Intro to Biological Databases

Lec˙04˙slides

Outline for today

1. Bioinformatics Databases
   - Flat files versus linked files
   - Highly-linked Structures
   - Relational databases
   - Data Quality
   - Quick NCBI Demo
2. Summary
3. Reading Assignment

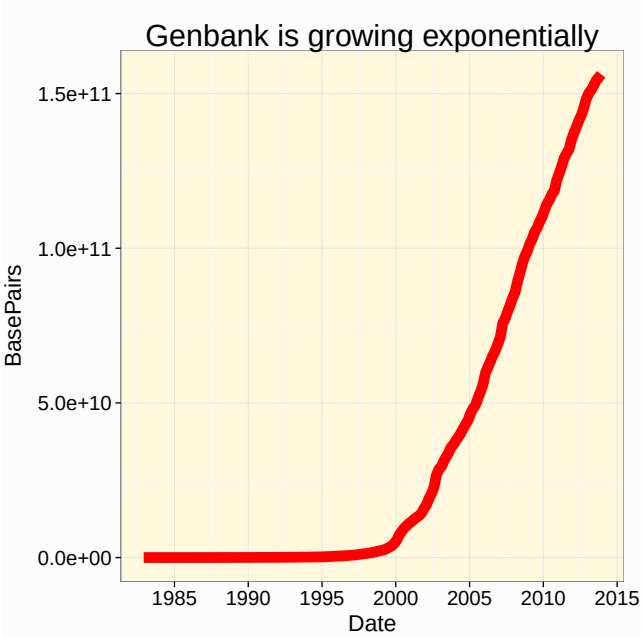
Outline

1 Bioinformatics Databases
Flat files versus linked files
Highly-linked Structures
Relational databases
Data Quality
Quick NCBI Demo

2 Summary

3 Reading Assignment

How do we do this efficiently?



simple query →

```
>mouse hexokinase
CCTGGTTAGTCGTTAC
TCATCGTTCGAGGCGG
...
```

Outline

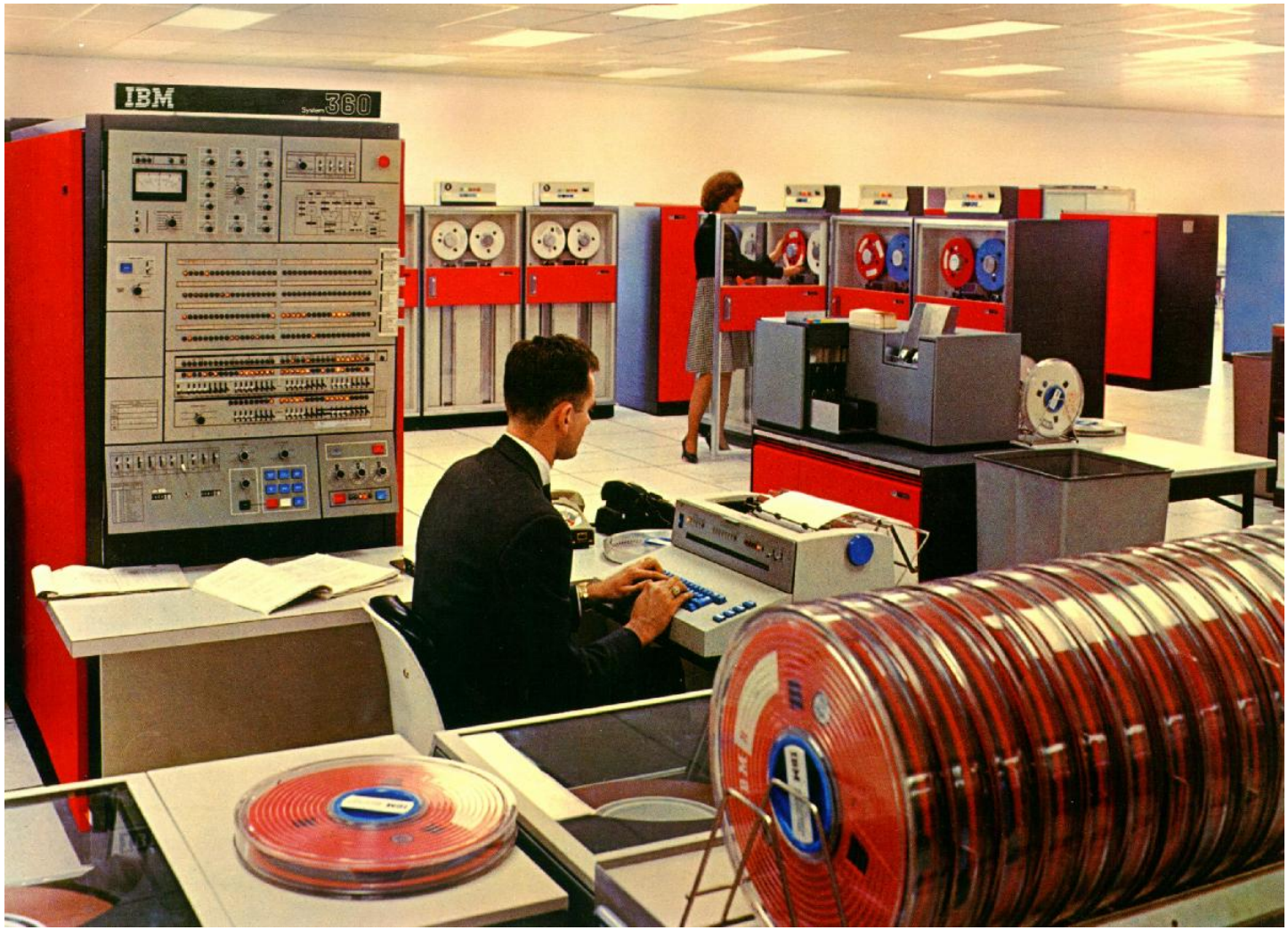
1 Bioinformatics Databases
Flat files versus linked files
Highly-linked Structures
Relational databases
Data Quality
Quick NCBI Demo
2 Summary
3 Reading Assignment

Simple **flat file** nucleotide sequence


```
>gi|24653803|Drosophila melanogaster hexokinase C
GTTTCCAAGGCGCACTGCATCTCAACGCCTGGCTCTTATCAGGCACCCAGG
GGCTCGCCAGGCGCCTGGTTAGTCGTTACTCATCGTTCGAGGCGGTTACAA
CAAGCAAAATGCTGGACGCGGAGGTGCGAGAACTTATGCAACCCTTTGTGT
GGAAGTGTACAGTCGCTTTTGCCTGGAAGTGGCCCGTGGACTTAAGCGGTC
GTCAAGTGTTTTCCCACGTACGTGCAGGATCTGCCCACGGGCGACGAGATG
ATCTCGGCGGTACCAACTTCCGAGTACTGCTCGTCTCGCTGAAAGGTCACC
TCAGATCTATGCCGTGCCAAAGGACCTGATGGTGGGGCCCGGTGTGGACCT
TGCCTGGCCAAATTTGTGGAGAAACACGACATGAAGACCGCATATCTGCCA
TCCCTTGCGTGCAACTAGGCCTTAAGGAGGGCATCCTGGTACGCTGGACTA
GGTTGAGGGCGAGGATGTGGGCCGCATGCTGCACGAGGCCATTCAGCGGCG
GTGGTGGCTATACTCAACGATACCACTGGCACCTTGATGTCCTGCGCCCAT
```

What kind of flat file is this?

Flat files are ancient computer history



http://static.cray-cyber.org/Comprooms/LARGE/IBM_360.jpg

Flat file databases

(A)

NAME	TELEPHONE	ADDRESS
S. Claus	0203 450	The North Pole, Lapland
M. Mouse	0202 453	Disneyworld, Florida
A. Moonman	0104 459	Craterland, The Moon

(B) GenBank Flat-File Format

```
LOCUS       SCU49845     5028 bp    DNA
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.
```

Properties of flat files

- Technically trivial – just a "document"
- Examples:
  - `>Hexokinase`
    `AACCTTGTCCAGGCATTACGGAGAA...`
  - Atomic coordinates of a protein model:

Atom	X	Y	Z
H213	213	423	322
N53	423	593	89
C67	235	675	865
⋮	⋮	⋮	

- Flat files commonly used for small items (bytes to megabytes)
  - Informal definition: 1 byte = 1 text character

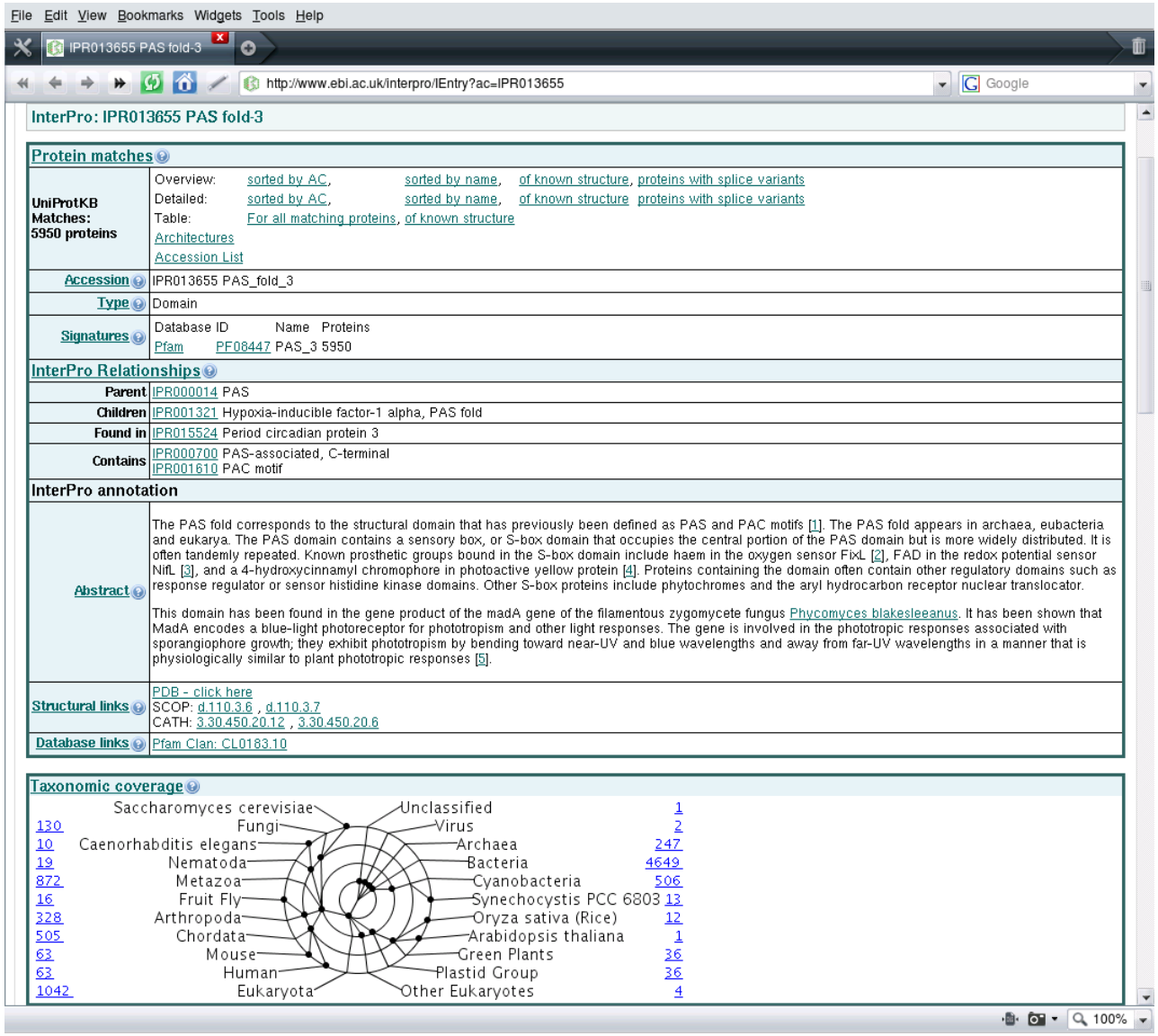
Big limitation: Flat files do not cross-reference other flat files

```
LOCUS       NM_079935 1578 bp mRNA linear
DEFINITION  Drosophila melanogaster hexokinase C
ACCESSION   NM_079935
REFERENCE
  AUTHORS   Hoskins,R.A., Carlson,J.W., Kennedy,C.,
            Acevedo,D., Evans-Holm,M., Frise,E.,
            Wan,K.H., Park,S., Mendez-Lago,M.,
            Rossi,F., Villasante,A., Dimitri,P.,
            Karpen,G.H. and Celniker,S.E.
  TITLE     Sequence finishing and mapping of Drosophila
            melanogaster heterochromatin
  JOURNAL   Science 316 (5831), 1625-1628 (2007)
  PUBMED    17569867
FEATURES    /protein_id="NP_524674.1"
```

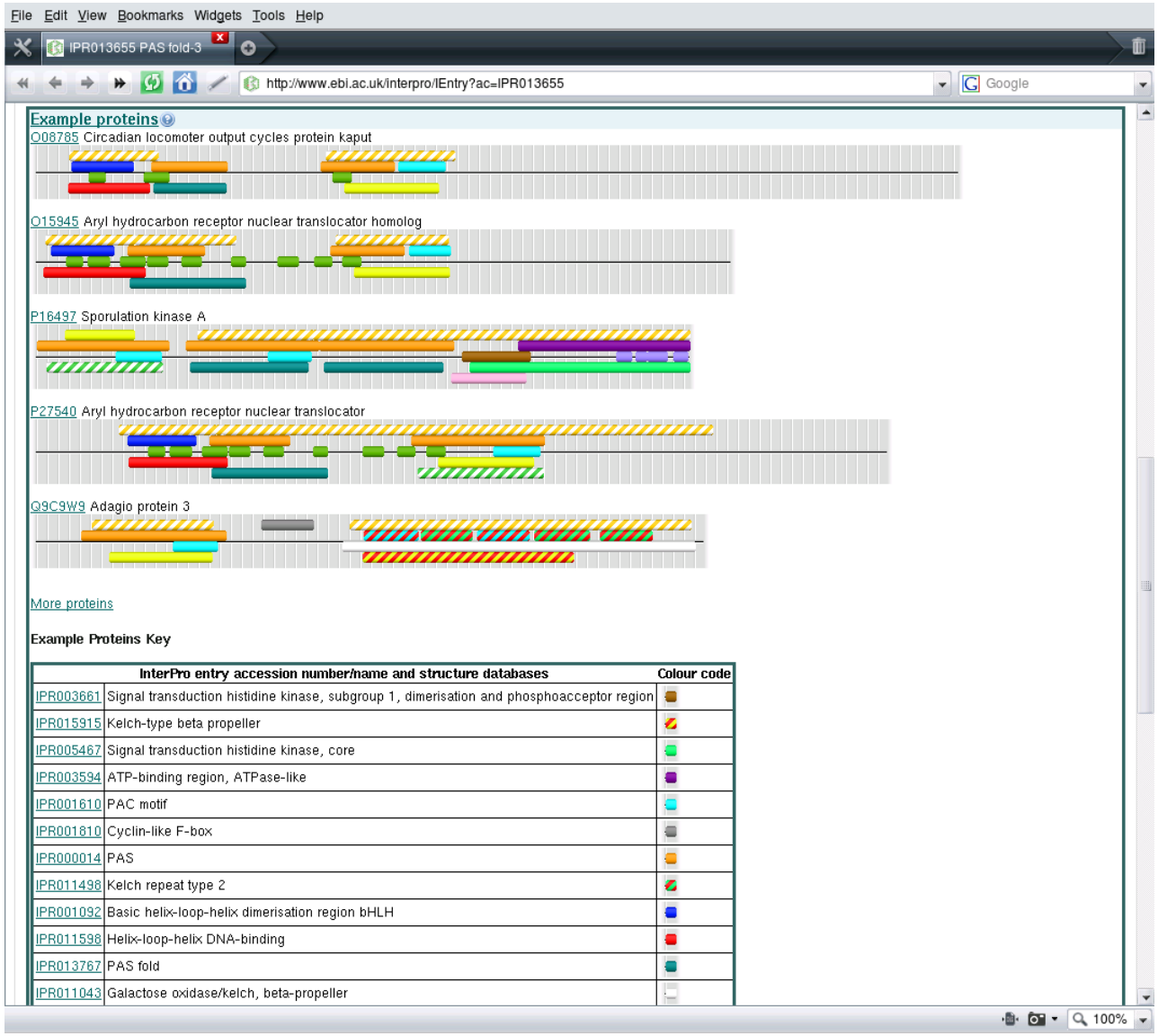
Outline

1 Bioinformatics Databases
Flat files versus linked files
Highly-linked Structures
Relational databases
Data Quality
Quick NCBI Demo
2 Summary
3 Reading Assignment

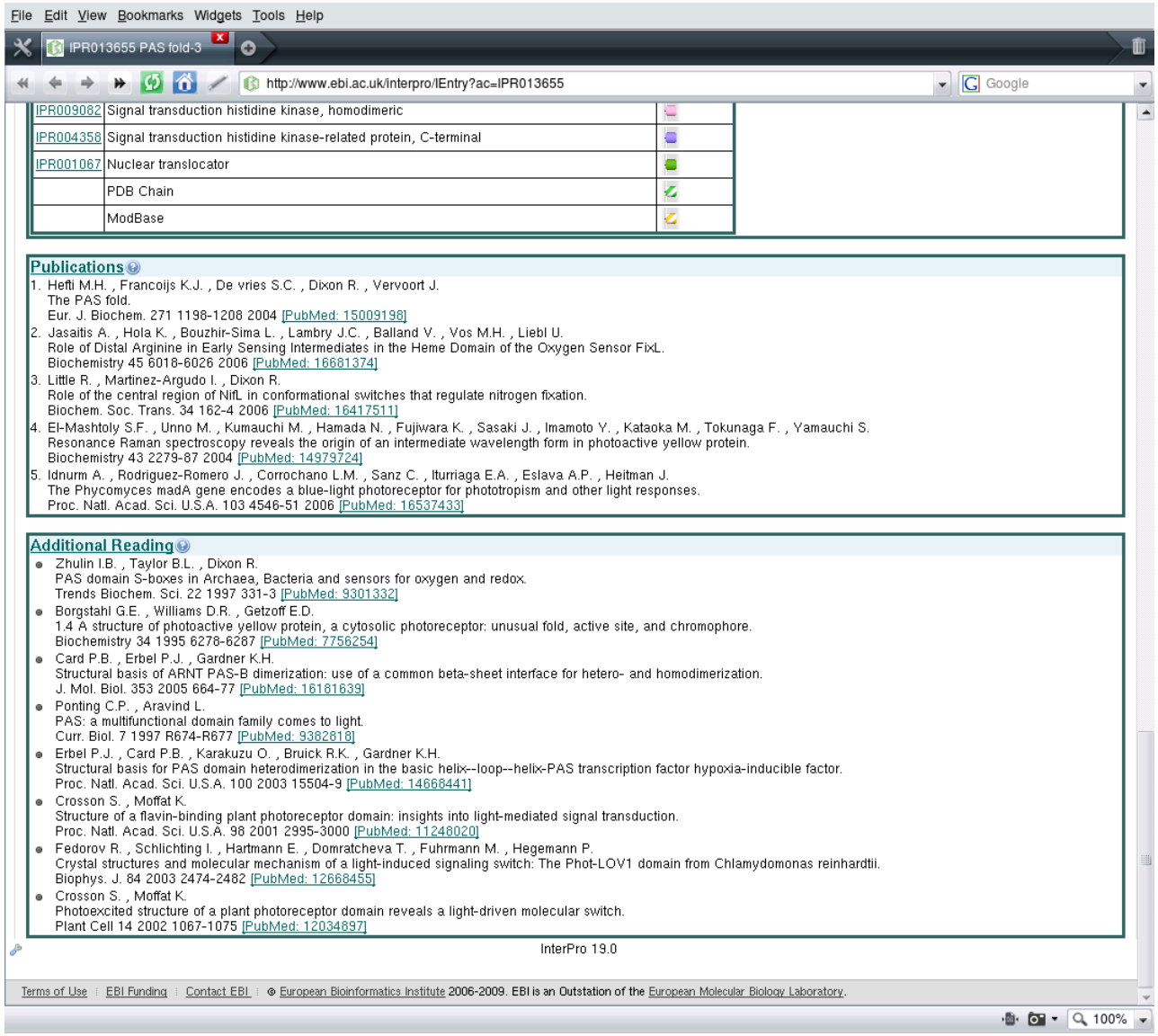
Contrast: We need highly-linked structures I



Contrast: We need highly-linked structures II



Contrast: We need highly-linked structures III



Contrast flat files with relational databases

A Relational database uses keys to relate data

protab1

Protein-code	Protein-name	Length	Species-origin
P1001	Hemoglobin	145	Bovine
P1002	Hemoglobin	136	Ovine
P1003	Eye Lens Protein	234	Human
.....			

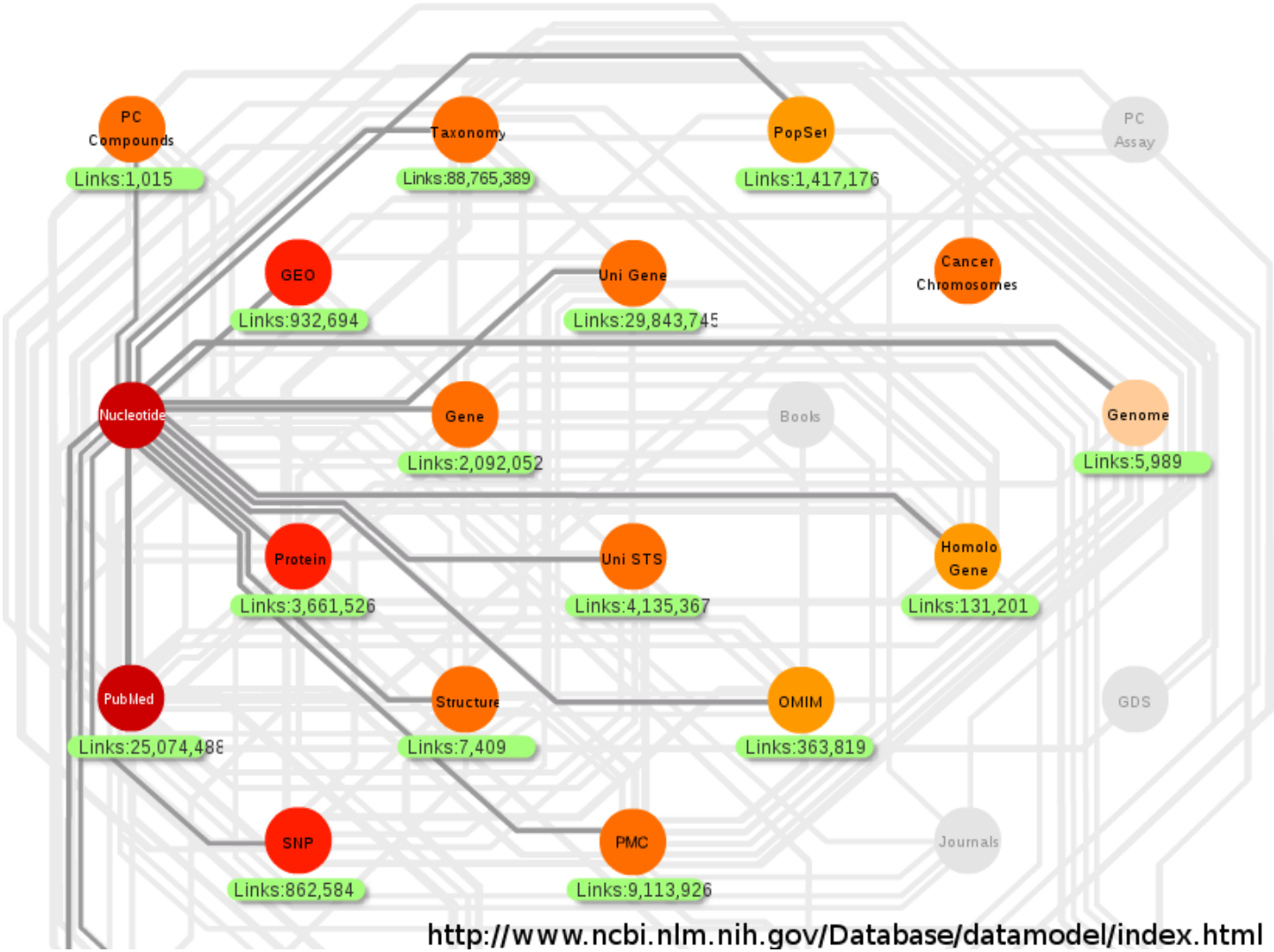
protab2

Protein-code	Protein-sequence
P1001	MDRTTHGFDLKLLSPRTVNQWLMLALFFGHS...
P1002	MDKTSHGFEIKLLTPKKLQQWLMIAIYFGHT...
P1003	SRTHEEEGKLMQWPPRPLYIALFTEPPYP...
.....	

Outline

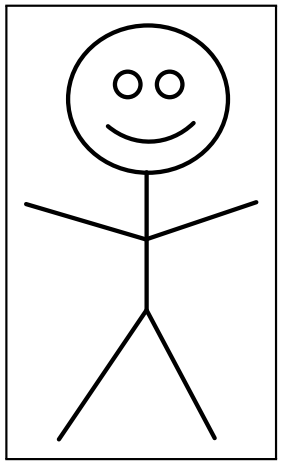
1 Bioinformatics Databases
Flat files versus linked files
Highly-linked Structures
Relational databases
Data Quality
Quick NCBI Demo
2 Summary
3 Reading Assignment

Relational databases link complex data



Metadata = annotation = "data about the data"

Data of a picture $\longrightarrow \longrightarrow \longrightarrow \longrightarrow \longrightarrow \longrightarrow \longrightarrow$



Title	Aunt Carla at the beach
Date	2014.02.07
Camera	Canon PowerShot SX280

Metadata about the picture:

String	Title
YYYY.MM.DD	Date
Brand Type Model	Camera

Data of a gene:

Sequence	`AACCGGTACCTAGAC...`
Name	hexokinase
Location	Chr 4 18247-19345

Metadata about the gene:

[ACGT]	Sequence
String	Name
"Chr" Chr Start-End	Location

Primary data

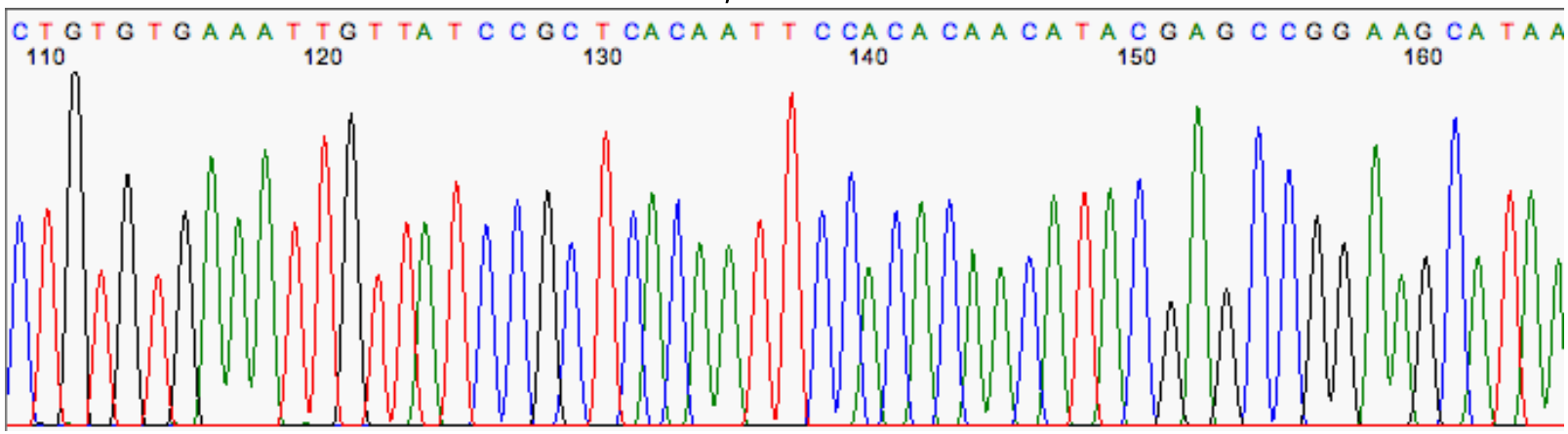
- Raw experimental data
  Examples:DNA sequence, 3D protein structure
- Primary data may be redundant. Example:
  1. Several labs sequence a gene
  2. These labs publish their sequences to a database
  3. Now: multiple copies of sequence in database
  4. Published sequences might differ! Metadata may differ!
     `TACG A TTA` versus `TACG C TTA` How can this happen?

Secondary data

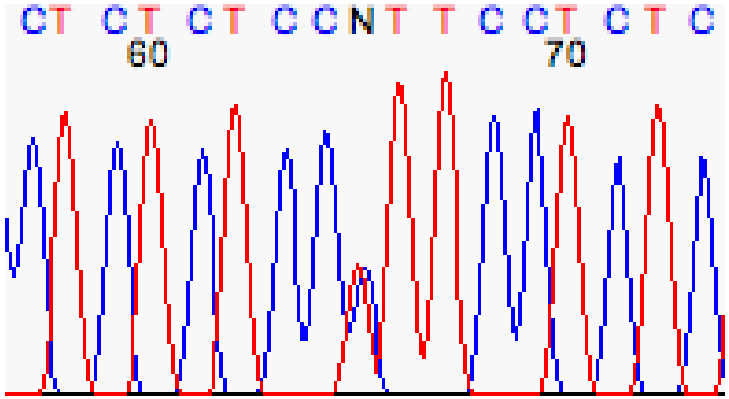
- Curated by experts
- Consensus of primary data
- Nonredundant
- **Always use curated secondary data if available**

DNA chromatograms reveal sequencing errors

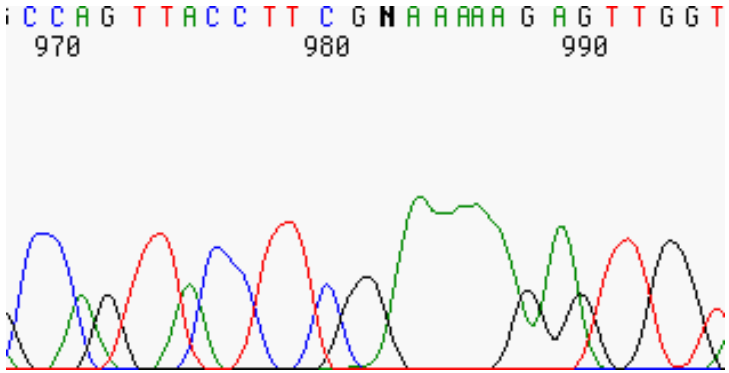
Normal, clean reads



Heterozygous SNP



Towards end of run



http://seqcore.brcf.med.umich.edu/doc/dnaseq/interpret.html

Annotations can be wrong too (and misleading)

Three labs sequence the same gene and put data in GenBank:

- Lab 1:

  Sequence `ACCGGACCTACCGGACCTACCGGACCTACCGGACCT`

  Function: Protein of unknown function

- Lab 2:

  Sequence `ACCGGACCTACCGGACCTACCGGACCTACCGGACCT`

  Function: Shares domains with HOX-family genes

- Lab 3:

  Sequence `ACCGGACCTACCGGACCTACCGGACCTACCGGACCT`

  Function: Appears related to glycolysis

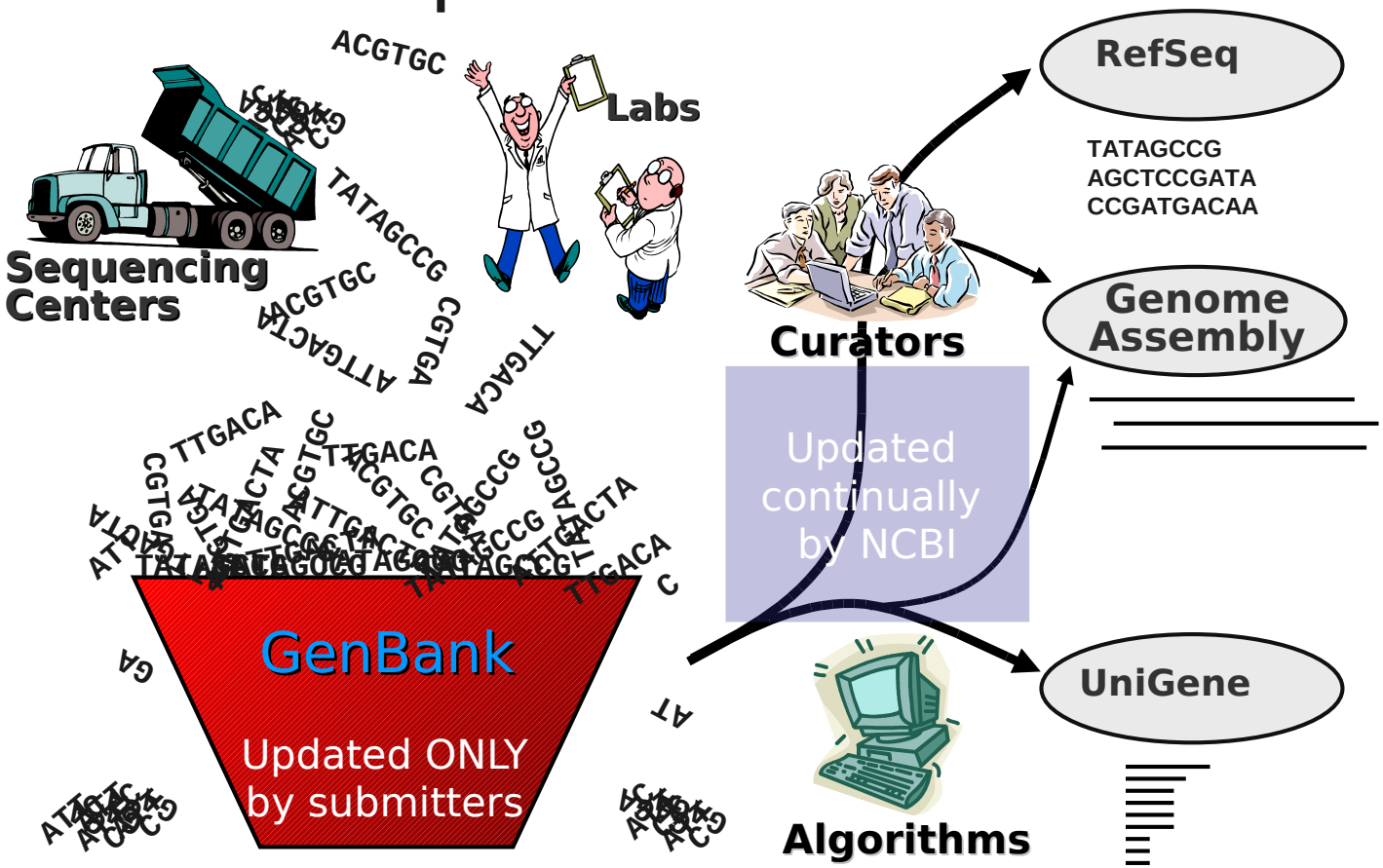


Image from ftp://ftp.ncbi.nih.gov/pub/FieldGuide/Slides/Archive/2008/USC.02.21.08/NCBI_part1.ppt

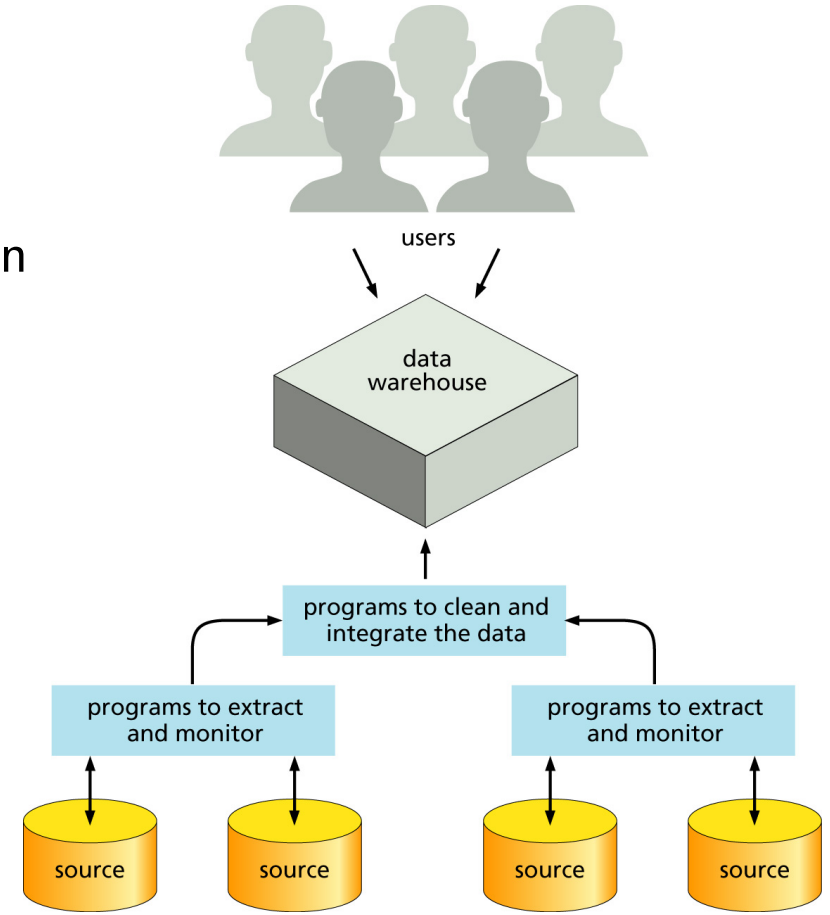
Biological data goes into huge **data warehouses**

Huge data warehouses:

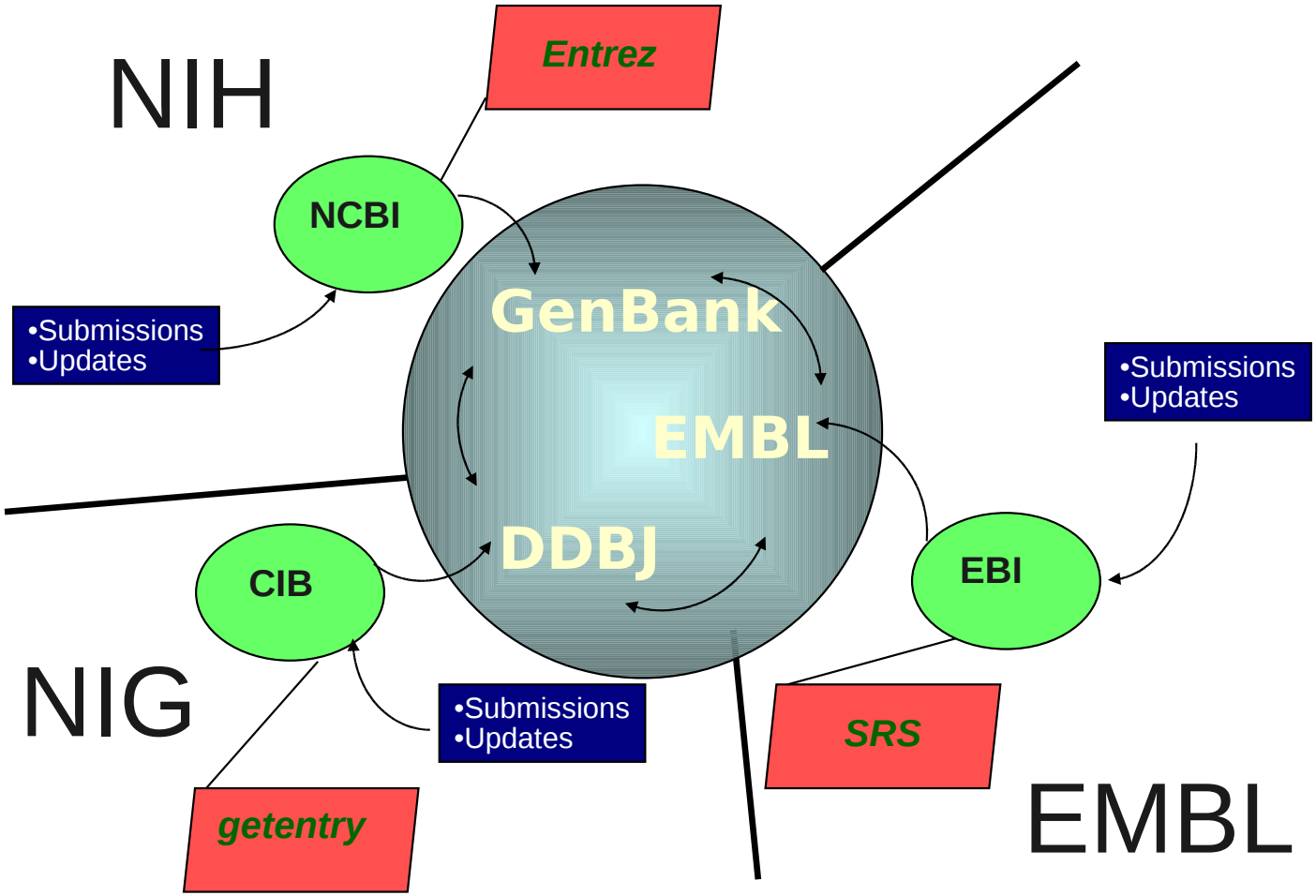
NCBI National Center for Biotechnology Information

EBI European Bioinformatics Institute

PDB Protein Data Bank

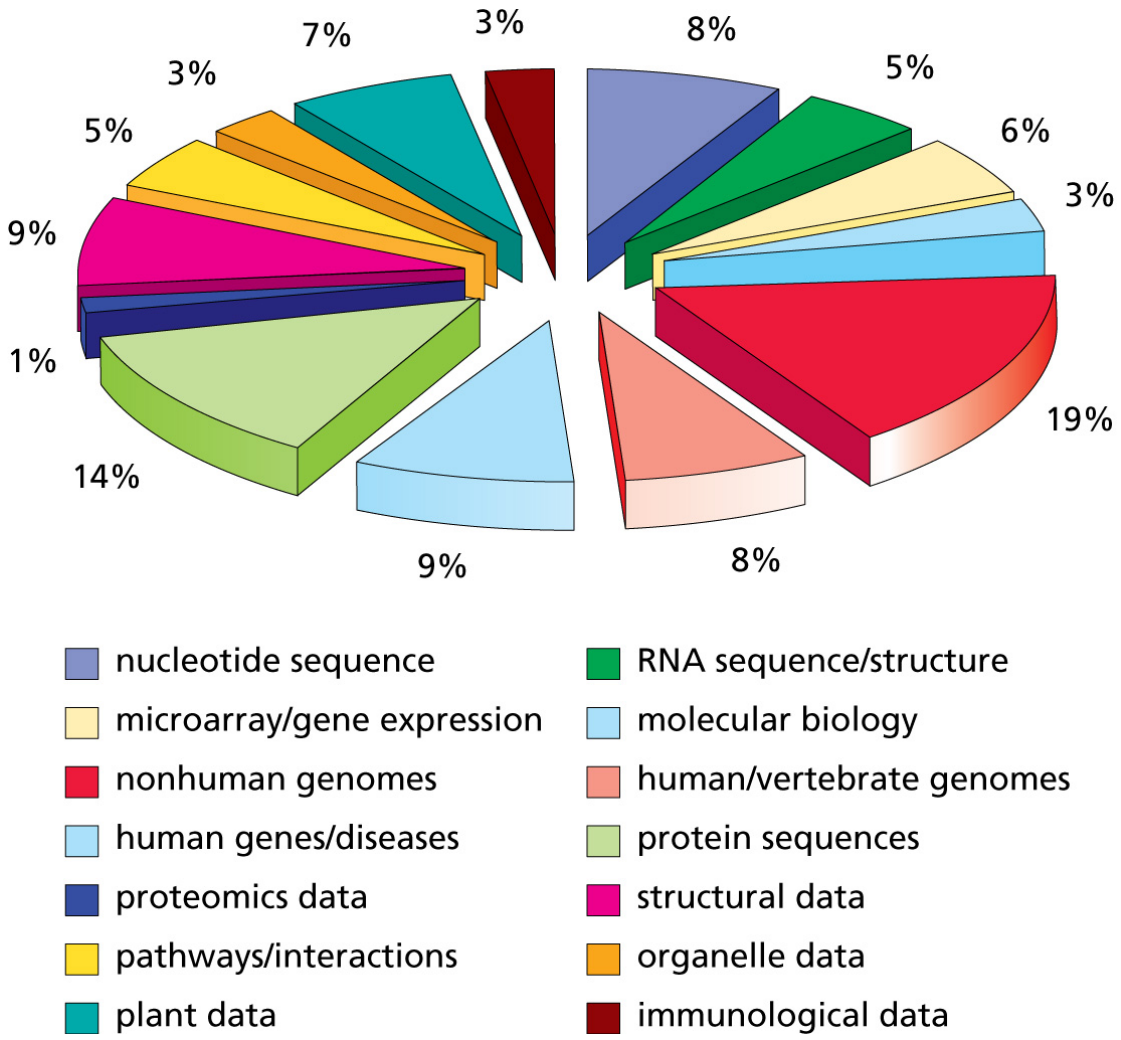


International Sequence Database Collaboration shares biological data between warehouses



ftp://ftp.ncbi.nih.gov/pub/FieldGuide/Slides/Archive/2008/USC.02.21.08/NCBI_part1.ppt

Wide variety of biological data available



Outline

1 Bioinformatics Databases
Flat files versus linked files
Highly-linked Structures
Relational databases
Data Quality
Quick NCBI Demo
2 Summary
3 Reading Assignment

- National Center for Biotechnology Information (NCBI): `http://www.ncbi.nlm.nih.gov/`
- NCBI is **HUGE**
- NCBI is just the tip of the iceberg.
- Lots of our work will be at NCBI.
- Feel free to use other tools too.
- Using NCBI, what can we quickly learn about hemoglobin?

Outline

1 Bioinformatics Databases
Flat files versus linked files
Highly-linked Structures
Relational databases
Data Quality
Quick NCBI Demo

2 Summary

3 Reading Assignment

- Metadata describes/defines the format of the data
- Primary versus secondary data
- Huge primary and secondary sequence databases
- Lots of biological data freely available
- NCBI is a great resource that we will be using A LOT!

Outline

1 Bioinformatics Databases
Flat files versus linked files
Highly-linked Structures
Relational databases
Data Quality
Quick NCBI Demo
2 Summary
3 Reading Assignment

Chapter 2

Section "Command-line Access to Data at NCBI"
to
End of the Chapter

Chapter 3

Section "Introduction"
to
Section "Scoring Matrices"

Pages	**Notes**
42–60	Read
69–79	Read