
   K00396.1 GI:178850

7. Homo sapiens apolipoprotein E, mRNA (cDNA clone MGC:1571 IMAGE:3355712), complete cds
   1,186 bp linear mRNA
   BC003557.1 GI:13097698

Homo sapiens apolipoprotein E (APOE), mRNA
1,223 bp linear mRNA
NM_000041.2 GI:48762938

RefSeqs: Annotation Reagents



Genomic DNA
(NC, NT, NW)

Model mRNA (XM)
(XR)

Model protein (XP)

= ?

Curated mRNA (NM)
(NR)

Curated Protein (NP)

RefSeq

GenBank
Sequences

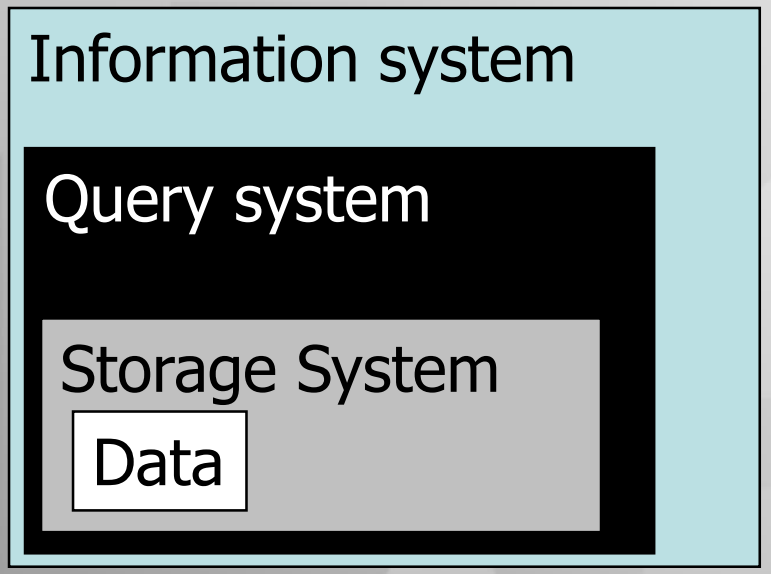
RefSeq Benefits

The University of Georgia

- Non-redundancy
- Updates to reflect current sequence data and *b*
- Data *validation*
- Format *consistency*
- Distinct accession series

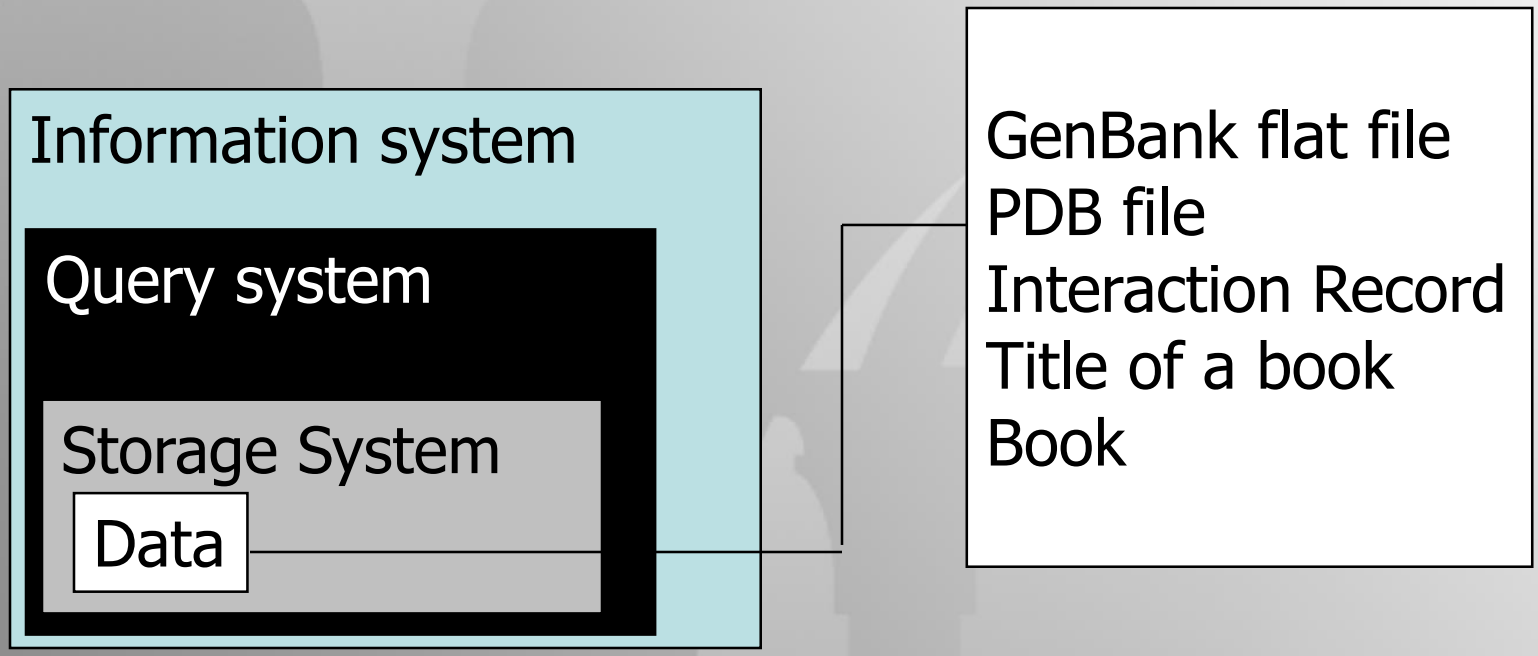
- **PubMed** is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. The United States National Library of Medicine (NLM) at the National Institutes of Health maintains the database as part of the Entrez system of information retrieval.

Databases Architecture



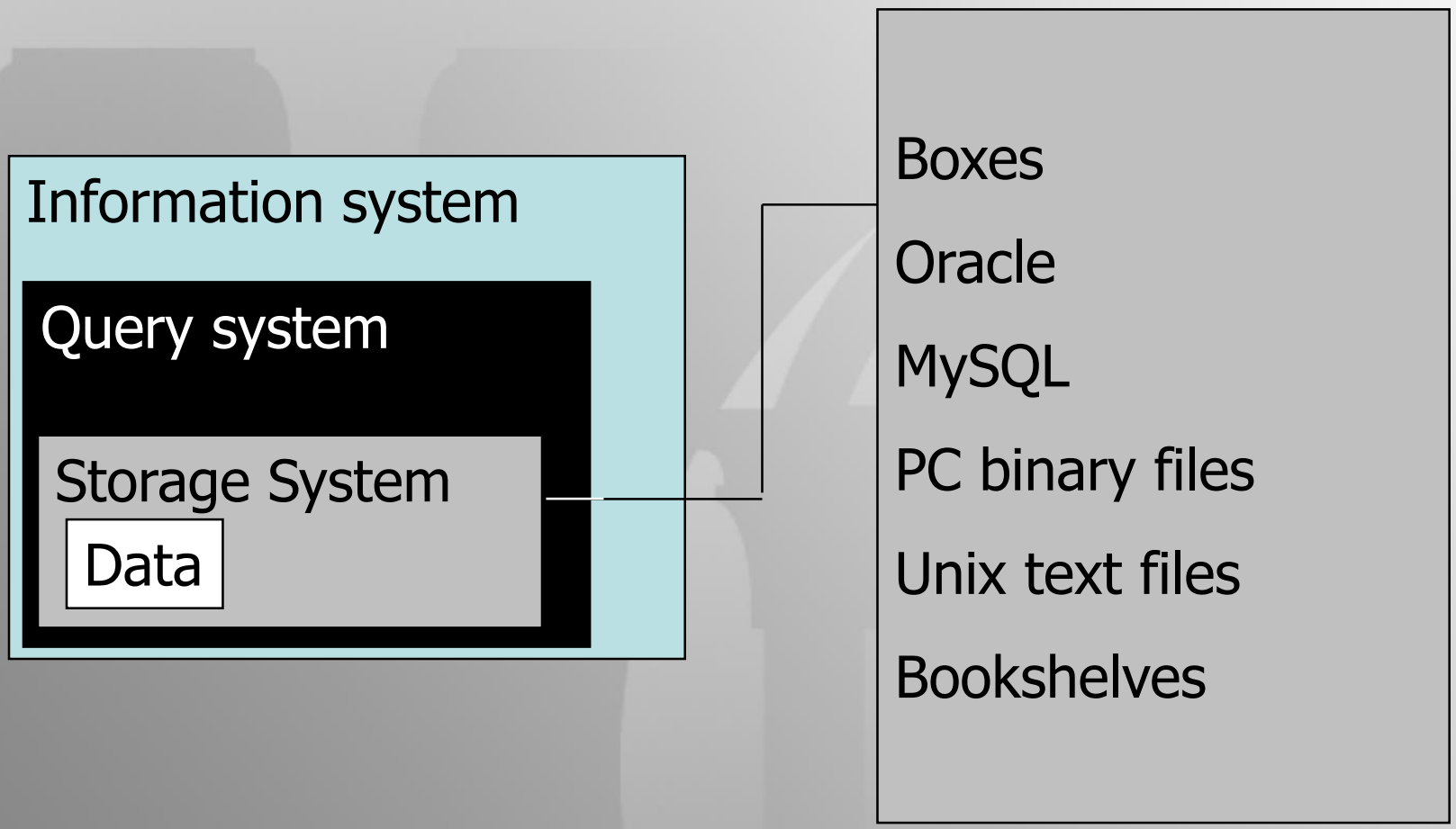
Databases Architecture

The University of Georgia



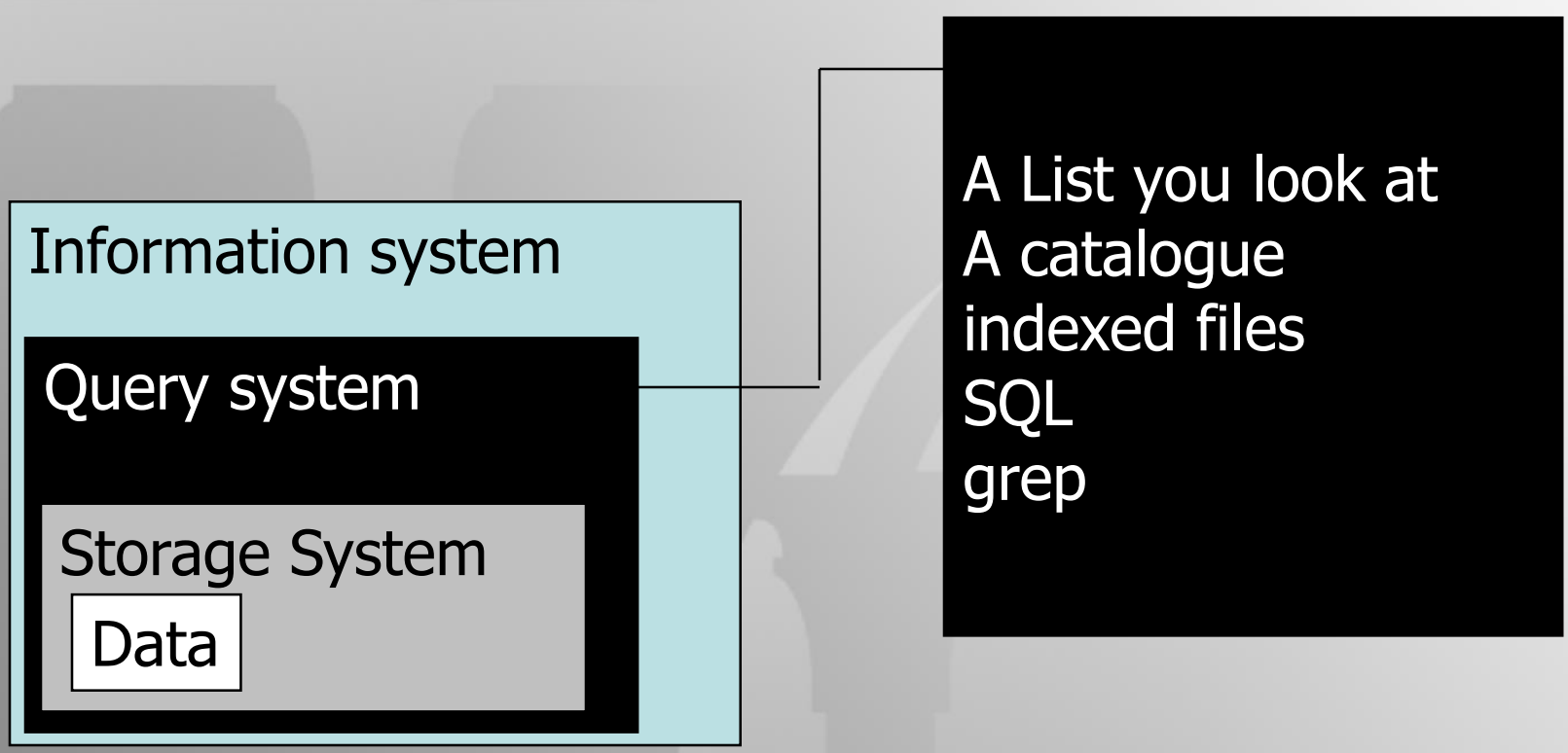
Databases Architecture

The University of Georgia



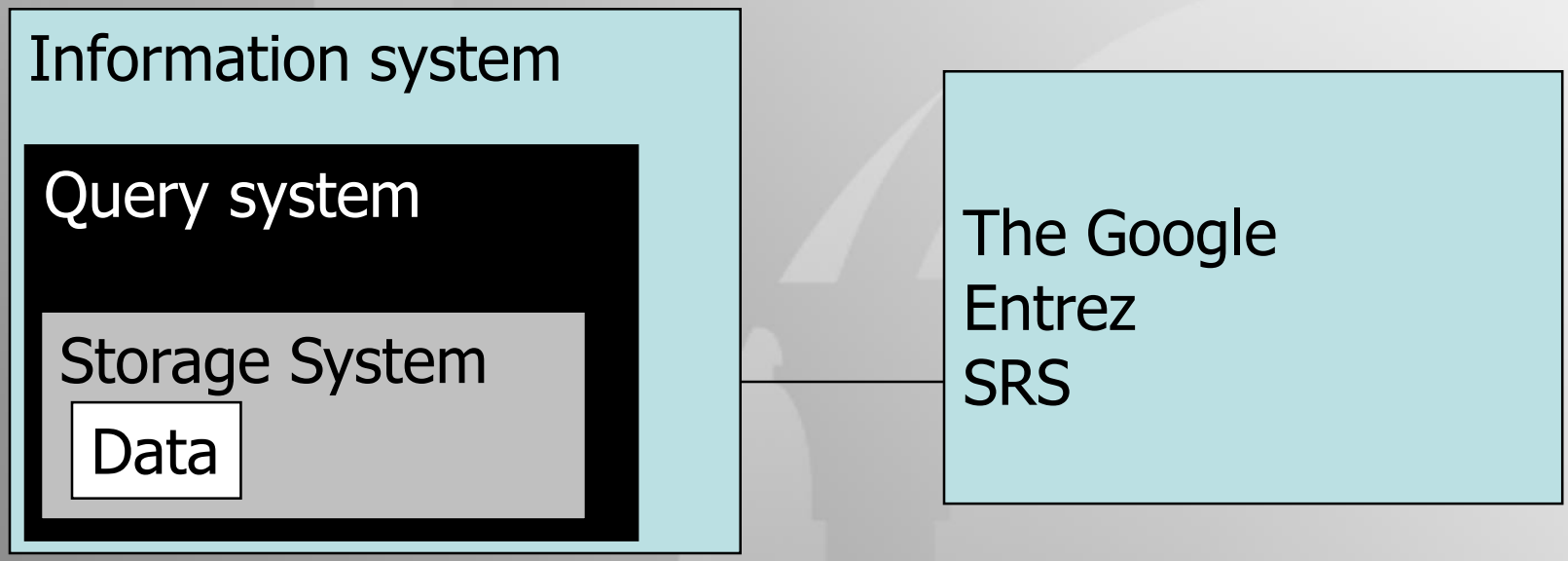
Databases Architecture

The University of Georgia



Databases Architecture

The University of Georgia

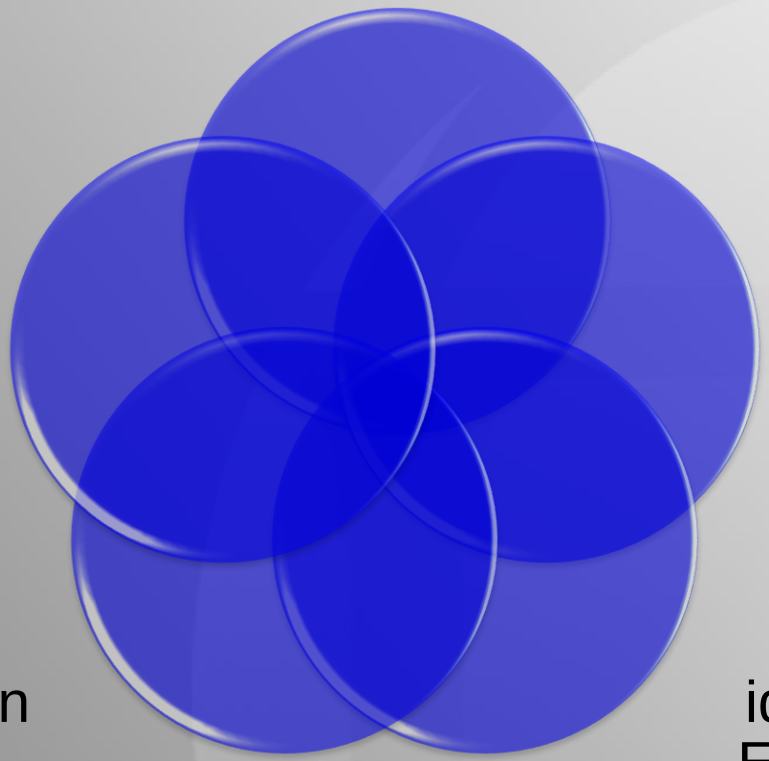


Aspects of genome analysis

The University of Georgia

Ab initio Gene prediction

Gene prediction via comparison, coding and regulatory regions



Locus

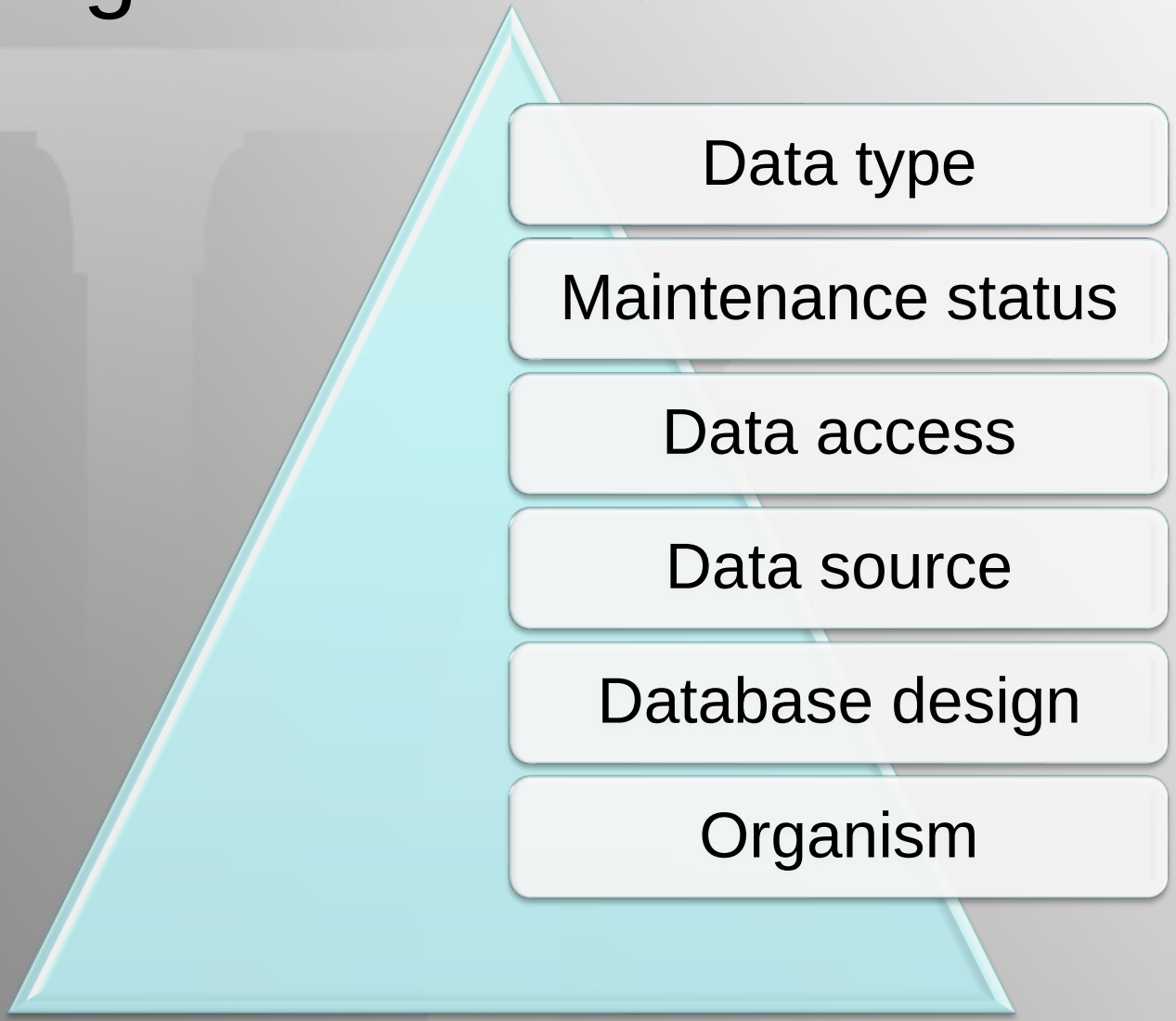
Gene prediction via EST

Gene identification by EST (expressed sequence tags)

Features of biological databases

1) Data heterogeneity
2) High volume data
3) Uncertainty
4) Data Curation
5) Large scale data integration
6) Data sharing
7) Dynamic and subject to change

Classification scheme for biological databases



THE ‘PERFECT’ DATABASE

The University of Georgia

- Comprehensive, but **easy to search**.
- Annotated, but not “too annotated”.
- A simple, easy to understand structure.
- **Cross-referenced.**
- Minimum redundancy.
- **Easy retrieval** of data.

Pitfalls of Biological Databases

The University of Georgia

- Errors in Sequence Databases
- Redundancy in the Primary Sequence Databases
- False or Incomplete Genes Annotations

Errors in Nucleotide Sequences

- sequencing errors
- frame-shifts
- Contaminated with sequences from cloning vectors
  - Exceptional Care for sequences produced before the 1990s

Redundancy

The University of Georgia

- repeated submission
  - identical or overlapping sequences by the same or different authors
  - revision of annotations
  - dumping of expressed sequence tags (EST) data
  - poor database management

NAR Articles

The University of Georgia

OXFORD Journals online

Nucleic Acids Research

HOME | HELP | FEEDBACK | SUBSCRIPTIONS | ARCHIVE | SEARCH | TABLE OF CONTENTS

B F Francis Ouelette || Change Password || View/Change User Information || CiteTrack Personal Alerts || Subscription HELP || **Sign Out**

Receive this page by email each issue: [Sign up for eTOCs]

[Cover Caption]

Other Issues:

Contents: Volume 31, Number 1 January 1 2003 [Index by Author]

- Editorial
- Articles

Find articles in this issue containing these words:

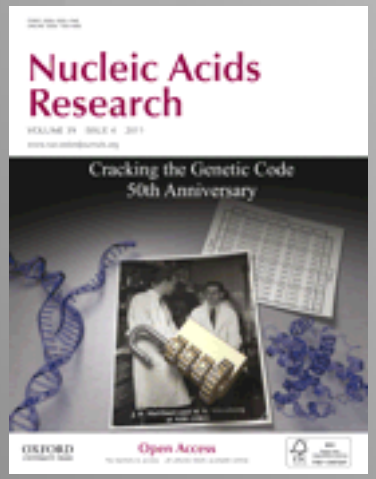
Enter [Search ALL Issues]

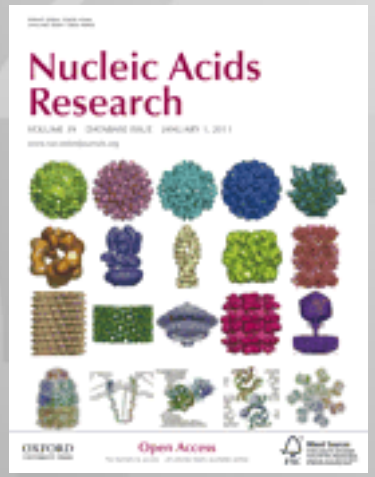
http://nar.oupjournals.org/content/vol31/issue1/

The definitive source….

The University of Georgia

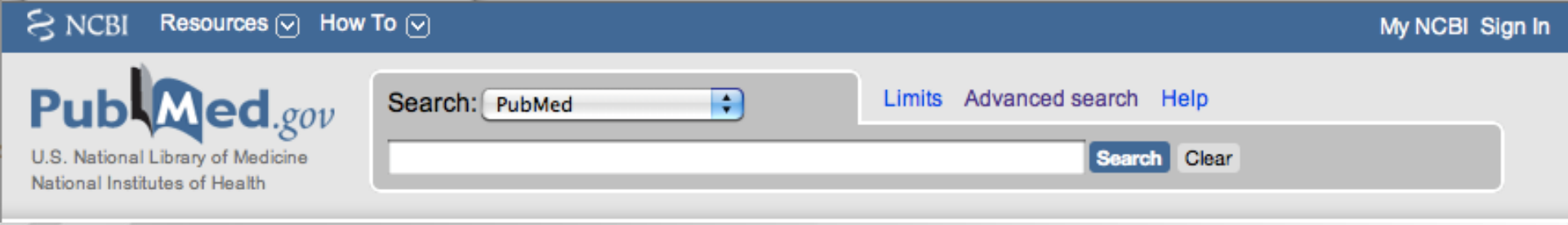
- More than 1300 DB
- http://nar.oxfordjournals.org/content/39/suppl_1.toc





DNA Sequence databases

- Main repositories:
  - GenBank (US)
    - (http://www.ncbi.nlm.nih.gov/Genbank/index.html)
  - EMBL (Europe)
    - (http://www.ebi.ac.uk/embl/)
  - DDBJ (Japan)
    - (http://www.ddbj.nig.ac.jp/)
- Primary databases
  - DNA sequences are identical



PubMed is…

- National Library of Medicine's search service
  - >20 million citations in MEDLINE
  - links to participating online journals
  - PubMed tutorial (via side bar)

Entrez integrates…

- the scientific literature;
- DNA and protein sequence databases;
- 3D protein structure data;
- population study data sets;
- assemblies of complete genomes

Genome Databases

- Focus on one organism or group of organisms:
  - Colibase (E. coli and related species) http://colibase.bham.ac.uk/
  - GDB (human) http://www.gdb.org/
  - Flybase (Drosophila) http://flybase.bio.indiana.edu/
  - WormBase (C. elegans) http://wormbase.org
  - AtDB (Arabidopsis) http://www.arabidopsis.org
  - SGD (S. cerevisiae) http://genome-www.stanford.edu/Saccharomyces/

SWISSPROT

The University of Georgia

http://www.ebi.ac.uk/swissprot/

- European/Swiss Bioinformatics Institute 1986
- Highly accurate, hand curated resource
- Aims:
  - Have a high level of annotation
    - Often by the people who have been working with the gene
  - Have a low level of redundancy
  - Have a high level of integration with other databases

TrEMBL

The University of Georgia

http://www.ebi.ac.uk/trembl/

- SWISSPROT's Big Brother
  - All genes which have been left out of SWISSPROT
  - Computer annotated rather than human annotated

PROSITE

The University of Georgia

http://ca.expasy.org/prosite/

- Families of proteins
- Can search using regular expressions
  - Similar to unix commands using wildcards, etc.
  - E.g., [AC]-x-V-x(4)-{ED}
  - Interpreted as:
  - [Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}
- Families exhibit these patterns
  - So we can search over families
- 1574 documents about 1308 different patterns

PFAM

The University of Georgia

http://pfam.sanger.ac.uk/

- Maintained by the Sanger Centre (Cambridge)
- Protein families aligned using HMMs
  - Hidden Markov Models (see later lecture)
- Given a new sequence
  - Find families which the sequence might fit into
- Sequence Coverage
  - 11912 families
  - Split into Pfam-A (high quality) and Pfam-B (low quality)

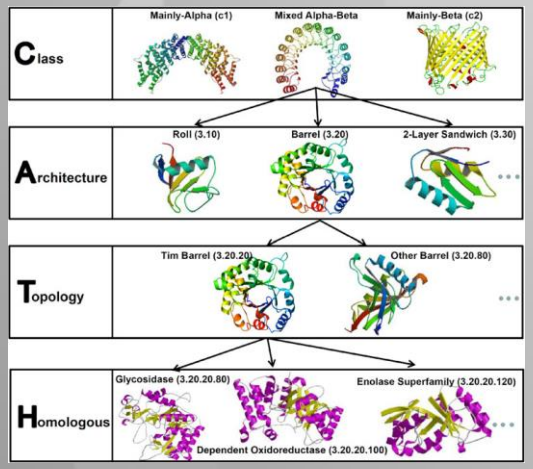
SCOP and CATH

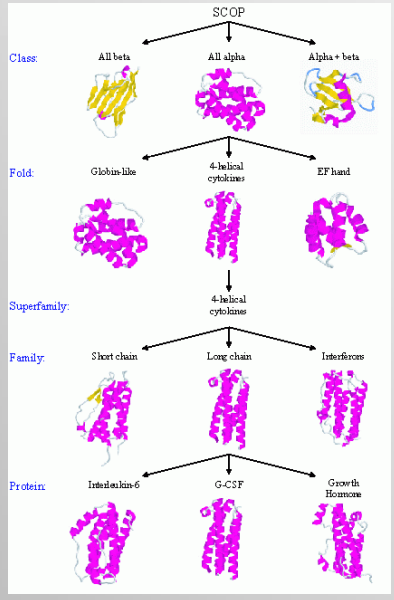
The University of Georgia

http://scop.mrc-lmb.cam.ac.uk/scop/
http://www.cathdb.info/

- SCOP
  - Structural Classification of Proteins
  - Hierarchically ordered and manually curated
  - 38221 PDB Entries
  - 110800 Domains

- CATH
  - Classification of protein domain structures
  - 124 folds
  - 226 Superfamily
  - 1148 Sequence family
  - 14473 Domain



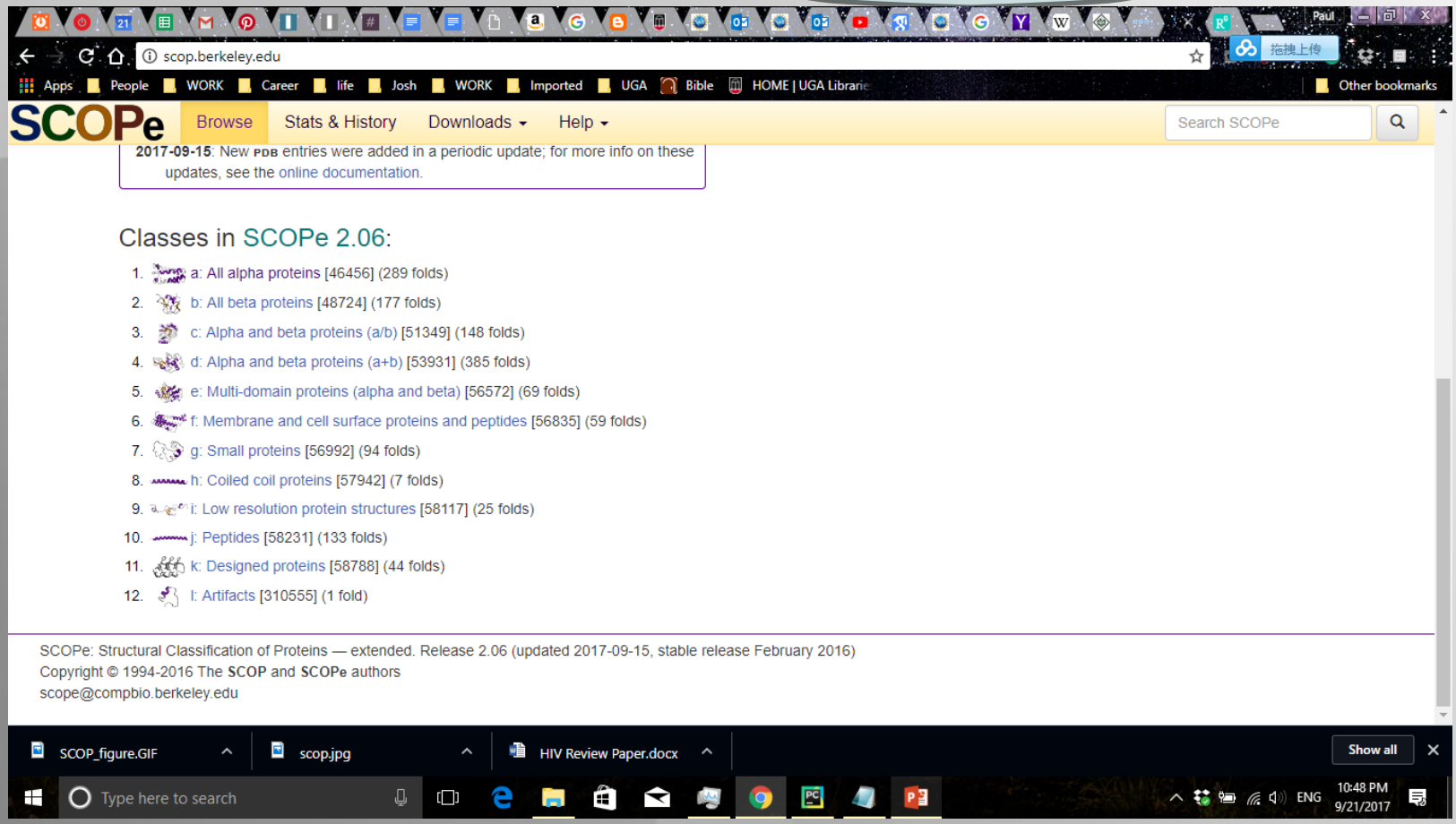


SCOP and CATH

The University of Georgia

http://scop.mrc-lmb.cam.ac.uk/scop/
http://www.cathdb.info/

- SCOP



PDB Format

The University of Georgia

- The PDB format consists of a collection of fixed format records that describe :
  - Atomic coordinates,
  - Chemical and biochemical features
  - Experimental details of the structure determination
  - Some structural features such as
    - Secondary structure assignments,
    - Hydrogen bonding
    - Biological assemblies
    - Active sites

PDB Format

The University of Georgia

```
SHEET    3   L 3 THR H 212  VAL H 214 -1  O  THR H 212   N  HIS H 207
SHEET    1   M 3 THR H 158  TRP H 161  0
SHEET    2   M 3 TYR H 201  HIS H 207 -1  O  ASN H 206   N  THR H 158
SHEET    3   M 3 LYS H 217  VAL H 218 -1  O  VAL H 218   N  TYR H 201
SHEET    1   N 3 TRP C  28  LEU C  29  0
SHEET    2   N 3 VAL C  13  ALA C  18 -1  N  VAL C  17   O  LEU C  29
SHEET    3   N 3 LEU C  36  ALA C  38 -1  O  LEU C  36   N  HIS C  15
SHEET    1   O 5 TRP C  28  LEU C  29  0
SHEET    2   O 5 VAL C  13  ALA C  18 -1  N  VAL C  17   O  LEU C  29
SHEET    3   O 5 TYR C 151  ALA C 156 -1  O  ILE C 154   N  ALA C  14
SHEET    4   O 5 GLY C  54  GLN C  67 -1  N  TYR C  59   O  GLY C 153
SHEET    5   O 5 PRO C 113  LEU C 126 -1  O  GLY C 122   N  ILE C  58
SHEET    1   P 5 GLU C  42  ARG C  44  0
SHEET    2   P 5 GLN C  47  VAL C  49 -1  O  VAL C  49   N  GLU C  42
SHEET    3   P 5 ARG C 131  ILE C 136 -1  O  LEU C 132   N  LEU C  48
SHEET    4   P 5 LEU C  76  ILE C  83 -1  N  THR C  79   O  GLU C 135
SHEET    5   P 5 LYS C  90  LYS C  98 -1  O  LEU C  94   N  ILE C  80
SSBOND   1 CYS L   23    CYS L   88                          1555   1555  2.05
SSBOND   2 CYS L  134    CYS L  194                          1555   1555  2.03
SSBOND   3 CYS H   22    CYS H   98                          1555   1555  2.05
SSBOND   4 CYS H  147    CYS H  203                          1555   1555  2.03
SSBOND   5 CYS C   69    CYS C  101                          1555   1555  2.06
CISPEP   1 SER L    7    PRO L    8          0         2.35
CISPEP   2 TRP L   94    PRO L   95          0         6.74
CISPEP   3 TYR L  140    PRO L  141          0         1.27
CISPEP   4 GLU H    1    VAL H    2          0         3.36
CISPEP   5 GLY H  140    GLY H  141          0         9.68
CISPEP   6 PHE H  153    PRO H  154          0        -3.35
CISPEP   7 GLU H  155    PRO H  156          0         0.91
CISPEP   8 CYS C  101    GLN C  102          0        -3.69
CRYST1  153.663  153.663   99.279  90.00  90.00 120.00 H 3           9
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.006508  0.003757  0.000000        0.00000
SCALE2      0.000000  0.007514  0.000000        0.00000
SCALE3      0.000000  0.000000  0.010073        0.00000
ATOM      1  N   ASP L   1      24.141  -2.203  -3.932  1.00 34.47           N
ATOM      2  CA  ASP L   1      25.160  -2.458  -2.905  1.00 44.54           C
ATOM      3  C   ASP L   1      25.772  -1.169  -2.301  1.00 43.71           C
```

how file was produced

the 17 numbered atoms

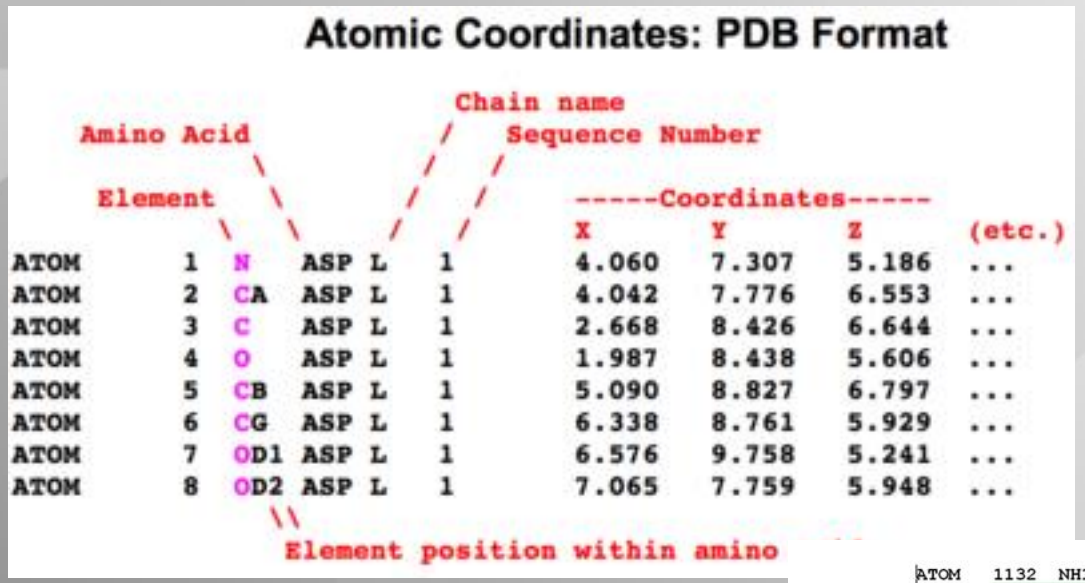
coordinates

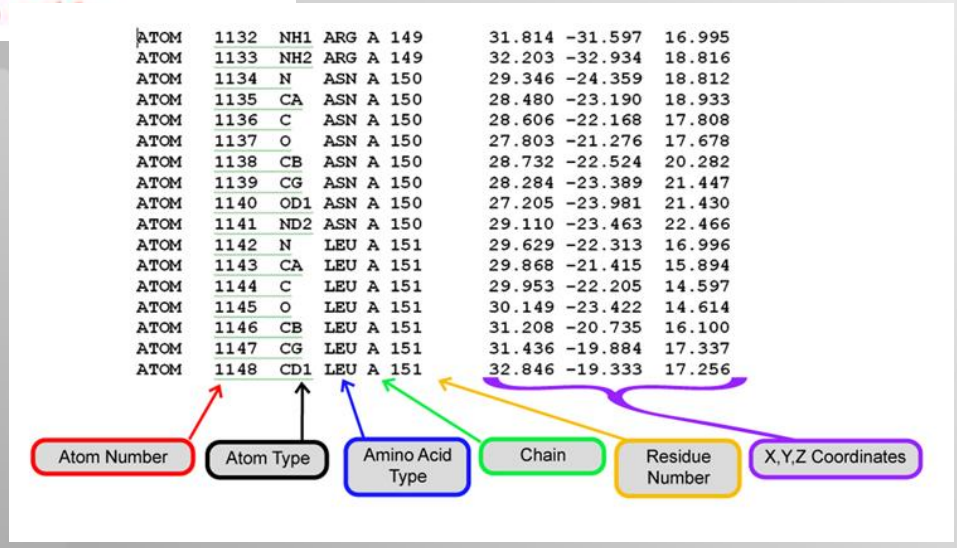
```
HEADER
REMARK Spartan ST exported MOO1      x       y       z
HETATM    1 Cu  UNK  0001        0.000   0.000   0.000
HETATM    2  N  UNK  0001       -0.870   0.000   1.728
HETATM    3  H  UNK  0001       -1.884   0.000   1.654
HETATM    4  H  UNK  0001       -0.607   0.825   2.263
HETATM    5  H  UNK  0001       -0.607  -0.825   2.263
HETATM    6  N  UNK  0001       -1.728   0.000  -0.870
HETATM    7  H  UNK  0001       -1.654   0.000  -1.884
HETATM    8  H  UNK  0001       -2.263   0.825  -0.607
HETATM    9  H  UNK  0001       -2.263  -0.825  -0.607
HETATM   10  N  UNK  0001        0.870   0.000  -1.728
HETATM   11  H  UNK  0001        1.884   0.000  -1.654
HETATM   12  H  UNK  0001        0.607   0.825  -2.263
HETATM   13  H  UNK  0001        0.607  -0.825  -2.263
HETATM   14  N  UNK  0001        1.728   0.000   0.870
HETATM   15  H  UNK  0001        1.654   0.000   1.884
HETATM   16  H  UNK  0001        2.263   0.825   0.607
HETATM   17  H  UNK  0001        2.263  -0.825   0.607
CONECT    1    2    6   10   14
CONECT    2    3    4    5    1
CONECT    3    2
CONECT    4    2
CONECT    5    2
CONECT    6    7    8    9    1
CONECT    7    6
CONECT    8    6
CONECT    9    6
CONECT   10   11   12   13    1
CONECT   11   10
CONECT   12   10
CONECT   13   10
CONECT   14   15   16   17    1
CONECT   15   14
CONECT   16   14
CONECT   17   14
END
```

How the 17 atoms are connected

PDB Format

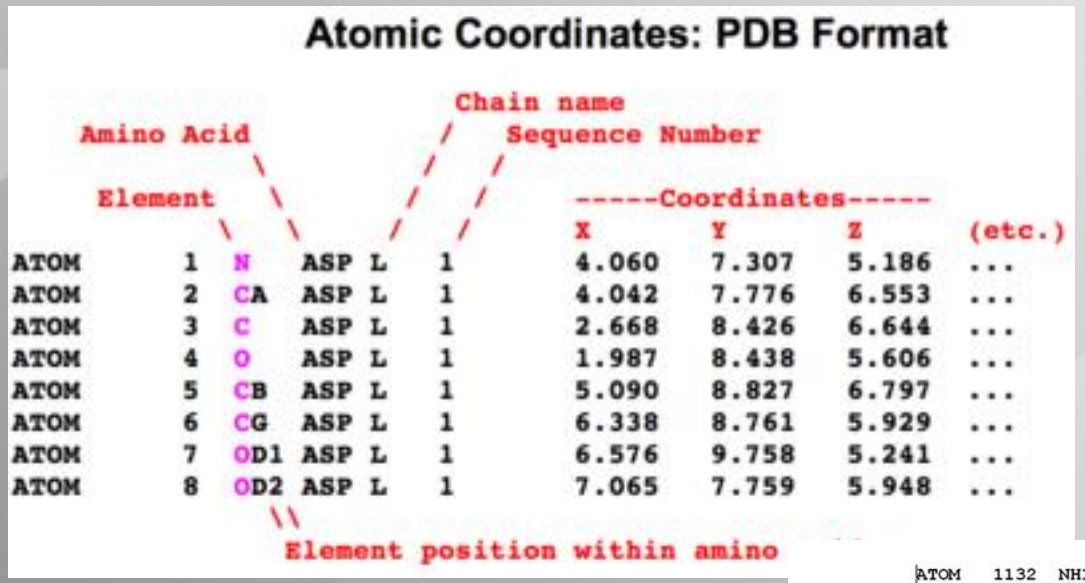
The University of Georgia

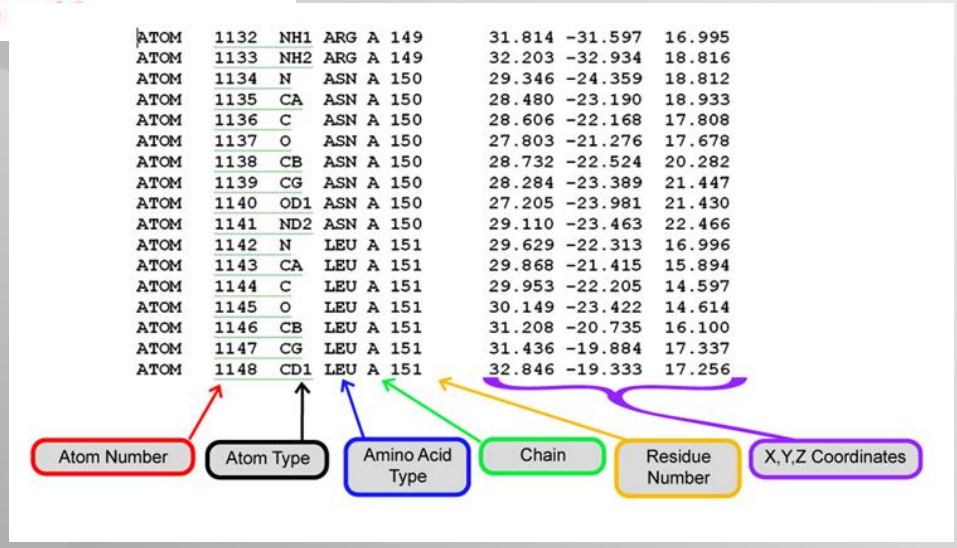




PDB Format

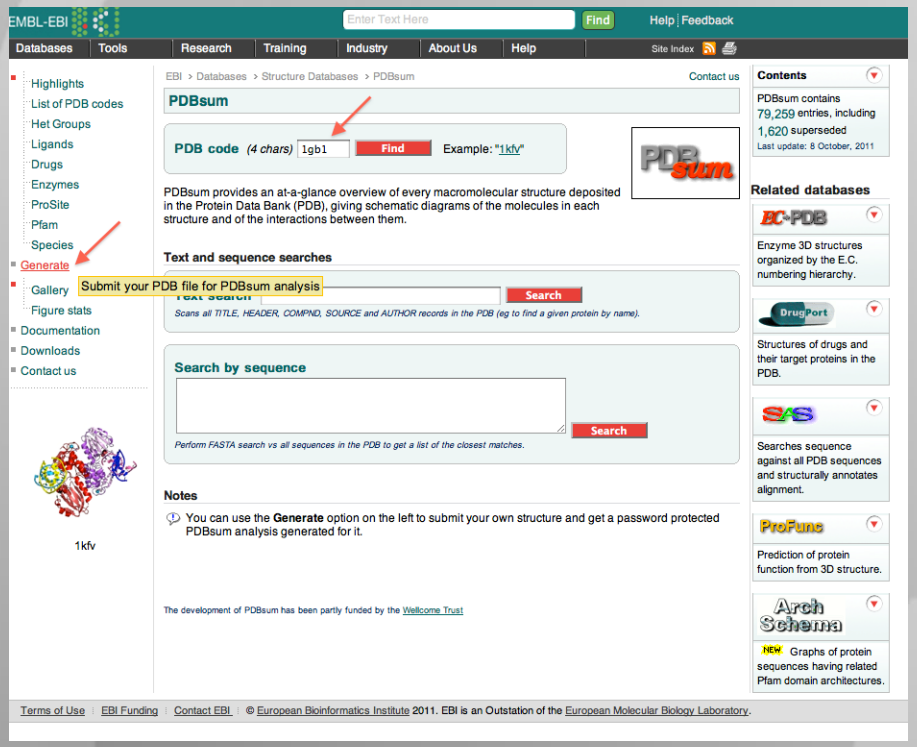
The University of Georgia

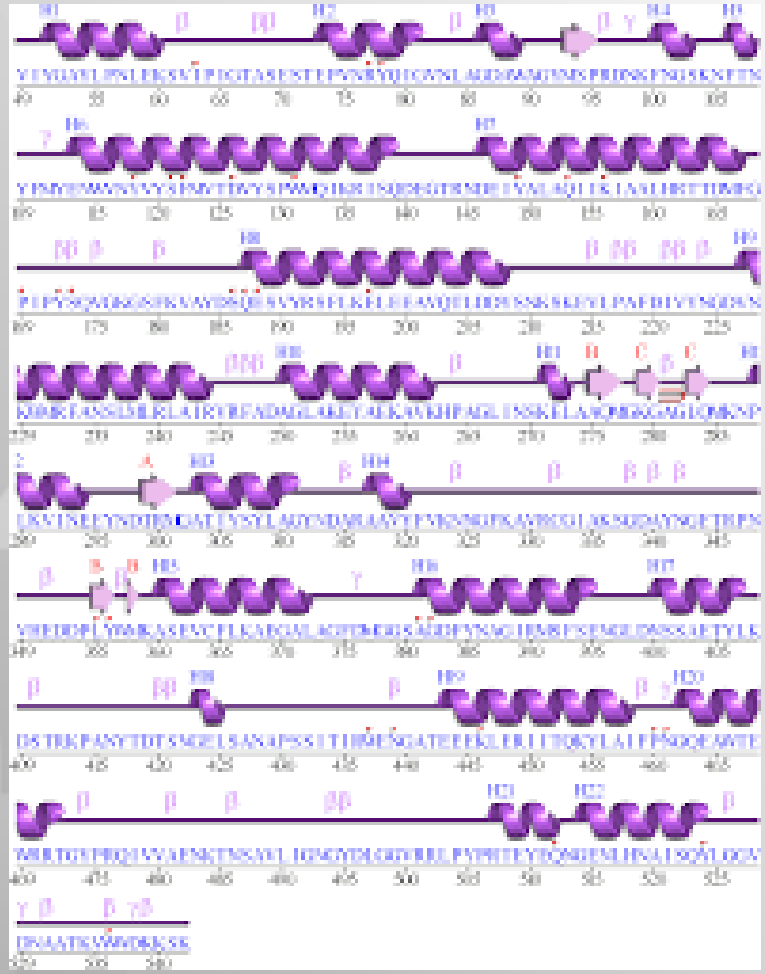




PDBsum

The University of Georgia





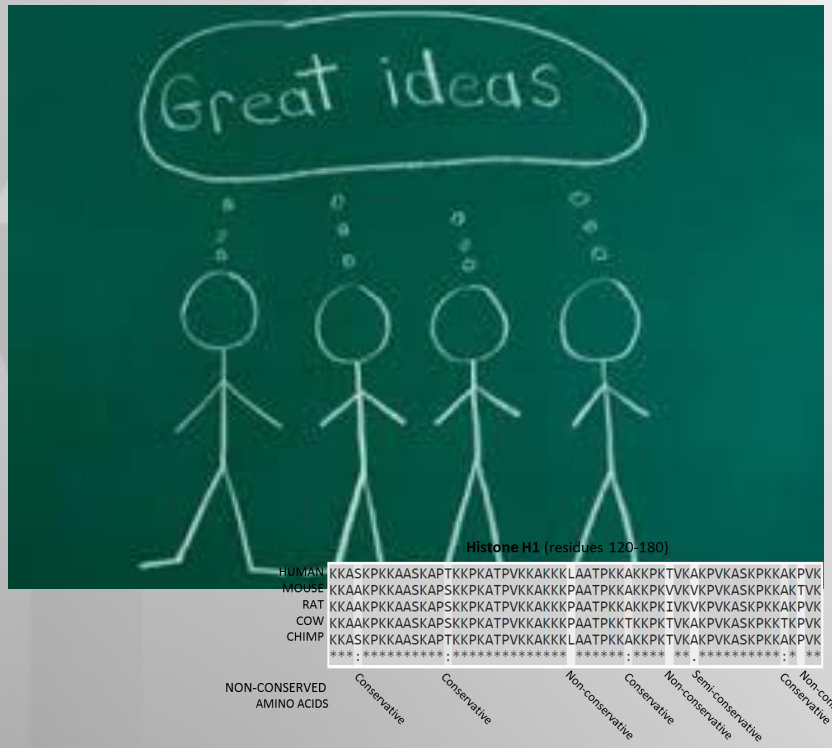
Python Time

The University of Georgia


Thank
You

Group Discussion

The University of Georgia




TIME
FOR A
BREAK

Sequence Databases

The University of Georgia

- Annotated sequence databases
  - SWISS-PROT, GenBank etc…
  - Usage: identifying function, retrieving information
- Low-annotation sequence databases
  - EST databases, high-throughput genome sequences
  - Usage: discovery of new genes

General Protein Databases

The University of Georgia

- SWISS-PROT
  - Manually curated
  - high-quality annotations, less data
- GenPept/TREMBL
  - Translated coding sequences from GenBank/EMBL
  - Few annotations, more up to date
- PIR
  - Phylogenetic-based annotations
- All 3 now combining efforts to form UniProt (http://www.uniprot.org)

Low-annotation Databases

- ESTs (Expressed Sequence Tags)
  - Low quality sequences generated by high - volume sequencing the 3’ or 5’ end of cDNAs
- High-throughput genome sequences
  - Produced by mass-sequencing of genomic DNA

Non-redundant Databases

The University of Georgia

- Sequence data only: cannot be browsed, can only be searched using a sequence
- Combine sequences from more than one database
- Examples:
  - NR Nucleic (genbank+EMBL+DDBJ+PDB DNA)
  - NR Protein (SWISS-PROT+TrEMBL+GenPept+PDB protein)

Sequence & Structure Databases

The University of Georgia

- PDB (Protein Databank)
  - Stores 3-dimensional atomic coordinates for biological molecules including protein and nucleic acids
  - Data obtained by X-ray crystallography, NMR, or computer modelling
  - http://www.rcsb.org/pdb/
- MMDB (Molecular Modelling database)
  - Over 28,000 3D macromolecular structures, including proteins and polynucleotides
  - (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Structure)
- SCOP (Structural Classification of Proteins)
  - Classification of proteins according to structural and evolutionary relationships

File Formats

The University of Georgia

- GenBank/GB, genbank flatfile format
- NBRF format
- EMBL, EMBL flatfile format
- Swissprot
- GCG, single sequence format of GCG software
- DNAStrider, for common Mac program
- Pearson/Fasta, a common format used by Fasta programs and others
- Phylip3.2, sequential format for Phylip programs
- Phylip, interleaved format for Phylip programs (v3.3, v3.4)
- Plain/Raw, sequence data only (no name, document, numbering)
- MSF multi sequence format used by GCG software
- PAUP"s multiple sequence (NEXUS) format
- ASN.1 format used by NCBI

EMBL Format

The University of Georgia

```
ID TRBG361 standard; mRNA; PLN; 1859 BP.
XX
AC X56734; S46826;
XX
SV X56734.1
XX
DT 12-SEP-1991 (Rel. 29, Created)
DT 15-MAR-1999 (Rel. 59, Last updated, Version 9)
XX
DE Trifolium repens mRNA for non-cyanogenic beta-
glucosidase
XX
KW beta-glucosidase.
XX
OS Trifolium repens (white clover)
OC Eukaryota; Viridiplantae; Streptophyta; Embryophyta;
Tracheophyta;
OC Spermatophyta; Magnoliophyta; eudicotyledons; core
eudicots; rosids;
OC eurosids I; Fabales; Fabaceae; Papilionoideae;
Trifolieae; Trifolium.
XX
RN [5]
RP 1-1859
RX MEDLINE; 91322517.
RX PUBMED; 1907511.
RA Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT "Nucleotide and derived amino acid sequence of the
cyanogenic
RT beta-glucosidase (linamarase) from white clover
(Trifolium repens L.).";
RL Plant Mol. Biol. 17(2):209-219(1991).
XX
RN [6]
RP 1-1859
RA Hughes M.A.;
RT ;
RL Submitted (19-NOV-1990) to the EMBL/GenBank/DDBJ
databases.
RL M.A. Hughes, UNIVERSITY OF NEWCASTLE UPON TYNE,
MEDICAL SCHOOL, NEW CASTLE
RL UPON TYNE, NE2 4HH, UK
XX
DR GOA; P26204.
DR MENDEL; 11000; Trirp;1162;11000.
DR SWISS-PROT; P26204; BGLS_TRIRP.
XX
FH Key Location/Qualifiers
FH
FT source 1..1859
FT /db_xref="taxon:3899"
FT /mol_type="mRNA"
FT /organism="Trifolium repens"
FT /tissue_type="leaves"
FT /clone_lib="lambda gt10"
FT /clone="TRE361"
FT CDS 14..1495
FT /db_xref="GOA:P26204"
FT /db_xref="SWISS-PROT:P26204"
FT /note="non-cyanogenic"
FT /EC_number="3.2.1.21"
FT /product="beta-glucosidase"
FT /protein_id="CAA40058.1"
FT /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGFI
FT FGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGIMK
FT DQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWDLPQ
FT VLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTNAPGR
FT CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNWLMPLD
FT DNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIVKNRLPKFSKFESSLVNGSFDF
FT IGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYVYPYMFIQ
FT EDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYRHLYYIRSA
FT IRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
FT mRNA 1..1859
FT /evidence=EXPERIMENTAL
XX
SQ Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt 60
cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag 120
tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga 180
aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata 240
tccagaaaaa ataagggatg gaagcaatgc agacatcacg gttgaccaat atcaccgcta 300
caaggaagat gttgggatta tgaaggatca aaatatggat tcgtatagat tctcaatctc 360
ttggccaaga atactcccaa agggaaagtt gagcggaggc ataaatcacg aaggaa
```

Genbank Format

```
LOCUS       SCU49845     5028 bp    DNA             PLN
21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial
cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina;
Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae;
Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function
is required for
            DNA damage-induced mutagenesis in Saccharomyces
cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  MEDLINE   95176709
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires
Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  MEDLINE   96194260
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale
University, New
            Haven, CT, USA
gene            687..3158
                /gene="AXL2"
CDS             687..3158
                /gene="AXL2"
                /note="plasma membrane glycoprotein"
                /codon_start=1
                /function="required for axial budding pattern of
                cerevisiae"
                /product="Axl2p"
                /protein_id="AAA98666.1"
                /db_xref="GI:1293615"

/translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF

TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN

VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE

VFNVTFDRSMFTNEESIVSYYGRSQLYNAPLPNWLFFDSGELKFTGTAPVINSAIAPE

TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDTGNVSYDLPLNYV

YLDDDPISSDKLGSINLLDAPDWVALDNATISGSVPDELLGKNSNPANFSVSIYDTYG

DVIYFNFEVVSTTDLFAISSLPNINATRGEWFSYYFLPSQFTDYVNTNVSLEFTNSSQ

DHDWVKFQSSNLTLAGEVPKNFDKLSLGLKANQGSQSQELYFNIIGMDSKITHSNHSA

NATSTRSSHHSTSTSSYTSSTYTAKISSTSAAATSSAPAALPAANKTSSHNKKAVAIA

CGVAIPLGVILVALICFLIFWRRRRENPDDENLPHAISGPDLNNPANKPNQENATPLN

NPFDDDASSYDDTSIARRLAALNTLKLDNHSATESDISSVDEKRDSLSGMNTYNDQFQ

SQSKEELLAKPPVQPPESPFFDPQNRSSSVYMDSEPAVNKSWRYTGNLSPVSDIVRDS

YGSQKTVDTEKLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK

HRNRHLQNIQDSQSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL
                VDFSNKSNVNVGQVKDIHGRIPEML"
BASE COUNT     1510 a   1074 c    835 g   1609 t
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac
ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca
gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt
gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa
taataaaccg
```

Swissprot format

Swiss-Prot: Q823P0

NiceProt - a user-friendly view of this Swiss-Prot entry

```
ID   DNA1_CHLCV     STANDARD;      PRT;   450 AA.
AC   Q823P0;
DT   10-OCT-2003 (Rel. 42, Created)
DT   10-OCT-2003 (Rel. 42, Last sequence update)
DT   25-OCT-2004 (Rel. 45, Last annotation update)
DE   Chromosomal replication initiator protein dnaA 1.
GN   Name=dnaA1; Synonyms=dnaA-1; OrderedLocusNames=CCA00368;
OS   Chlamydophila caviae.
OC   Bacteria; Chlamydiae; Chlamydiales; Chlamydiaceae; Chlamydophila.
OX   NCBI_TaxID=83557;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=GPIC;
RX   MEDLINE=22569155; PubMed=12682364 [NCBI, ExPASy, EBI, Israel, Japan]; DOI=10.1093/n
RA   Read T.D., Myers G.S.A., Brunham R.C., Nelson W.C., Paulsen I.T.,
RA   Heidelberg J.F., Holtzapple E.K., Khouri H.M., Federova N.B.,
RA   Carty H.A., Umayam L.A., Haft D.H., Peterson J.D., Beanan M.J.,
RA   White O., Salzberg S.L., Hsia R.-C., McClarty G., Rank R.G.,
RA   Bavoil P.M., Fraser C.M.;
RT   "Genome sequence of Chlamydophila caviae (Chlamydia psittaci GPIC):
RT   examining the role of niche-specific genes in the evolution of the
RT   Chlamydiaceae.";
RL   Nucleic Acids Res. 31:2134-2147(2003).
CC   -!- FUNCTION: Plays an important role in the initiation and regulation
CC       of chromosomal replication. Binds to the origin of replication; it
CC       binds specifically double-stranded DNA at a 9 bp consensus (dnaA
CC       box): 5'-TTATC(C/A)A(C/A)A-3'. DnaA binds to ATP and to acidic
CC       phospholipids (By similarity).
CC   -!- SIMILARITY: Belongs to the dnaA family.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AE016995; AAP05115.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
DR   HSSP; P03004; 1J1V. [HSSP ENTRY / SWISS-3DIMAGE / PDB]
DR   TIGR; CCA00368; -.
DR   HAMAP; MF_00377; -; 1.
DR   InterPro; IPR003593; AAA ATPase.
DR   InterPro; IPR001957; Bac DnaA.
DR   InterPro; IPR010921; Trp repress rep.
DR   InterPro; Graphical view of domain structure.
DR   Pfam; PF00308; Bac DnaA; 1.
DR   Pfam; Graphical view of domain structure.
DR   PRINTS; PR00051; DNAA.
DR   SMART; SM00382; AAA; 1.
DR   TIGRFAMs; TIGR00362; DnaA; 1.
DR   PROSITE; PS01008; DNAA; 1.
DR   ProDom [Domain structure / List of seq. sharing at least 1 domain]
DR   HOBACGEN [Family / Alignment / Tree]
DR   BLOCKS; Q823P0.
DR   ProtoNet; Q823P0.
DR   ProtoMap; Q823P0.
DR   PRESAGE; Q823P0.
DR   DIP; Q823P0.
DR   ModBase; Q823P0.
DR   SMR; Q823P0.
DR   SWISS-2DPAGE; GET REGION ON 2D PAGE.
KW   ATP-binding; Complete proteome; DNA replication; DNA-binding.
FT   NP_BIND     156    163       ATP (Potential).
SQ   SEQUENCE   450 AA;  51099 MW;  CF440A7B300210D8 CRC64;
     MLTCSDCSTW EQFVNYVKTR CSKTAFENWI SPIQIIEETQ EKIRLEVPNI FVQNYLLDNY
     KQDLCSFVPL DAQGEPALEF VVAEIKKAPA QPIAPREPQE SPAETFEESK DFELKLNAAY
     RFDNFIEGPS NQFVKSAAVG IAGRPGRSYN PLFIHGGVGL GKTHLLHAVG HYVREHHKNL
     RVHCITTEAF INDLVQHLRL KSIDKMKNFY RSLDLLLVDD IQFLQNRQNF EEEFCNTFET
     LINLNKQIVI TSDKPPGQLK LSERIIARME WGLVAHVGIP DLETRVAILQ HKAEQKGLHI
     PNEIAFYIAD HIYGNVRQLE GAINKLTAYC RLFGKTLTES IVRDTLRELF RSPSKQKVSV
     ESILKSVATV FQVKLQDLKG TSRSKELVLA RQVAMYLAKT LITDSLVAIG SAFGKTHSTV
     LYACKTIEQK IEKDETLTRQ ISLCKNHIVG
//
```

Specialized Sequence Databases

- Focus on a specific type of sequences
- Sequences are often modified or specially annotated
- Usage depends on the database
- Examples:
  - Ribosomal RNA databases
  - Immunology databases

Protein domain databases

The University of Georgia

- Pfam (http://www.sanger.ac.uk/Software/Pfam/)
  - Collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families
- SMART **(a Simple Modular Architecture Research Tool)**
  - Identification and annotation of genetically mobile domains and the analysis of domain architectures
  - (http://smart.embl-heidelberg.de/help/smart_about.shtml
- CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi)
  - Combines SMART and Pfam databases
  - Easier and quicker search