

Bioinformatics

Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality
Quick NCBI
Demo

Summary

Reading
Assignment

Intro to Biological Databases

Lec'04'slides

Bioinformatics Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality

Quick NCBI
Demo

Summary

Reading
Assignment

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

- 2 Summary

- 3 Reading Assignment

Bioinformatics Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality

Quick NCBI
Demo

Summary

Reading
Assignment

① Bioinformatics Databases

Flat files versus linked files

Highly-linked Structures

Relational databases

Data Quality

Quick NCBI Demo

② Summary

③ Reading Assignment

Sequence databases: The Big Picture

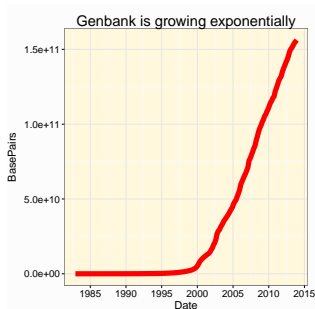
Bioinformatics Databases

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

Summary

Reading Assignment

How do we do this efficiently?



simple query →

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

Bioinformatics
Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality

Quick NCBI
Demo

Summary

Reading
Assignment

- 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

Bioinformatics

Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality

Quick NCBI
Demo

Summary

Reading
Assignment

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

What kind of flat file is this?

Flat files are ancient computer history

Bioinformatics
Databases

Flat files versus
linked files

Highly-linked
Structures

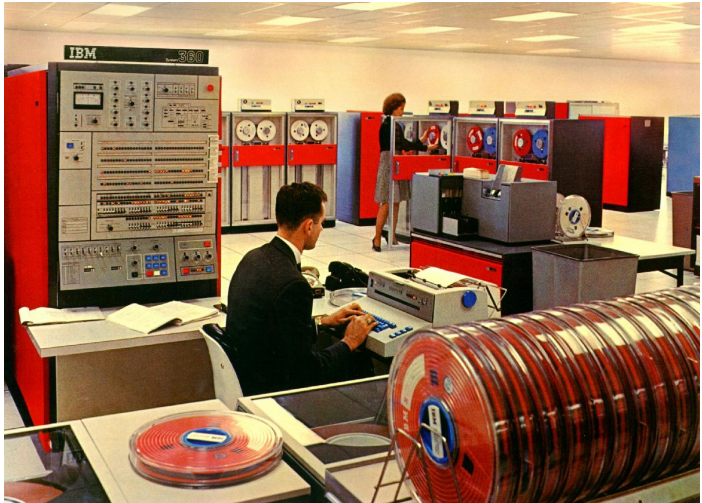
Relational
databases

Data Quality

Quick NCBI
Demo

Summary

Reading
Assignment



(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.
    
```


- 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

Flat files do not cross-reference other flat files

Bioinformatics

Databases

Flat files versus
linked filesHighly-linked
StructuresRelational
databasesData Quality
Quick NCBI
Demo

Summary

Reading
Assignment

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"

Bioinformatics
Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality
Quick NCBI
Demo

Summary

Reading
Assignment

- 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

Bioinformatics
Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality

Quick NCBI

Demo

Summary

Reading
Assignment

File Edit View Bookmarks Widgets Tools Help

IPR013655 PAS fold-3

http://www.ebi.ac.uk/interpro/entry?ac=IPR013655

InterPro: IPR013655 PAS fold-3

Protein matches

Overview: [sorted by AC](#), [sorted by name](#), [of known structure](#), [proteins with splice variants](#)
 Detailed: [sorted by AC](#), [sorted by name](#), [of known structure](#), [proteins with splice variants](#)
 Table: [For all matching proteins, of known structure](#)

UniProtKB
Matches:
5990 proteins

[Architectures](#)
[Accession](#)
[Accession](#)
[Accession](#)

[Accession](#) IPR013655 PAS_fold_3

[Type](#) Domain

[Signatures](#)

Database ID Name Proteins
 Pfam PF08447 PAS_3 5950

InterPro Relationships

Parent [IPR000014](#) PAS
 Children [IPR001321](#) Hypoxia-inducible factor-1 alpha, PAS fold
 Found in [IPR015524](#) Period circadian protein 3
 Contains [IPR000700](#) PAS-associated, C-terminal
[IPR001610](#) PAC motif

InterPro annotation

[Abstract](#)

The PAS fold corresponds to the structural domain that has previously been defined as PAS and PAC motifs [1]. The PAS fold appears in archaea, eubacteria and eukarya. The PAS domain contains a sensory box, or S-box domain that occupies the central portion of the PAS domain but is more widely distributed. It is often tandemly repeated. Known prosthetic groups bound in the S-box domain include haem in the oxygen sensor FxL [2], FAD in the redox potential sensor HSL [3], and 4-hydroxyoctanoyl chromophore in photolabile yellow protein [4]. Proteins containing the domain often contain other regulatory domains such as response regulator or sensor histidine kinase domains. Other S-box proteins include phytochromes and the aryl hydrocarbon receptor nuclear translocator.

This domain has been found in the gene product of the *hnaA* gene of the filamentous zygomycete fungus *Phycomyces blakesleeensis*. It has been shown that *hnaA* encodes a blue-light photoreceptor for phototropism and other light responses. The gene is involved in the phototropic responses associated with sporangium growth; they exhibit phototropism by bending toward near-UV and blue wavelengths and away from far-UV wavelengths in a manner that is physiologically similar to plant phototropic responses [5].

[Structural links](#)

[PDB - click here](#)
 SCOP: [d110.3.6](#), [d110.3.7](#)
 CATH: [3.30.450.20.12](#), [3.30.450.20.6](#)

[Database links](#)

[Pfam](#) [Clan](#) [CL0183.10](#)

Taxonomic coverage

Saccharomyces cerevisiae 130
 Fungi 10
 Caenorhabditis elegans 19
 Nematoda 872
 Metazoa 16
 Fruit Fly 328
 Arthropoda 505
 Chordata 63
 Mouse 63
 Human 1042
 Eukaryota 1042
 Unclassified 1
 Virus 2
 Archaea 242
 Bacteria 4649
 Cyanobacteria 506
 Synechocystis PCC 6803 13
 Oryza sativa (Rice) 12
 Arabidopsis thaliana 1
 Green Plants 36
 Plastid Group 36
 Other Eukaryotes 4

Contrast: We need highly-linked structures II

Bioinformatics
Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality

Quick NCBI
Demo

Summary

Reading
Assignment

File Edit View Bookmarks Widgets Tools Help

PR013655 PAS 100-3

http://www.ebi.ac.uk/interpro/entry?ac=IPR013655

Example proteins

Q08785 Circadian locomotor output cycles protein kaput

Q115345 Aryl hydrocarbon receptor nuclear translocator homolog

P216497 Sporulation kinase A

P27540 Aryl hydrocarbon receptor nuclear translocator

Q9C9W9 Adagio protein 3

More proteins

Example Proteins Key

InterPro entry accession number	name and structure databases	Colour code
IPR003661	Signal transduction histidine kinase, subgroup 1, dimerisation and phosphoacceptor region	
IPR015315	Kelch-type beta propeller	
IPR005467	Signal transduction histidine kinase, core	
IPR003594	ATP-binding region, ATPase-like	
IPR001610	PAC motif	
IPR001810	Cyclin-like F-box	
IPR000014	PAS	
IPR011439	Kelch repeat type 2	
IPR001092	Basic helix-loop-helix dimerisation region bHLH	
IPR011538	Helix-loop-helix DNA-binding	
IPR013767	PAS fold	
IPR011043	Galactose oxidase/kelch, beta-propeller	

Contrast: We need highly-linked structures III

Bioinformatics
Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality

Quick NCBI
Demo

Summary

Reading
Assignment

File Edit View Bookmarks Widgets Tools Help

PR013655 PAS_fold

http://www.ebi.ac.uk/interpro/entry/ac=IPR013655

Accession	Description	Links
IPR005082	Signal transduction histidine kinase, homodimeric	PDB Chain
IPR004359	Signal transduction histidine kinase-related protein, C-terminal	ModBase
IPR010627	Nuclear translocator	
IPR013655	PAS_fold	

Publications

- Heff M.H., François K.J., De vries S.C., Dixon R., Vervoort J. The PAS fold. *Eur. J. Biochem.* 271 1198-1208 2004 [[PubMed: 15009198](#)]
- Jasalski A., Hols K., Bouzhir-Sima L., Lambry J.C., Ballard V., Vos M.H., Liebi U. Role of Distal Arginine in Early Sensing Intermediates in the Heme Domain of the Oxygen Sensor Fod. *Biochemistry* 45 6019-6026 2006 [[PubMed: 16601374](#)]
- Little R., Martinez-Argente I., Dixon R. Role of the central region of NtrB in conformational switches that regulate nitrogen fixation. *Biochem. Soc. Trans.* 34 162-4 2006 [[PubMed: 16417511](#)]
- El-Mashtoly S.F., Unno M., Kumauchi M., Hamada N., Fujiwara K., Saraki J., Imamoto Y., Katsuka M., Tokunaga F., Yamauchi S. Resonance Raman spectroscopy reveals the origin of an intermediate wavelength form in photoactive yellow protein. *Biochemistry* 43 2279-87 2004 [[PubMed: 14575724](#)]
- Idnani A., Rodriguez-Romero J., Corrochano L.M., Sanz C., Iturraga E.A., Eslava A.P., Heitman J. The *Phycomyces mada* gene encodes a blue-light photoreceptor for phototropism and other light responses. *Proc. Natl. Acad. Sci. U.S.A.* 103 4546-51 2006 [[PubMed: 16537430](#)]

Additional Reading

- Zhulin I.B., Taylor B.L., Dixon R. PAS domain S-boxes in Archaea, Bacteria and sensors for oxygen and redox. *Trends Biochem. Sci.* 22 197 331-3 [[PubMed: 5301332](#)]
- Bongstien G.E., Williams D.R., Getoff E.D. 1.4 A structure of photoactive yellow protein, a cytosolic photoreceptor: unusual fold, active site, and chromophore. *Biochemistry* 34 1995 6278-6287 [[PubMed: 7756254](#)]
- Card P.B., Erbel P.J., Gardner K.H. Structural basis of hPRV PAS-B dimerization: use of a common beta-sheet interface for hetero- and homodimerization. *J. Mol. Biol.* 353 2005 664-77 [[PubMed: 16101635](#)]
- Ponding C.P., Aravind L. PAS: a multifunctional domain family comes to light. *Curr. Biol.* 7 1997 R674-R677 [[PubMed: 9380814](#)]
- Erbel P.J., Card P.B., Karakuzu O., Bruck R.K., Gardner K.H. Structural basis for PAS domain heterodimerization in the basic helix-loop-helix-PAS transcription factor hypoxia-inducible factor. *Proc. Natl. Acad. Sci. U.S.A.* 100 2003 15504-9 [[PubMed: 14660441](#)]
- Crosson S., Moffat K. Structure of a flavin-binding plant photoreceptor domain: insights into light-mediated signal transduction. *Proc. Natl. Acad. Sci. U.S.A.* 98 2001 2995-3000 [[PubMed: 11248020](#)]
- Fedorov R., Schlichting I., Hartmann E., Domratcheva T., Fuhrmann M., Hegemann P. Crystal structures and molecular mechanism of a light-induced signaling switch: The Phot-LOV1 domain from *Chlamydomonas reinhardtii*. *Biophys. J.* 84 2003 2474-2482 [[PubMed: 12660455](#)]
- Crosson S., Moffat K. Photoexcited structure of a plant photoreceptor domain reveals a light-driven molecular switch. *Plant Cell* 14 2002 1097-1075 [[PubMed: 12034637](#)]

InterPro 19.0

Terms of Use | EBI Funding | Contact EBI | European Bioinformatics Institute 2006-2009. EBI is an institution of the European Molecular Biology Laboratory.

100%

Contrast flat files with relational databases

Bioinformatics
Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality

Quick NCBI
Demo

Summary

Reading
Assignment

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	MDRTHGFDLKLSPRTVNQWLMLALFFGHS...
P1002	MDKTSHGFEIKLLTPKKLQQWLMIAIYFGHT...
P1003	SRTHEEEGKLMQWPPRPLYIALFTEPPYP...
.....	

Bioinformatics Databases

Flat files versus
linked files

Highly-linked
Structures

**Relational
databases**

Data Quality
Quick NCBI
Demo

Summary

Reading
Assignment

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

Bioinformatics Databases

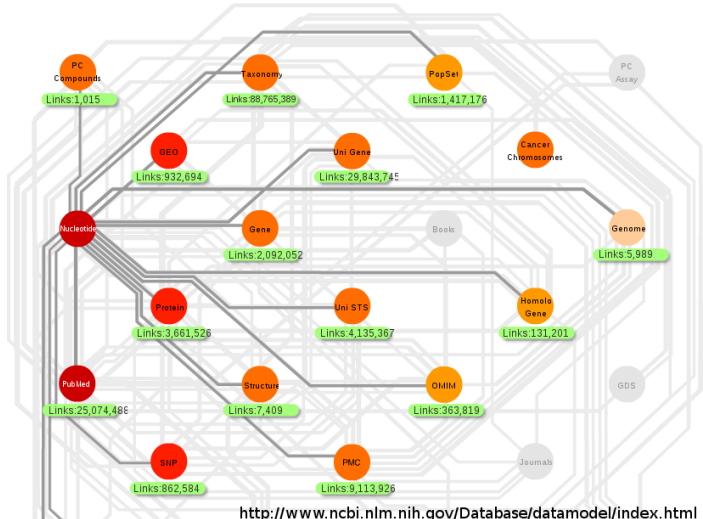
Flat files versus
linked files
Highly-linked
Structures

Relational databases

Data Quality
Quick NCBI
Demo

Summary

Reading
Assignment



Metadata = annotation = “data about the data”

Bioinformatics

Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality

Quick NCBI
Demo

Summary

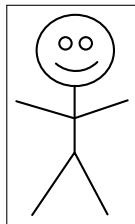
Reading
Assignment

Data of a picture → → → → → → →

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

Bioinformatics

Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality

Quick NCBI
Demo

Summary

Reading
Assignment

① Bioinformatics Databases

Flat files versus linked files

Highly-linked Structures

Relational databases

Data Quality

Quick NCBI Demo

② Summary

③ Reading Assignment

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

Bioinformatics

Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

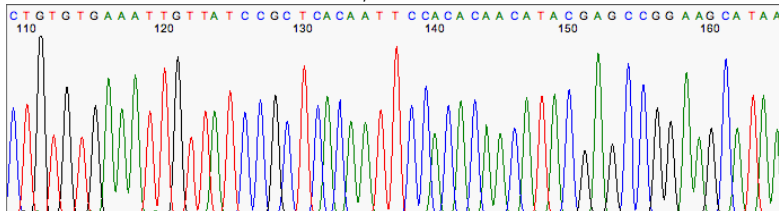
Data Quality

Quick NCBI
Demo

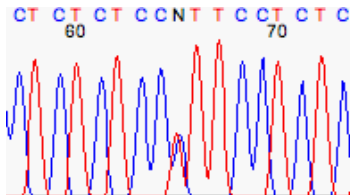
Summary

Reading
Assignment

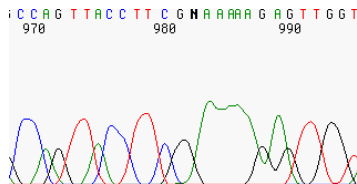
Normal, clean reads



Heterozygous SNP



Towards end of run



Annotations can be wrong too (and misleading)

Bioinformatics
Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality

Quick NCBI
Demo

Summary

Reading
Assignment

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

Primary vs. Derivative Sequence Databases



Biological data goes into huge **data warehouses**

Bioinformatics
Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality
Quick NCBI
Demo

Summary

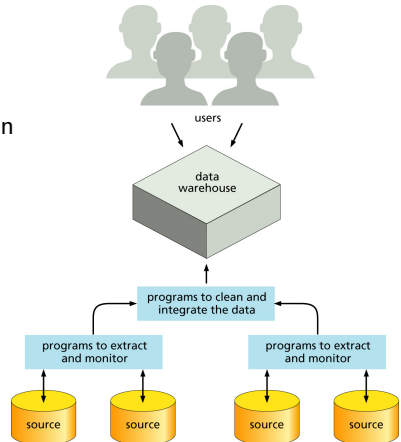
Reading
Assignment

Huge data warehouses:

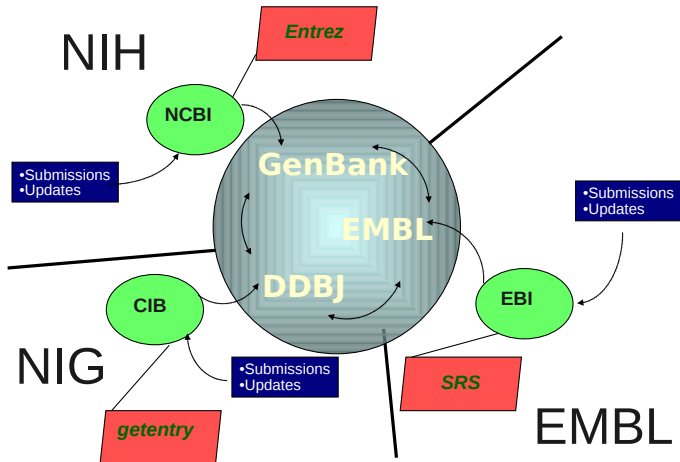
NCBI National Center for
Biotechnology Information

EBI European Bioinformatics
Institute

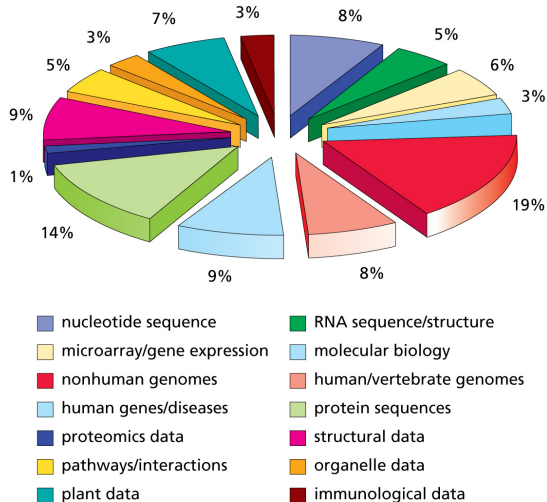
PDB Protein Data Bank



International Sequence Database Collaboration shares biological data between warehouses



Wide variety of biological data available



Bioinformatics
Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality

Quick NCBI
Demo

Summary

Reading
Assignment

Bioinformatics Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality

Quick NCBI
Demo

Summary

Reading
Assignment

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?

Bioinformatics

Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality

Quick NCBI
Demo

Summary

Reading
Assignment

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

2 Summary

3 Reading Assignment

Bioinformatics

Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality
Quick NCBI
Demo

Summary

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!

Bioinformatics

Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality
Quick NCBI
Demo

Summary

Reading
Assignment

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

- 2 Summary

- 3 Reading Assignment

Bioinformatics

Databases

Flat files versus
linked files

Highly-linked
Structures

Relational
databases

Data Quality

Quick NCBI
Demo

Summary

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