

# Understanding and Using Biological Databases

Francis Ouellette

*francis@bioinformatics.ubc.ca*

# Lecture 1.4 Objectives

- Able to recognize various data formats, and know what their primary use is.
- Know, understand and utilize all types of sequence identifiers.
- Know and understand various feature types present in the GenBank flat files.
- Know and understand the various GenBank divisions.

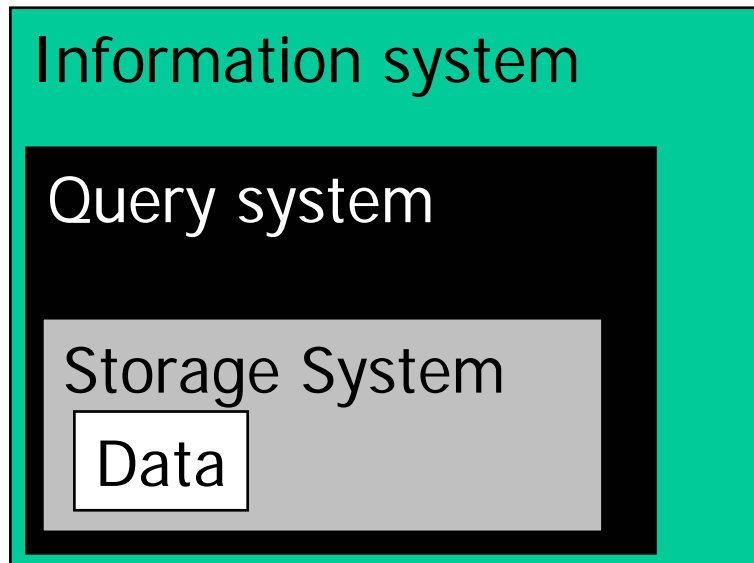
# Outline

- Information landscape
- Data type
- Sequence Databases
- Data Formats
- Other “databases” and “datasets”
- GenBank dissection
  - identifiers
  - divisions

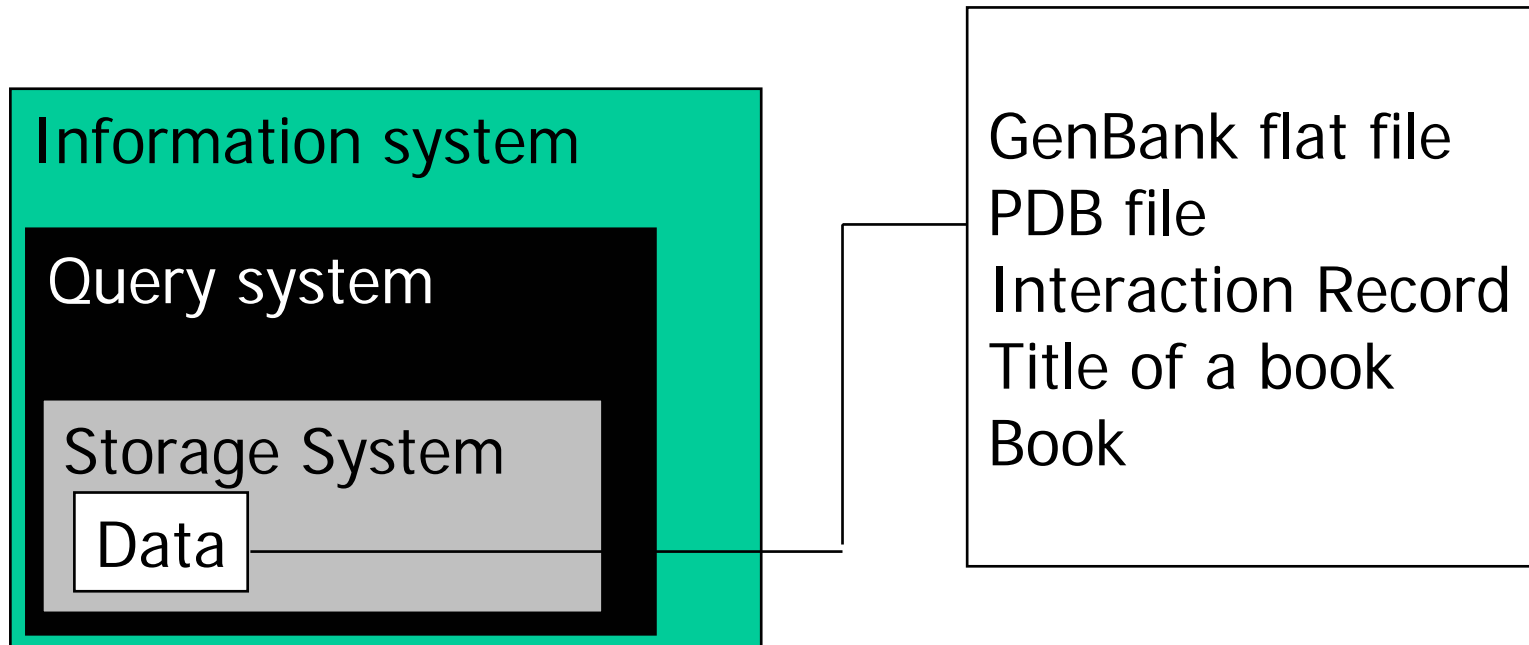
# The reagent: databases

- Organized array of information
- Place where you put things in, and (if all is well) you should be able to get them out again.
- Resource for other databases and tools.
- Simplify the information space by specialization.
- Bonus: Allows you to make discoveries.

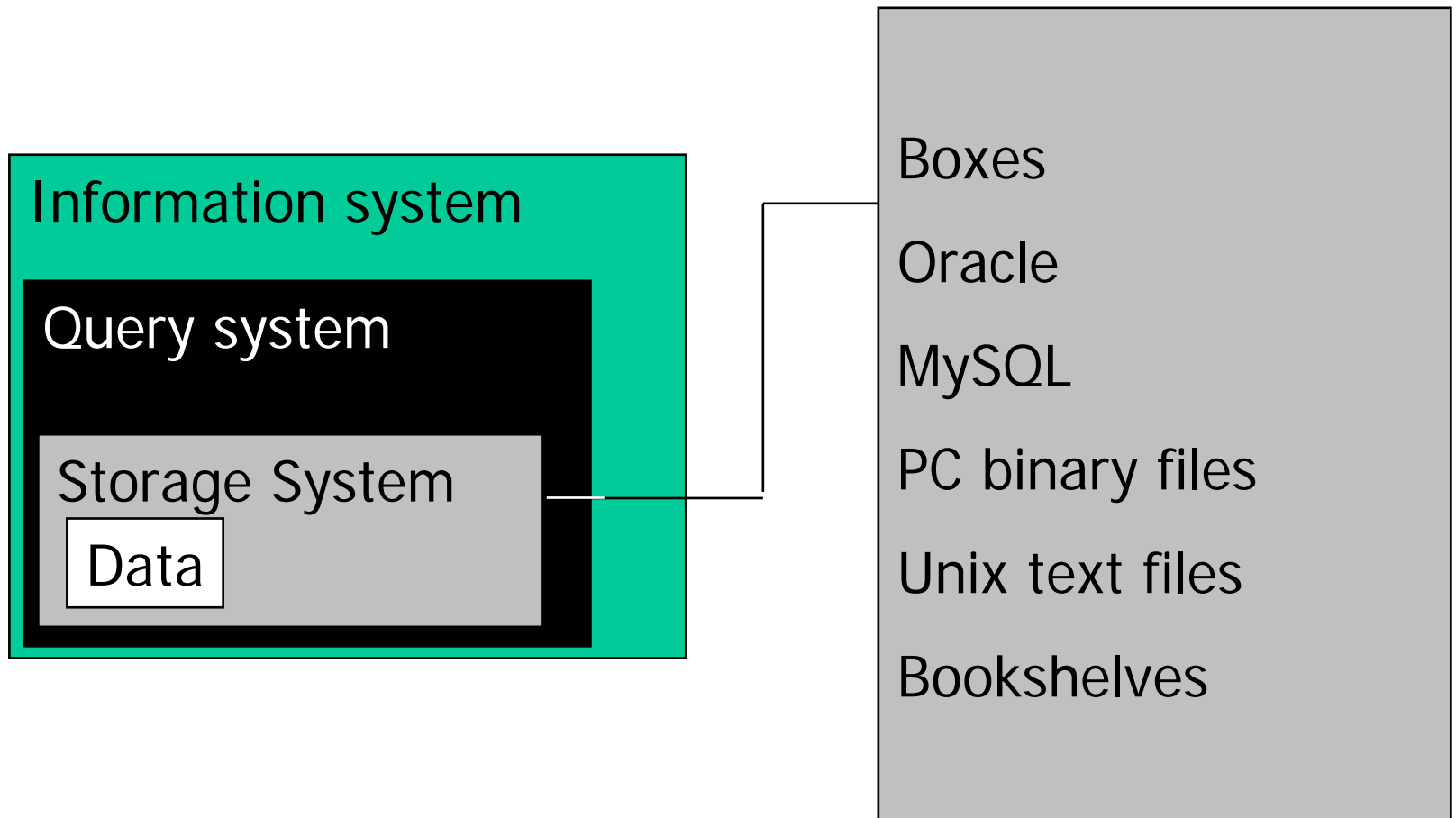
# Databases



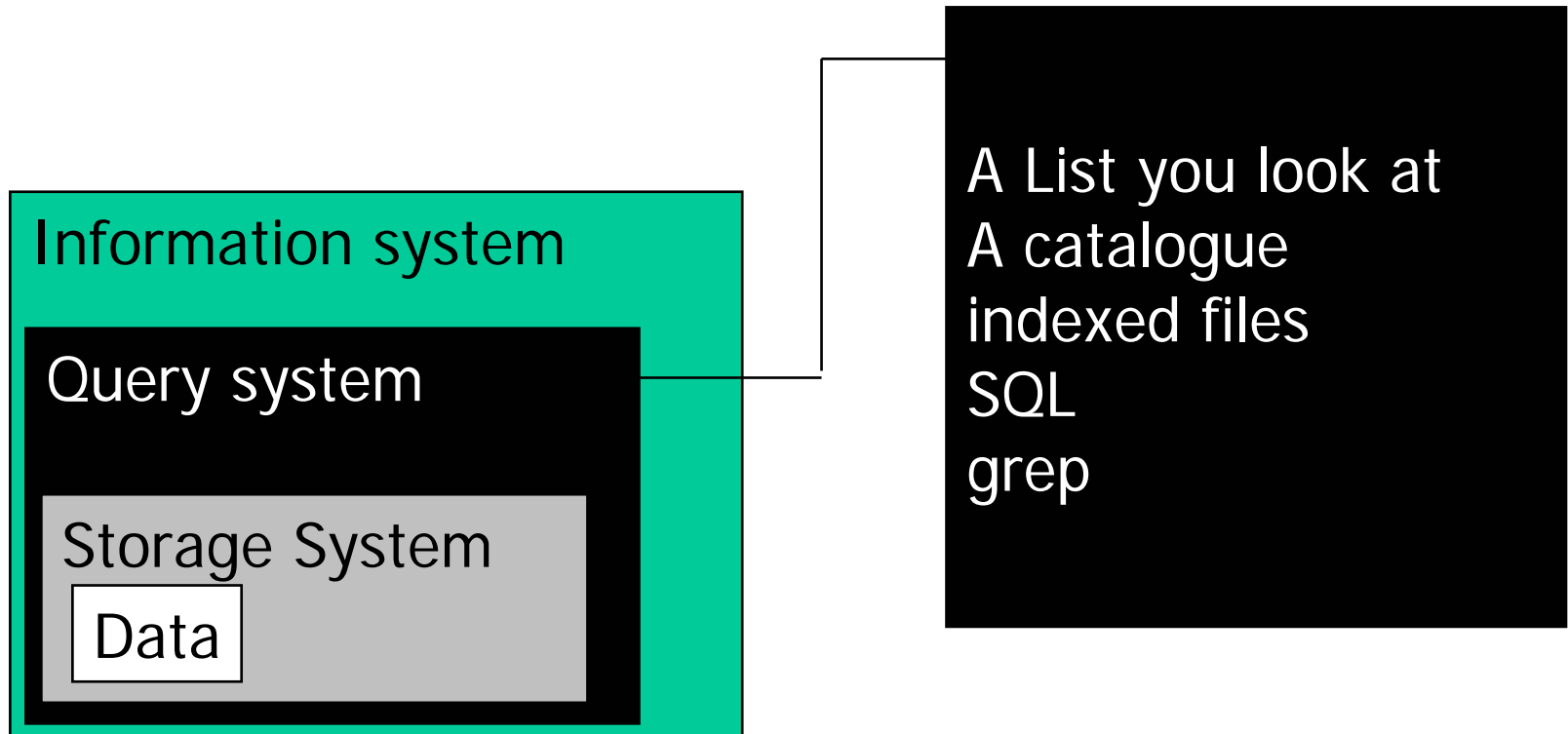
# Databases



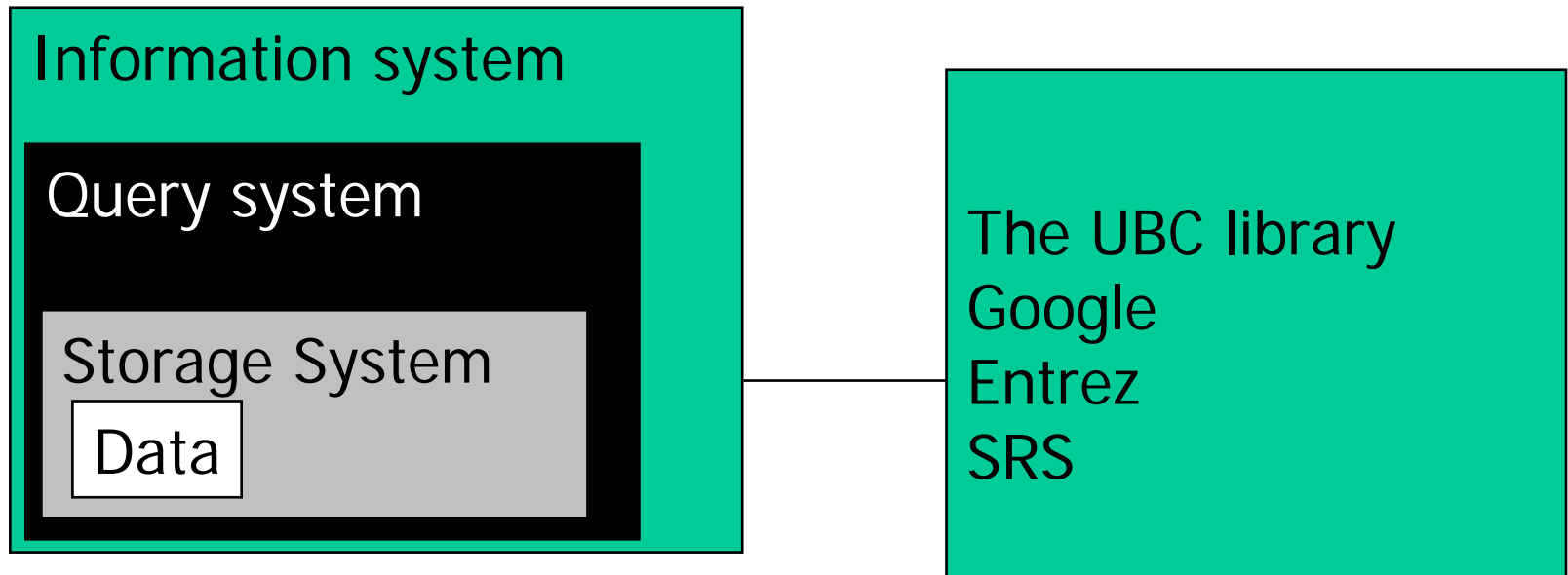
# Databases



# Databases



# Databases



# Bioinformatics Information Space

July 17, 1999

- Nucleotide sequences: 4,456,822
- Protein sequences: 706,862
- 3D structures: 9,780
- Human Unigene Clusters: 75,832
- Maps and Complete Genomes: 10,870
- Different species node: 52,889
- dbSNP 6,377
- RefGenes 515
- human contigs > 250 kb 341 (4.9MB)
- PubMed records: 10,372,886
- OMIM records: 10,695

# The challenge of the information space:

Feb 10 2004

Nucleotide records	36,653,899
Protein sequences	4,436,362
3D structures	19,640
Interactions & complexes	52,385
Human Unigene Cluster	118,517
Maps and Complete Genomes	6,948
Different taxonomy Nodes	283,121
Human dbSNP	13,179,601
Human RefSeq records	22,079
bp in Human Contigs > 5,000 kb (116)	2,487,920,000
PubMed records	12,570,540
OMIM records	15,138

# From a CBW student course evaluation:

“I could probably live the rest of my life happily without ever seeing the ‘growth of GenBank’ curve ... again.”

# Databases

- Primary (archival)
  - GenBank/EMBL/DDBJ
  - UniProt
  - PDB
  - Medline (PubMed)
  - BIND
- Secondary (curated)
  - RefSeq
  - Taxon
  - UniProt
  - OMIM
  - SGD

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

Nucl. Acids. Res. -- Table of Contents (31 [1]) - Netscape 6

File Edit View Search Go Bookmarks Tasks Help

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

**Nucleic Acids Research** OXFORD Journals online

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: [\[Sign up for eTOCs\]](#)

**Contents: Volume 31, Number 1** January 1 2003 [\[Index by Author\]](#)

- [+ Editorial](#)
- [+ Articles](#)

[\[Cover Caption\]](#)

Other Issues:  

Find articles in this issue containing these words:

Enter [\[Search ALL Issues\]](#)

Document: Done (17.025 secs)

[http://nar.oupjournals.org/content/vol32/suppl\\_1/](http://nar.oupjournals.org/content/vol32/suppl_1/)

Nucl. Acids. Res. -- Table of Contents (32 [Database Issue]) - Netscape

File Edit View Go Bookmarks Tools Window Help

[http://nar.oupjournals.org/content/vol32/suppl\\_1/index.shtml](#) Search

HOME HELP FEEDBACK SUBSCRIPTIONS ARCHIVE SEARCH TABLE OF CONTENTS

QUICK SEARCH: [advanced]

Go	Author:	Keyword(s):
<input type="text"/>	<input type="text"/>	<input type="text"/>
Year:	Vol:	Page:

OXFORD Journals online

Receive this page by email each issue: [\[Sign up for eTOCs\]](#)

**Contents: Volume 32, Database Issue** January 1 2004 [\[Index by Author\]](#)

- [Editorial](#)
- [Articles](#)
- [Commercial Database Articles](#)

[\[Cover Caption\]](#)

Other Issues: [←](#) [→](#)

Find articles in this issue containing these words:  Enter [\[Search ALL Issues\]](#)

To see an article, click its [Full Text] link. To review many abstracts, check the boxes to the left of the titles you want, and click the 'Get All Checked Abstract(s)' button. To see one abstract at a time, click its [Abstract] link.

**Editorial:** [+](#)

**FREE EDITORIAL**  
Nucl. Acids. Res. 2004 32: D1. [\[FREE Full Text\]](#)

**Articles:** [+](#)

Michael Y. Galperin  
**FREE The Molecular Biology Database Collection: 2004 update**  
Nucl. Acids. Res. 2004 32: D3-D22. [\[Abstract\]](#) [\[FREE Full Text\]](#) [\[Database Listing\]](#)

# Sequence Databases

- Primary DNA
  - DDBJ/EMBL/GenBank
- Primary protein
  - GenPept/TrEMBL
- Curated DB
  - RefSeq (Genomic, mRNA and protein)
  - Swiss-Prot & PIR -> UniProt (protein)

# What is GenBank?

GenBank is the NIH genetic sequence database of all publicly available DNA and derived protein sequences, with annotations describing the biological information these records contain.

<http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html>

Benson *et al.*, 2004, *Nucleic Acids Res.* **32**:D23-D26

**NIH**

*Entrez*

NCBI

GenBank

EMBL

DDBJ

CIB

EBI

- Submissions
- Updates

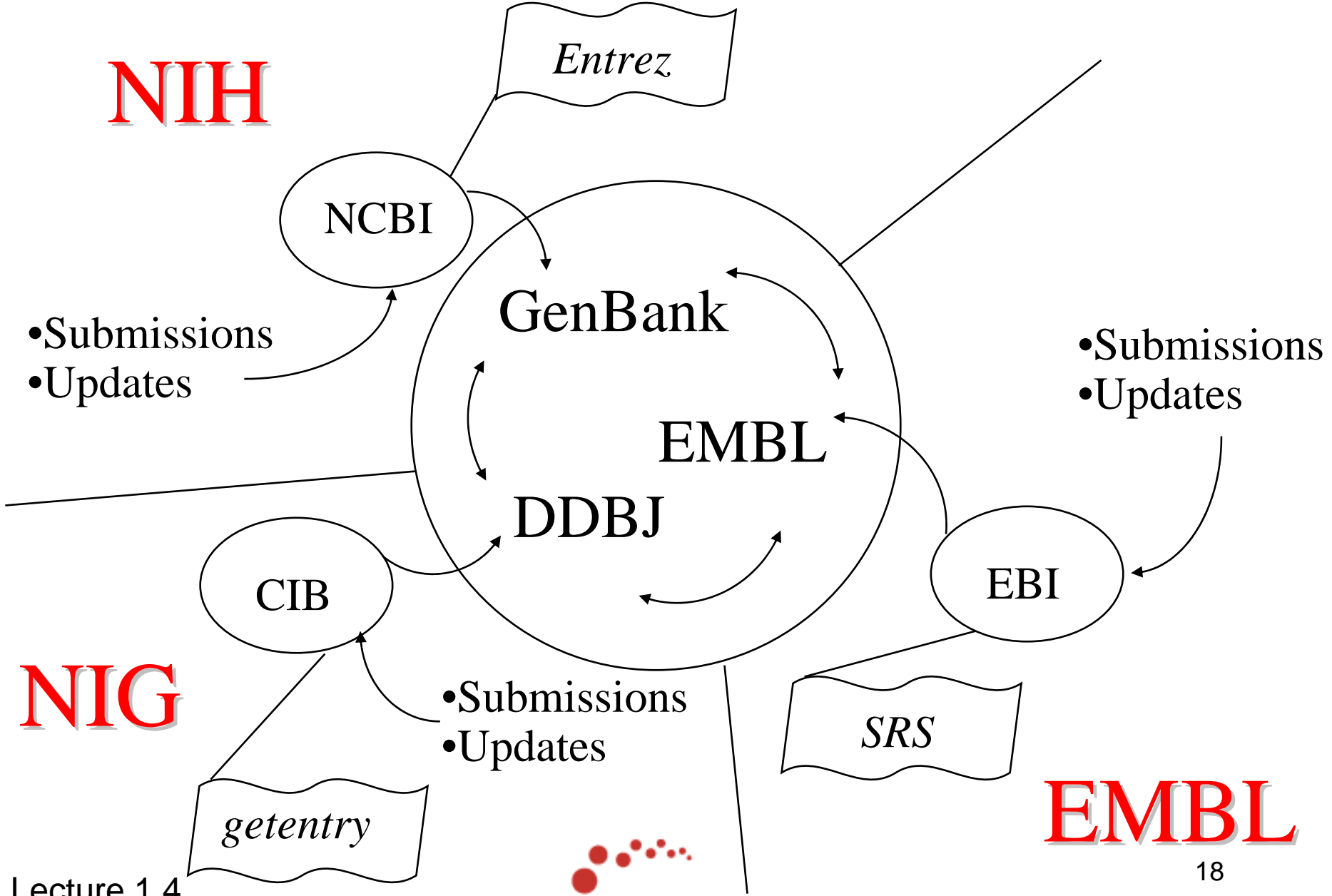
- Submissions
- Updates

- Submissions
- Updates

*SRS*

*getentry*

**EMBL**



# GenBank Flat File (GBFF)

```
LOCUS       MUSNGH             1803 bp     mRNA             ROD             29-AUG-1997
DEFINITION  Mouse neuroblastoma and rat glioma hybridoma cell line NG108-15
            cell TA20 mRNA, complete cds.
ACCESSION   D25291
RID        gi1850791
KEYWORDS    neurite extension activity; growth arrest; TA20.
SOURCE      Murinae gen. sp. mouse neuroblastoma-rat glioma hybridoma
            cell_line:NG108-15 cDNA to mRNA.
ORGANISM    Murinae gen. sp.
            Eukaryotes; mitochondrion eukaryotes; Metazoa; Chordata;
            Vertebrata; Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae;
            Murinae.
REFERENCE   1 (sites)
AUTHORS     Tohda,C., Nagai,S., Tohda,M. and Nomura,Y.
TITLE       A novel factor, TA20, involved in neuronal differentiation: cDNA
            cloning and expression
JOURNAL     Neurosci. Res. 23 (1), 21-27 (1995)
MEDLINE     96064354
REFERENCE   2 (bases 1 to 1803)
AUTHORS     Tohda,C.
TITLE       Direct Submission
JOURNAL     Submitted (18-NOV-1993) to the DDBJ/EMBL/GenBank databases. Chihiro
            Tohda, Toyama Medical and Pharmaceutical University, Research
            Institute for Wakan-yaku, Analytical Research Center for
            Ethnomedicines; 2630 Suititani, Toyama, Toyama 930-01, Japan
            (E-mail:CHIHIRO@ms.toyama-mpu.ac.jp, Tel:+81-764-34-2281(ex.2841),
            Fax:+81-764-34-5057)
COMMENT     On Feb 26, 1997 this sequence version replaced gi:793764.
FEATURES    Location/Qualifiers
            source
                1..1803
                /organism="Murinae gen. sp."
                /note="source origin of sequence, either mouse or rat, has
                not been identified"
                /db_xref="taxon:39108"
                /cell_line="NG108-15"
                /cell_type="mouse neuroblastoma-rat glioma hybridoma"
            misc_signal
                156..163
                /note="AP-2 binding site"
            GC_signal
                647..655
                /note="Sp1 binding site"
            TATA_signal
                694..701
            gene
                748..1311
                /gene="TA20"
            CDS
                748..1311
                /gene="TA20"
                /function="neurite extension activity and growth arrest
                effect"
                /codon_start=1
                /db_xref="PID:d1005516"
                /db_xref="PID:g793765"
                /translation="MMKLVVPSRSLFNSFNHYSPLSHITLIRYNNSLFISNTHLSRR
                KLRVTFPIYTKRSLNIFYLLIPSCRTRLLMIYIYRNLMKHMSTVVRSHSHSIYRL
                RFSMRTNIIILRCHSYKPPFISHPYWNPSRMNLRGLLSRQSHLOPLIRFPLHLLIY
                RQFSRSPPLFPFRIRKQPNRILKCR"
            polyA_site
                1803
BASE COUNT  507 a    458 c    311 g    527 t
ORIGIN
1   tcagtttttt tttttttt tttttttt tttttttt tttttttt tttttttt ttgttctg
61   tccgtttaca tctgtaga tccaagcc cagccaac aattggctg cttagaat
121  cctccttggt gaaccagta tacttgcc atgaaccca gccacctatg gctagtagg
181  agaagctcaa ctgtaggcct gacttggaa gagaatgca atgctgtat cgaacttca
241  catggtggac ctctggccag agtcagcag ccgagggttc tttccgggc tgcctccca
301  ctgctgact ctgctcagt gctccatca tctggcggc cgttatgct attgctctc
361  cctctgtcac agcaatgac caacttagc gtagggagc agcctggg tctctaggc
421  gtttccatg ggcctgtg acaatccaa agatgagcc tccaaacc agaatcaga
481  ggcccagctc attgtaaaa acactctct gttggatga atggtacag ggcgtttca
541  gacaaagaa agctttctg tcactccat gagaacgct gcaatcact ttcgaaag
601  gaggagccca gaataacgt gtagggcat gacgatgc cggagagag cggagccca
661  ggaagcagaa agcaaaaba cacaccact attaaatt ataacact cactcttga
721  cctactgccc ccaccaca ttcatcatg atgaacttt ggtccctc taggattctg
781  ctaaatagtc caaatcatta caggttttt cttagccata cactaacat cagatacaat
841  aacagcttt toatcagaa caaacattg tcgagacta aatacggg gactaatcg
901  atatacagc gaagaaggag cctcaataa ttatttgc tcatctcc atgtcggagc
961  aggttatat tatgatcat atactttat agaaacctg acactggag taactctat
1021 gttcgcagc atagcaca gatttatag ctactctct ccatggagc aaatacaat
1081 aggtgtgcc acagtatta caaacctct atcagccat ccatatatt gaaacaacct
1141 agtcaatga atttggggg gctctcagt agacaagcc acctgacc gattctctc
1201 ttccacttc atcttaact ttatctcgc ggcctcaga atgttcaac tctctctct
1261 ccacgaaca ggtacaaca acccaacag attaaacta gatgcgata aaattcatt
1321 tcaccctac tatacaca agataccta ggtactca tcatcttt aattctca
1381 acctgatat tattttccc agacataca ggaagccag acaactac acctcgaat
1441 ccactaaca ccccaacca tattaaacc gaatgatatt tctatttc atagccatt
1501 ctacgtcaa tcccacaa actgaaggt gctcagct taactctac tatctaaatt
1561 ttacgctaa taccttctc tcatctcca agcaacgaa gctaatatt ccgcccaat
1621 acacaaatt tgaactgat cctagtacc acctacta tettaacct aattggggc
1681 caaccagat accaccatt attactcgt gccactagc ctcaatcca tactctcaa
1741 tcatctaat tctatacca atctcagaa ttatcgaga caaaataca aaattatc
1801 cat
```

## Header

- Title
- Taxonomy
- Citation

## Features (AA seq)

## DNA Sequence

# Types of files in GenBank

- From one-gene investigators
  - Often a very well annotated cDNA
  - A genomic segment from an new invertebrate
  - A mitochondria or virus
- From population/phylogenetic analysis
  - rRNA amplicon from environmental sampling
- From Genome Centers:
  - Gene expression:
    - Expressed Sequence Tags
    - Full Length Insert cDNA
  - Genome sequencing projects
    - WGS
    - HTG
    - CON

# UniProt

- New protein sequence database that is the result of a merge from SWISS-PROT and PIR. It will be **the** annotated curated protein sequence database.
- Data in UniProt is primarily derived from coding sequence annotations in EMBL (GenBank/DDDBJ) nucleic acid sequence data.
- UniProt is a Flat-File database just like EMBL and GenBank
- Flat-File format is SwissProt-like, or EMBL-like

# Swiss-Prot

ID CYS3\_YEAST STANDARD; PRT; 393 AA.  
AC P31373;  
DT 01-JUL-1993 (REL. 26, CREATED)  
DE CYSTATHIONINE GAMMA-LYASE (EC 4.4.1.1) (GAMMA-CYSTATHIONASE).  
GN CYS3 OR CY11 OR STR1 OR YAL012W OR FUN35.  
OS TAXONOMY  
OC SACCHAROMYCETACEAE; SACCHAROMYCES.  
RX CITATION  
CC -- CATALYTIC ACTIVITY: L-CYSTATHIONINE + H(2)O = L-CYSTEINE +  
CC NH(3) + 2-OXOBUTANOATE.  
CC -- COFACTOR: PYRIDOXAL PHOSPHATE.  
CC -- PATHWAY: FINAL STEP IN THE TRANS-SULFURATION PATHWAY SYNTHESIZING  
CC L-CYSTEINE FROM L-METHIONINE.  
CC -- SUBUNIT: HOMOTETRAMER.  
CC -- SUBCELLULAR LOCATION: CYTOPLASMIC.  
CC -- SIMILARITY: BELONGS TO THE TRANS-SULFURATION ENZYMES FAMILY.

## DISCLAMOR

DR DATABASE cross-reference

KW CYSTEINE BIOSYNTHESIS; LYASE; PYRIDOXAL PHOSPHATE.

FT INIT\_MET 0 0

FT BINDING 203 203 PYRIDOXAL PHOSPHATE (BY SIMILARITY).

SQ SEQUENCE 393 AA; 42411 MW; 55BA2771 CRC32;

TLQESDKFAT KAIHAGEHVD VHGSVIEPIS LSTTFKQSSP ANPIGTYEYS RSQNPENRENL  
ERAAVALENA QYGLAFSSGS ATTATILQSL PQGSHAVSIG DVYGGTHRYF TKVANAHGVE  
TSFTNDLLND LPQLIKENTK LVWIETPTNP TLKVTDIQKV ADLIKKHAAG QDVILVVDNT  
FLSPYISNPL NFGADIVVHS ATKYINGHSD VVLGVLATNN KPLYERLQFL QNAIGAIPSP  
FDAWLTHRGL KTLHLRVRQA ALSANKIAEF LAADKENVVA VNYPLKTHP NYDVVLKQHR  
DALGGGMISF RIKGGAEAS KFASTRFLT LAESLGGIES LLEVPVMTH GGIPKEAREA  
SGVFDDLVRV SVGIEDTDDL LEDIKQALKQ ATN

//

ID CYS3\_YEAST STANDARD; PRT; 393 AA.  
AC P31373;  
DT 01-JUL-1993 (REL. 26, CREATED)  
DT 01-JUL-1993 (REL. 26, LAST SEQUENCE UPDATE)  
DT 01-NOV-1995 (REL. 32, LAST ANNOTATION UPDATE)  
DE CYSTATHIONINE GAMMA-LYASE (EC 4.4.1.1) (GAMMA-CYSTATHIONASE).  
GN CYS3 OR CY11 OR STR1 OR YAL012W OR FUN35.  
OS SACCHAROMYCES CEREVISIAE (BAKER'S YEAST).  
OC RHIZARIOTA; FUNGI; ASCOMYCOTA; HEMIASCOMYCETES; SACCHAROMYCETALES;  
OC SACCHAROMYCETACEAE; SACCHAROMYCES.  
RN [1]  
RP SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.  
RX MEDLINE: 92250430. [NCBI, EXPASY, Israel, Japan]  
RA ONO B.-I., TANAKA K., NAITO K., HEIKE C., SHINODA S., YAMAMOTO S.,  
RA OHMORI S., OSHIMA T., TOH-E A.;  
RT "Cloning and characterization of the CYS3 (CY11) gene of  
RT Saccharomyces cerevisiae.";  
RL J. BACTERIOL. 174:3339-3347(1992).  
RN [2]  
RP SEQUENCE FROM N.A., AND CHARACTERIZATION.  
RX STRAIN-DBY939;  
RX MEDLINE: 9328665. [NCBI, EXPASY, Israel, Japan]  
RA YAMAGATA S., D'ANDREA R.J., FUJISAKI S., ISAJI M., NAKAMURA K.;  
RT "Cloning and bacterial expression of the CYS3 gene encoding  
RT cystathionine gamma-lyase of Saccharomyces cerevisiae and the  
RT physicochemical and enzymatic properties of the protein.";  
RL J. BACTERIOL. 175:4800-4808(1993).  
RN [3]  
RP SEQUENCE FROM N.A.  
RX STRAIN-S288C / AB972;  
RX MEDLINE: 93289814. [NCBI, EXPASY, Israel, Japan]  
RA BARTON A.B., KABACK D.B., CLARK M.W., KENG T., OUELLETTE B.F.F.,  
RA STORMS R.K., ZENG B., ZHONG W.W., FORTIN N., DELANEY S., BUSSEY H.;  
RT "Physical localization of yeast CYS3, a gene whose product resembles  
RT the rat gamma-cystathionase and Escherichia coli cystathionine gamma-  
RT synthase enzymes.";  
RL YEAST 9:363-369(1993).  
RN [4]  
RP SEQUENCE FROM N.A.  
RX STRAIN-S288C / AB972;  
RX MEDLINE: 93209532. [NCBI, EXPASY, Israel, Japan]  
RA OUELLETTE B.F.F., CLARK M.W., KENG T., STORMS R.K., ZHONG W.W.,  
RA ZENG B., FORTIN N., DELANEY S., BARTON A.B., KABACK D.B., BUSSEY H.;  
RT "Sequencing of chromosome I from Saccharomyces cerevisiae: analysis  
RT of a 32 kb region between the LTR1 and SPO7 genes.";  
RL GENOME 36:32-42(1993).  
RN [5]  
RP SEQUENCE OF 1-18, AND CHARACTERIZATION.  
RX MEDLINE: 93289817. [NCBI, EXPASY, Israel, Japan]  
RA ONO B.-I., ISHII N., NAITO K., MIYOSHI S.-I., SHINODA S., YAMAMOTO S.,  
RA OHMORI S.;  
RT "Cystathionine gamma-lyase of Saccharomyces cerevisiae: structural  
RT gene and cystathionine gamma-synthase activity.";  
RL YEAST 9:389-397(1993).  
CC -- CATALYTIC ACTIVITY: L-CYSTATHIONINE + H(2)O = L-CYSTEINE +  
CC NH(3) + 2-OXOBUTANOATE.  
CC -- COFACTOR: PYRIDOXAL PHOSPHATE.  
CC -- PATHWAY: FINAL STEP IN THE TRANS-SULFURATION PATHWAY SYNTHESIZING  
CC L-CYSTEINE FROM L-METHIONINE.  
CC -- SUBUNIT: HOMOTETRAMER.  
CC -- SUBCELLULAR LOCATION: CYTOPLASMIC.  
CC -- SIMILARITY: BELONGS TO THE TRANS-SULFURATION ENZYMES 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](mailto:license@isb-sib.ch).)  
CC -----  
DR EMBL: L05146; AAC04945.1; -. [EMBL / GenBank / DDBJ] [CodingSequence]  
DR EMBL: L04459; AA05217.1; -. [EMBL / GenBank / DDBJ] [CodingSequence]  
DR EMBL: D14143; BA03190.1; -. [EMBL / GenBank / DDBJ] [CodingSequence]  
DR PIR: S31228; S31228.  
DR YEPD: 5280; -.  
DR SGD: L0000470; CYS3. [SGD / YPD]  
DR PFAM: PF01053; Cys\_Met\_Meta\_PP\_1.  
DR PROSITE: PS00868; CYS\_MET\_TAB\_PP\_1.  
DR DOMO: P31373.  
DR PRODOM [Domain structure / List of seq. sharing at least 1 domain]  
DR PROTOPA: P31373.  
DR PIRNAME: P31373.  
DR SWISS-2DPAGE; GET REGION ON 2D PAGE.  
KW CYSTEINE BIOSYNTHESIS; LYASE; PYRIDOXAL PHOSPHATE.  
FT INIT\_MET 0 0  
FT BINDING 203 203 PYRIDOXAL PHOSPHATE (BY SIMILARITY).  
SQ SEQUENCE 393 AA; 42411 MW; 55BA2771 CRC32;  
TLQESDKFAT KAIHAGEHVD VHGSVIEPIS LSTTFKQSSP ANPIGTYEYS RSQNPENRENL  
ERAAVALENA QYGLAFSSGS ATTATILQSL PQGSHAVSIG DVYGGTHRYF TKVANAHGVE  
TSFTNDLLND LPQLIKENTK LVWIETPTNP TLKVTDIQKV ADLIKKHAAG QDVILVVDNT  
FLSPYISNPL NFGADIVVHS ATKYINGHSD VVLGVLATNN KPLYERLQFL QNAIGAIPSP  
FDAWLTHRGL KTLHLRVRQA ALSANKIAEF LAADKENVVA VNYPLKTHP NYDVVLKQHR  
DALGGGMISF RIKGGAEAS KFASTRFLT LAESLGGIES LLEVPVMTH GGIPKEAREA  
SGVFDDLVRV SVGIEDTDDL LEDIKQALKQ ATN

//

# Swiss-Prot

NiceProt View of SWISS-PROT: P31373 - Netscape

File Edit View Go Communicator Help

[ExPASy Home page](#) [Site Map](#) [Search ExPASy](#) [Contact us](#) [SWISS-PROT](#)

Mirror sites: [Taiwan](#) [Australia](#)

## NiceProt View of SWISS-PROT: P31373

General information about the entry	
Entry name	CYS3_YEAST
Primary accession number	P31373
Secondary accession number(s)	None
Entered in SWISS-PROT in	Release 26, July 1993
Sequence was last modified in	Release 26, July 1993
Annotations were last modified in	Release 32, November 1995
Name and origin of the protein	
Protein name	CYSTATHIONINE GAMMA-LYASE
Synonym(s)	<a href="#">EC 4.4.1.1</a> GAMMA-CYSTATHIONASE
Gene name(s)	CYS3 OR CYI1 OR STR1 OR YAL012W OR FUN35
From	<a href="#">SACCHAROMYCES CEREVISIAE (BAKER'S YEAST)</a>
Taxonomy	EUKARYOTA; FUNGI; ASCOMYCOTA; HEMIASCOMYCETES; SACCHAROMYCETALES; SACCHAROMYCETACEAE; SACCHAROMYCES.

Document: Done

# Swiss-Prot

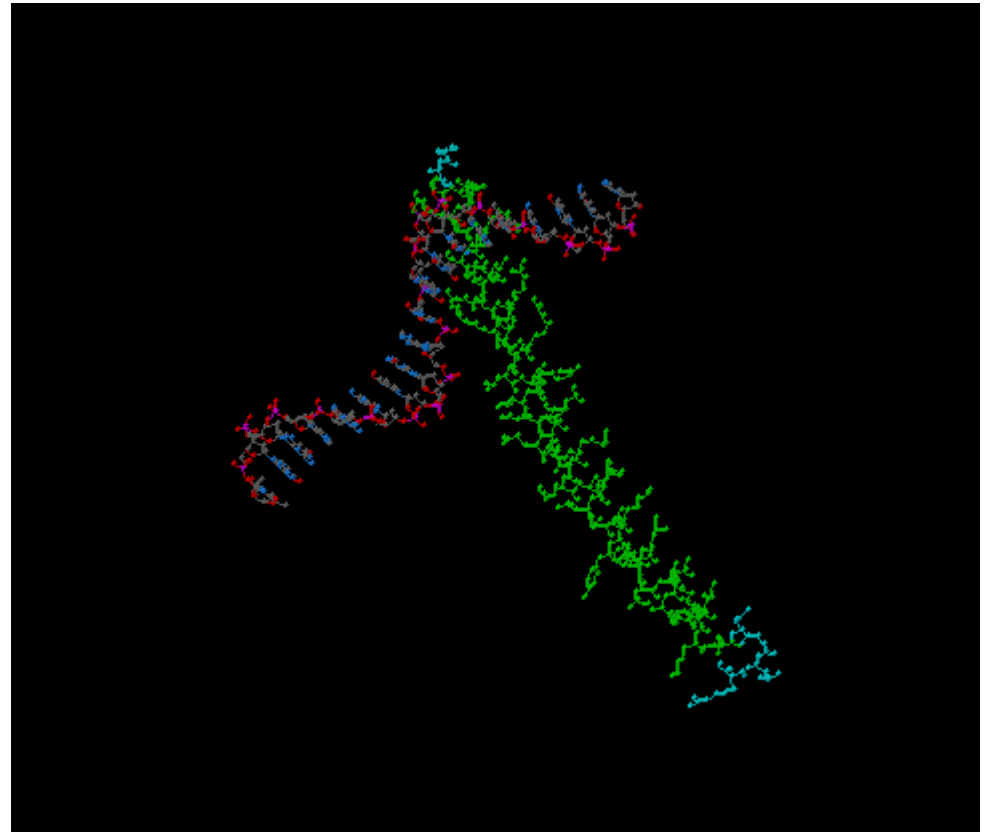
- SWISS-PROT incorporates:
  - Function of the protein
  - Post-translational modification
  - Domains and sites.
  - Secondary structure.
  - Quaternary structure.
  - Similarities to other proteins;
  - Diseases associated with deficiencies in the protein
  - Sequence conflicts, variants, etc.

# TREMBL

- TrEMBL is a computer-annotated protein sequence database supplementing the SWISS-PROT Protein Sequence Data Bank.
- TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL Nucleotide Sequence Database not yet integrated in SWISS-PROT.
- TrEMBL can be considered as a preliminary section of SWISS-PROT. For all TrEMBL entries which should finally be upgraded to the standard SWISS-PROT quality, SWISS-PROT accession numbers have been assigned.

# PDB

- Protein DataBase
  - Protein and NA 3D structures
  - Sequence present
  - YAFFF



# PDB

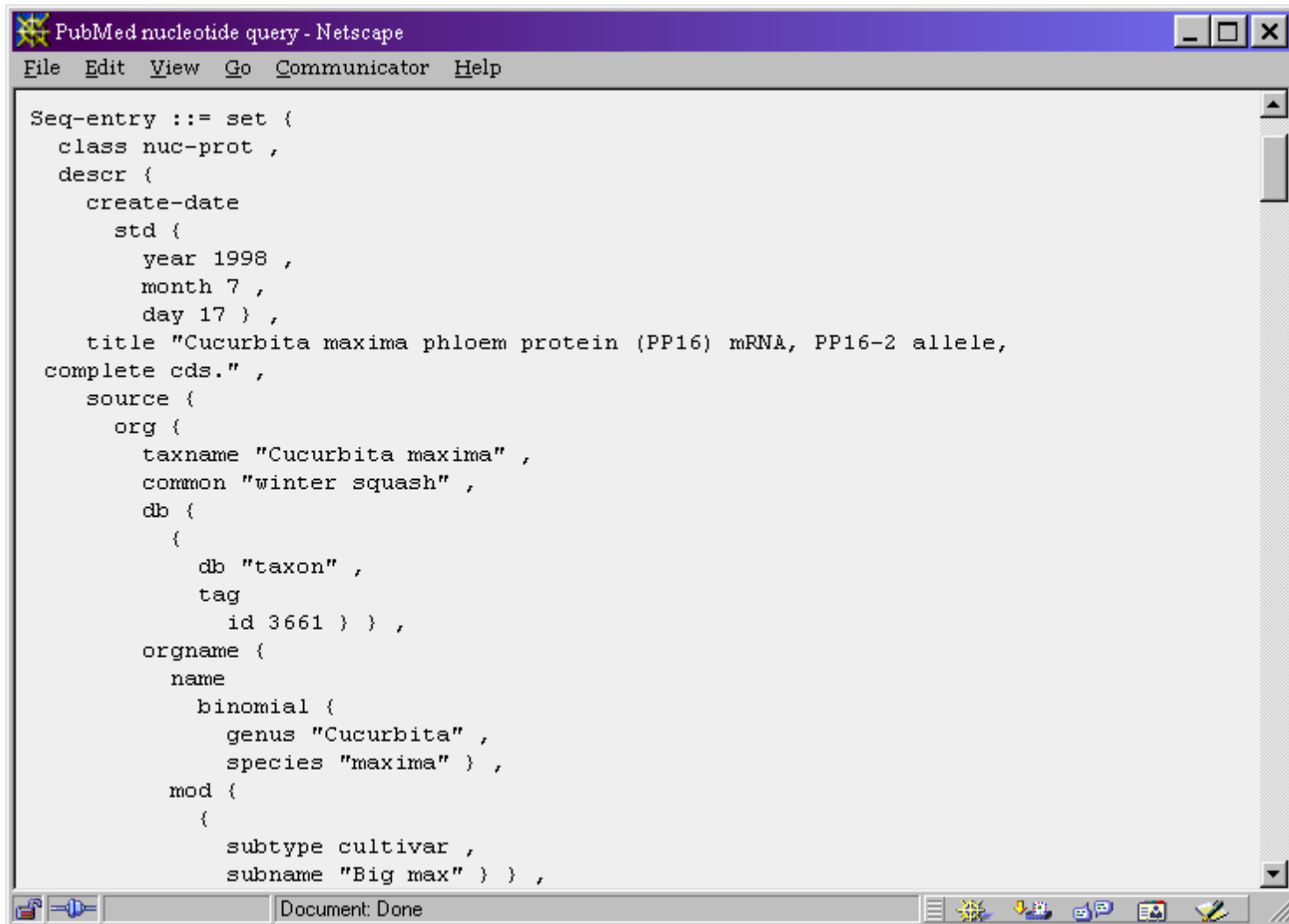
- HEADER
- COMPND
- SOURCE
- AUTHOR
- DATE
- JRNL
- REMARK
- SECRES
- ATOM COORDINATES

```
HEADER LEUCINE ZIPPER 15-JUL-93 IDGC IDGC 2
COMPND GCN4 LEUCINE ZIPPER COMPLEXED WITH SPECIFIC IDGC 3
COMPND 2 ATF/CREB SITE DNA IDGC 4
SOURCE GCN4: YEAST (SACCHAROMYCES CEREVISIAE); DNA: SYNTHETIC IDGC 5
AUTHOR T.J.RICHMOND IDGC 6
REVDAT 1 22-JUN-94 LDGC 0 IDGC 7
JRNL AUTH P.KONIG,T.J.RICHMOND IDGC 8
JRNL TITL THE X-RAY STRUCTURE OF THE GCN4-BZIP BOUND TO IDGC 9
JRNL TITL 2 ATF/CREB SITE DNA SHOWS THE COMPLEX DEPENDS ON DNA IDGC 10
JRNL TITL 3 FLEXIBILITY IDGC 11
JRNL REF J.MOL.BIOL. V. 233 139 1993 IDGC 12
JRNL REFN ASTM JMOBAK UK ISSN 0022-2836 0070 IDGC 13
REMARK 1 IDGC 14
REMARK 2 IDGC 15
REMARK 2 RESOLUTION. 3.0 ANGSTROMS. IDGC 16
REMARK 3 IDGC 17
REMARK 3 REFINEMENT. IDGC 18
REMARK 3 PROGRAM X-PLOR IDGC 19
REMARK 3 AUTHORS BRUNGER IDGC 20
REMARK 3 R VALUE 0.216 IDGC 21
REMARK 3 RMSD BOND DISTANCES 0.020 ANGSTROMS IDGC 22
REMARK 3 RMSD BOND ANGLES 3.86 DEGREES IDGC 23
REMARK 3 IDGC 24
REMARK 3 NUMBER OF REFLECTIONS 3296 IDGC 25
REMARK 3 RESOLUTION RANGE 10.0 - 3.0 ANGSTROMS IDGC 26
REMARK 3 DATA CUTOFF 3.0 SIGMA(F) IDGC 27
REMARK 3 PERCENT COMPLETION 98.2 IDGC 28
REMARK 3 IDGC 29
REMARK 3 NUMBER OF PROTEIN ATOMS 456 IDGC 30
REMARK 3 NUMBER OF NUCLEIC ACID ATOMS 386 IDGC 31
REMARK 4 IDGC 32
REMARK 4 GCN4: TRANSCRIPTIONAL ACTIVATOR OF GENES ENCODING FOR AMINO IDGC 33
REMARK 4 ACID BIOSYNTHETIC ENZYMES. IDGC 34
REMARK 5 IDGC 35
REMARK 5 AMINO ACIDS NUMBERING (RESIDUE NUMBER) CORRESPONDS TO THE IDGC 36
REMARK 5 281 AMINO ACIDS OF INTACT GCN4. IDGC 37
REMARK 6 IDGC 38
REMARK 6 BZIP SEQUENCE 220 - 281 USED FOR CRYSTALLIZATION. IDGC 39
REMARK 7 IDGC 40
REMARK 7 MODEL FROM AMINO ACIDS 227 - 281 SINCE AMINO ACIDS 220 - IDGC 41
REMARK 7 226 ARE NOT WELL ORDERED. IDGC 42
REMARK 8 IDGC 43
REMARK 8 RESIDUE NUMBERING OF NUCLEOTIDES: IDGC 44
REMARK 8 5' T G G A G A T G A C G T C A T C T C C IDGC 45
REMARK 8 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 1 2 3 4 5 6 7 8 9 IDGC 46
REMARK 9 IDGC 47
REMARK 9 THE ASYMMETRIC UNIT CONTAINS ONE HALF OF PROTEIN/DNA IDGC 48
REMARK 9 COMPLEX PER ASYMMETRIC UNIT. IDGC 49
REMARK 10 IDGC 50
REMARK 10 MOLECULAR DYAD AXIS OF PROTEIN DIMER AND PALINDROMIC HALF IDGC 51
REMARK 10 SITES OF THE DNA COINCIDES WITH CRYSTALLOGRAPHIC TWO-FOLD IDGC 52
REMARK 10 AXIS. THE FULL PROTEIN/DNA COMPLEX CAN BE OBTAINED BY IDGC 53
REMARK 10 APPLYING THE FOLLOWING TRANSFORMATION MATRIX AND IDGC 54
REMARK 10 TRANSLATION VECTOR TO THE COORDINATES X Y Z: IDGC 55
REMARK 10 IDGC 56
REMARK 10 0 -1 0 X 117.32 X SYMM IDGC 57
REMARK 10 -1 0 0 Y + 117.32 = Y SYMM IDGC 58
REMARK 10 0 0 -1 Z 43.33 Z SYMM IDGC 59
SEQRES 1 A 62 ILE VAL PRO GLU SER SER ASP PRO ALA ALA LEU LYS ARG IDGC 60
SEQRES 2 A 62 ALA ARG ASN THR GLU ALA ALA ARG ARG SER ARG ALA ARG IDGC 61
SEQRES 3 A 62 LYS LEU GLN ARG MET LYS GLN LEU GLU ASP LYS VAL GLU IDGC 62
SEQRES 4 A 62 GLU LEU LEU SER LYS ASN TYR HIS LEU GLU ASN GLU VAL IDGC 63
SEQRES 5 A 62 ALA ARG LEU LYS LYS LEU VAL GLY GLU ARG IDGC 64
SEQRES 1 B 19 T G G A G A T G A C G T C IDGC 65
SEQRES 2 B 19 A T C T C IDGC 66
HELIX 1 A ALA A 228 LYS A 276 1 IDGC 67
CRYST1 58.660 58.660 86.660 90.00 90.00 90.00 P 41 21 2 8 IDGC 68
ORIGX1 1.000000 0.000000 0.000000 0.00000 IDGC 69
ORIGX2 0.000000 1.000000 0.000000 0.00000 IDGC 70
ORIGX3 0.000000 0.000000 1.000000 0.00000 IDGC 71
SCALE1 0.017047 0.000000 0.000000 0.00000 IDGC 72
SCALE2 0.000000 0.017047 0.000000 0.00000 IDGC 73
SCALE3 0.000000 0.000000 0.011539 0.00000 IDGC 74
ATOM 1 N PRO A 227 35.313 108.011 15.140 1.00 38.94 IDGC 75
ATOM 2 CA PRO A 227 34.172 107.658 15.972 1.00 39.82 IDGC 76
ATOM 842 C5 C B 9 57.692 100.286 22.744 1.00 29.82 IDGC 916
ATOM 843 C6 C B 9 58.128 100.193 21.465 1.00 30.65 IDGC 917
TER 844 C B 9 IDGC 918
MASTER 46 0 0 1 0 0 0 6 842 2 0 7 IDGC 919
END IDGC 920
```

# Format

- ASN.1
- Flat Files
  - DNA
  - Protein
- FASTA
  - DNA
  - Protein

# Abstract Syntax Notation (ASN.1)



The screenshot shows a Netscape browser window titled "PubMed nucleotide query - Netscape". The main content area displays an ASN.1 sequence entry for a Cucurbita maxima protein. The entry is structured as follows:

```
Seq-entry ::= set {
  class nuc-prot ,
  descr {
    create-date
    std {
      year 1998 ,
      month 7 ,
      day 17 } ,
    title "Cucurbita maxima phloem protein (PP16) mRNA, PP16-2 allele,
complete cds." ,
    source {
      org {
        taxname "Cucurbita maxima" ,
        common "winter squash" ,
        db {
          {
            db "taxon" ,
            tag
            id 3661 } } ,
        orgname {
          name
          binomial {
            genus "Cucurbita" ,
            species "maxima" } ,
          mod {
            {
              subtype cultivar ,
              subname "Big max" } } ,
```

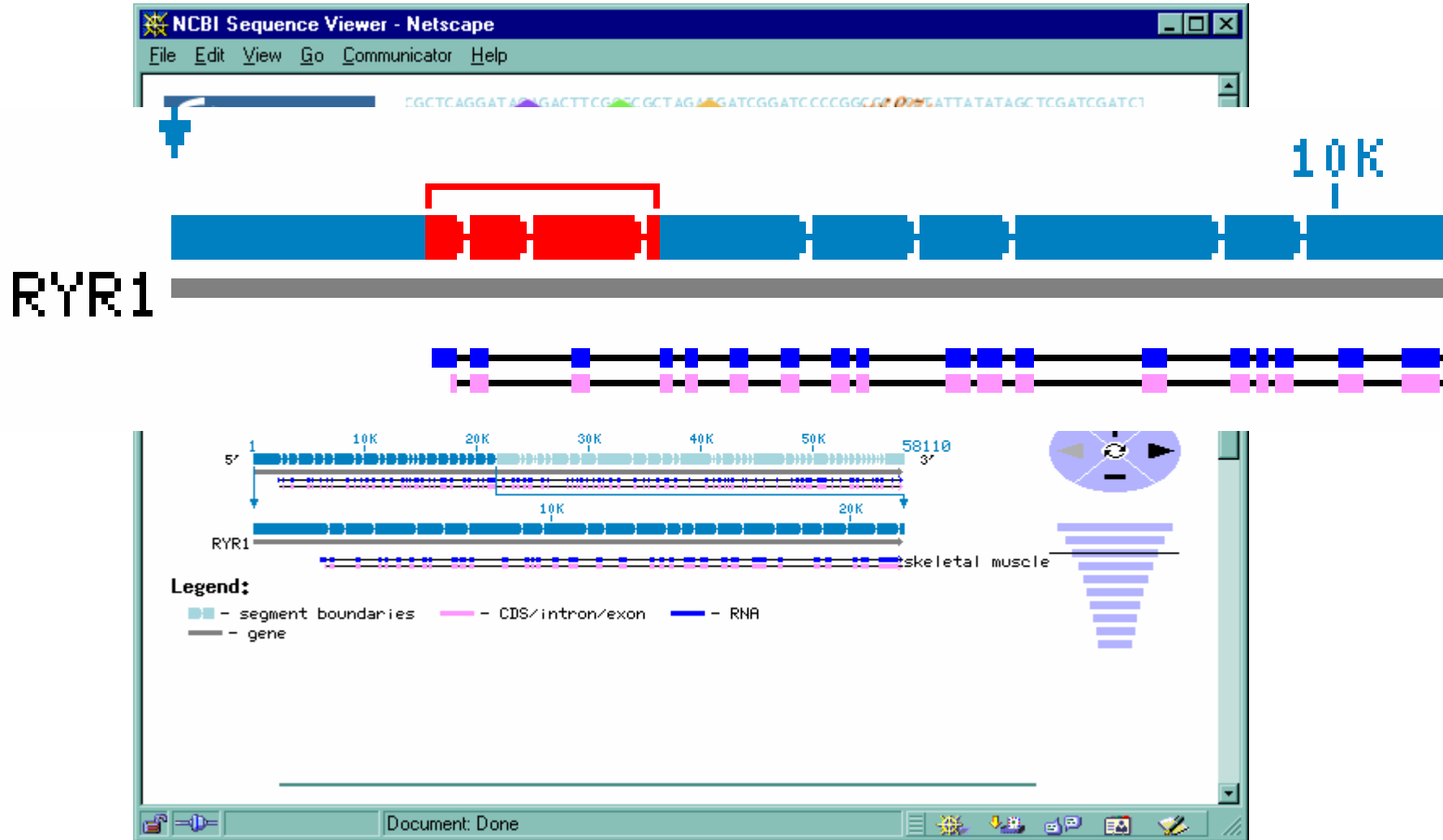
The browser's status bar at the bottom indicates "Document: Done".

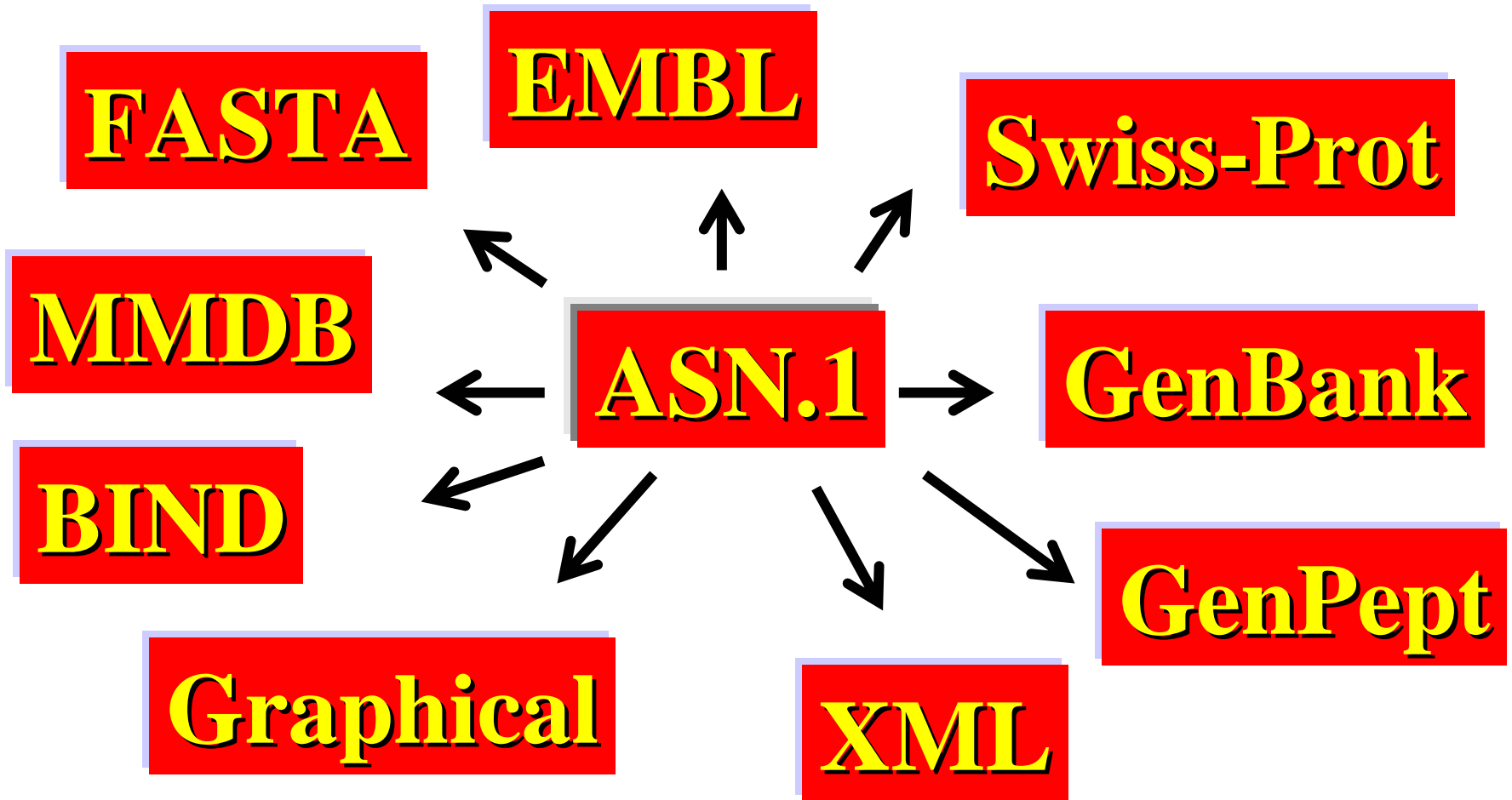
# FASTA

>

```
MSEYQPSL FALNPMGF SPLDGSKSTNENV SASTSTAKPMVGQLIFDKFIKTEEDPI  
IKQDTPSNLDFDFALPQTATAPDAKTVLPIPELDDAVVESFFSSSTDSTPMFEYEN  
LEDNSKEWTSLFDNDIPVTTDDVSLADKAIESTEEVSLVPSNLEVSTTSFLPTPVL  
EDAKLTQTRKVKKPNSVVKKSHHVGKDDERLDHLGVVAYNRKQRSIPLSPIVPES  
SDPAALKRARNTAARRSRARKLQRMKQLEDKVEELLSKNYHLENEVARLKKLVGE  
R
```

# Graphical Representation





# Organismal Divisions

Used in which database?

BCT	Bacterial	DDBJ - GenBank
FUN	Fungal	EMBL
HUM	Homo sapiens	DDBJ - EMBL
INV	Invertebrate	all
MAM	Other mammalian	all
ORG	Organelle	EMBL
PHG	Phage	all
PLN	Plant	all
PRI	Primate (also see HUM)	all (not same data in all)
PRO	Prokaryotic	EMBL
ROD	Rodent	all
SYN	Synthetic and chimeric	all
VRL	Viral	all
VRT	Other vertebrate	all

# Functional Divisions

**PAT** Patent

**EST** Expressed Sequence Tags

**STS** Sequence Tagged Site

**GSS** Genome Survey Sequence

**HTG** High Throughput Genome (unfinished)

**HTC** High throughput cDNA (unfinished)

**CON** Contig assembly instructions

Organismal divisions:

**BCT** **FUN** **INV** **MAM** **PHG** **PLN**

**PRI** **ROD** **SYN** **VRL** **VRT**

# Guiding Principals

In GenBank, records are grouped for various reasons: understand this is key to using and fully taking advantage of this database.

# Identifiers

- You need identifiers which are stable through time
- Need identifiers which will always refer to specific sequences
- Need these identifiers to track history of sequence updates
- Also need feature and annotation identifiers

# LOCUS, Accession, NID and protein\_id

**LOCUS:** Unique string of 10 letters and numbers in the database. Not maintained amongst databases, and is therefore a poor sequence identifier.

**ACCESSION:** A unique identifier to that record, citable entity; does not change when record is updated. A good record identifier, ideal for citation in publication.

**VERSION:** : New system where the accession and version play the same function as the accession and gi number.

**Nucleotide gi:** Geninfo identifier (gi), a unique integer which will change every time the sequence changes.

**PID:** Protein Identifier: g, e or d prefix to gi number. Can have one or two on one CDS.

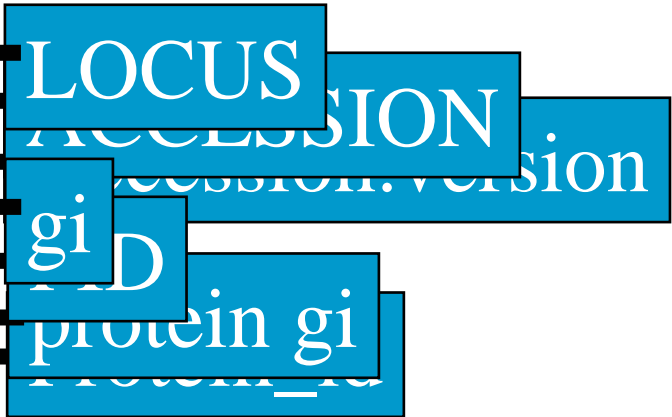
**Protein gi:** Geninfo identifier (gi), a unique integer which will change every time the sequence changes.

**protein\_id:** Identifier which has the same structure and function as the nucleotide Accession.version numbers, but slightly different format.

# LOCUS, Accession, gi and PID

LOCUS	HSU40282	1789 bp	mRNA	PRI	21-MAY-1998
DEFINITION	Homo sapiens integrin-linked kinase (ILK) mRNA, complete cds.				
ACCESSION	U40282				
VERSION	U40282.1 GI:3150001				

LOCUS: HSU40282  
ACCESSION: U40282  
VERSION: U40282.1  
GI: 3150001  
PID: g3150002  
Protein gi: 3150002  
protein\_id: AAC16892.1



```
CDS          157..1515
              /gene="ILK"
              /note="protein serine/threonine kinase"
              /codon_start=1
              /product="integrin-linked kinase"
              /protein_id="AAC16892.1"
              /db_xref="PID:g3150002"
              /db_xref="GI:3150002"
```

# EST: Expressed Sequence Tag

Expressed Sequence Tags are short (300-500 bp) single reads from mRNA (cDNA) which are produced in large numbers. They represent a snapshot of what is expressed in a given tissue, and developmental stage.

Also see: <http://www.ncbi.nlm.nih.gov/dbEST/>  
<http://www.ncbi.nlm.nih.gov/UniGene/>

# STS

Sequenced Tagged Sites, are operationally unique sequence that identifies the combination of primer pairs used in a PCR assay that generate a mapping reagent which maps to a single position within the genome.

Also see: <http://www.ncbi.nlm.nih.gov/dbSTS/>

<http://www.ncbi.nlm.nih.gov/genemap/>

# GSS: Genome Survey Sequences

Genome Survey Sequences are similar in nature to the ESTs, except that its sequences are genomic in origin, rather than cDNA (mRNA).

The GSS division contains:

- random "single pass read" genome survey sequences.
- single pass reads from cosmid/BAC/YAC ends (these could be chromosome specific, but need not be)
- exon trapped genomic sequences
- Alu PCR sequences

Also see: <http://www.ncbi.nlm.nih.gov/dbGSS/>

# HTG: High Throughput Genome

High Throughput Genome Sequences are unfinished genome sequencing efforts records. Unfinished records have gaps in the nucleotides sequence, low accuracy, and no annotations on the records.

Also see: <http://www.ncbi.nlm.nih.gov/HTGS/>

Ouellette and Boguski (1997) *Genome Res.* 7:952-955

# HTGS in GenBank

phase 0 → ← ← → → HTG  
Acc = AC000003 gi = 1235673

phase 1 → ← → → ← HTG  
Acc = AC000003 gi = 1556454

phase 2 → → → → HTG  
Acc = AC000003 gi = 2182283

phase 3 → → → → PRI  
Acc = AC000003 gi = 2204282

# HTGS in GenBank

- Unfinished Record
  - Sequencing will be unfinished
  - Phase 1 or phase 2
  - HTG division
  - **KEYWORDS: HTG; HTGS\_PHASE1 or 2**
- Finished record
  - Sequencing will be finished
  - Phase 3
  - Organismal division it belongs to **PRI, INV or PLN**
  - **KEYWORDS: HTG**

# HTC in GenBank

- GenBank division for unfinished high-throughput cDNA sequencing (HTC).
- HTC sequences may have 5'UTR and 3'UTR at their ends, partial coding regions, and **introns**.
- A keyword of "HTC" will be present, in addition to division code "HTC". Those HTC sequences that undergo finishing (eg, re-sequencing) will move to the appropriate taxonomic GenBank division and the "HTC" keyword will be removed.

# Top 5 organisms in the HTC division

64106	<i>Mus musculus</i>
62848	<i>Anopheles gambiae</i>
9119	<i>Zea mays</i>
7732	<i>Homo sapiens</i>
2957	<i>Schmidtea mediterranea</i>

# WGS in GenBank

- Contigs from ongoing Whole Genome Shotgun sequencing projects
- The nucleotides from WGS projects go into the BLAST 'wgs' database, whereas the proteins go into the BLAST nr database.
- More info, and how to submit to this division:  
<http://www.ncbi.nlm.nih.gov/Genbank/wgs.html>
- Accession format is 4+2+6

# CON in GenBank

- Points to files that make the contig, does not actually contain sequence
- ‘Invented’ by NCBI to deal with tracking of segmented sets and 350 KB limit in DDBJ/EMBL/GenBank

# CON in GenBank

```
LOCUS      AH007743      7832 bp      DNA                      CON      26-MAY-1999
DEFINITION Gallus gallus ornithine transcarbamylase (OTC) gene, complete cds.
ACCESSION  AH007743
VERSION    AH007743.1  GI:4927367
KEYWORDS   .
SOURCE     chicken.
  ORGANISM Gallus gallus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Archosauria;
            Aves; Neognathae; Galliformes; Phasianidae; Phasianinae; Gallus.

[....]
FEATURES             Location/Qualifiers
     source           1..7832
                     /organism="Gallus gallus"
                     /db_xref="taxon:9031"
                     /chromosome="1"
CONTIG             join(AF065630.1:1..1903,gap(),AF065631.1:1..435,gap(),
AF065632.1:1..509,gap(),AF065633.1:1..722,gap(),AF065634.1:1..707,
gap(),AF065635.1:1..836,gap(),AF065636.1:1..1614,gap(),
AF065637.1:1..605,gap(),AF065638.1:1..501)

//
```

```
join(AF065630.1:1..1903,  
     gap(),  
     AF065631.1:1..435,  
     gap(),  
     AF065632.1:1..509,  
     gap(),  
     AF065633.1:1..722,  
     gap(),  
     AF065634.1:1..707,  
     ...
```

# Sequences NOT in GenBank

- SNPs
- SAGE tags
- RefSeq (Genomic, mRNA, or protein)
- Consensus sequences

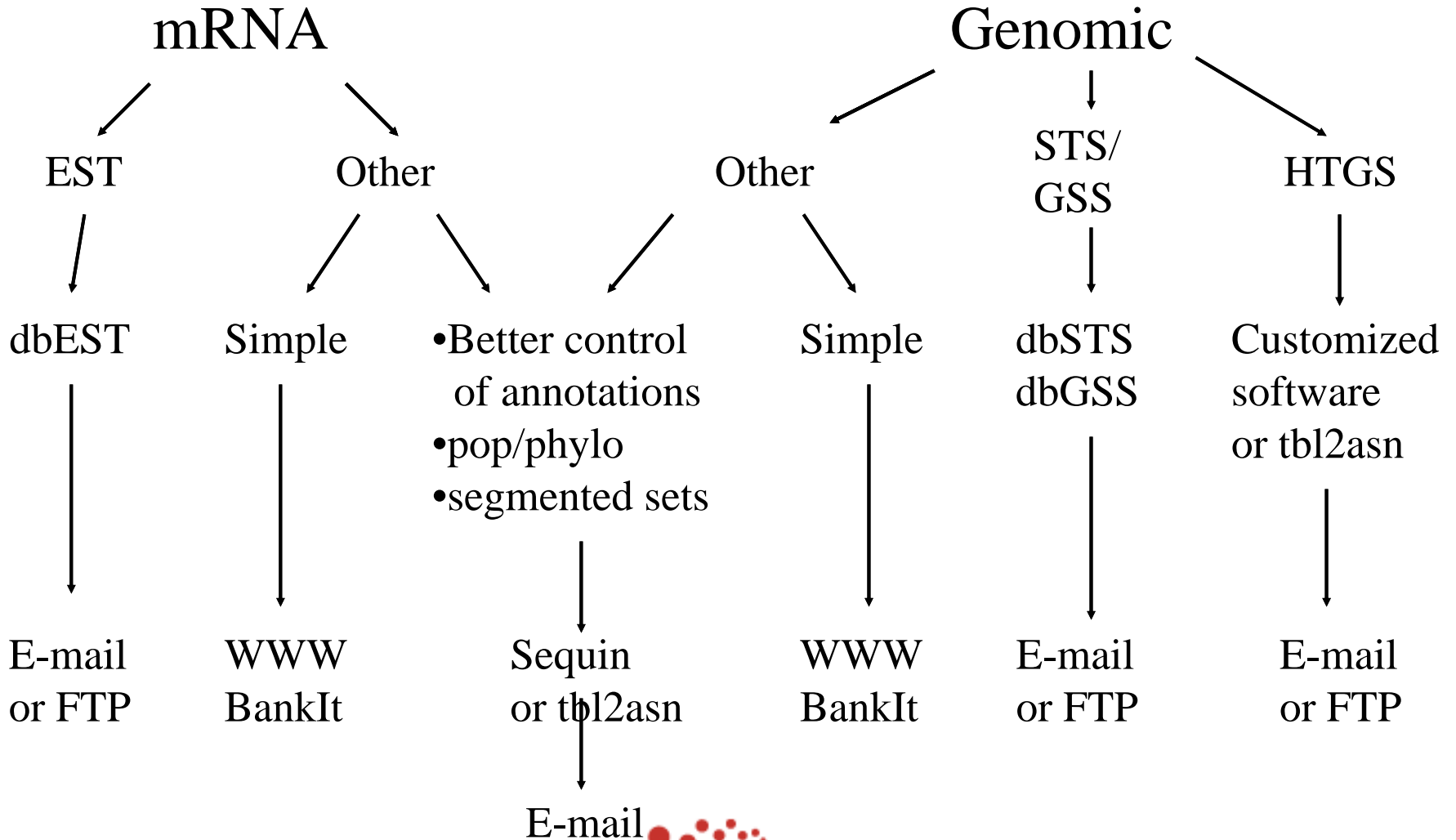
# Sequences to Public Databases

- No longer publish sequences in Journal
- Electronic format , is most useful
- Allows validations testing of data
- best way to move Science forward
- Sequences sent to DDBJ/EMBL/GenBank are exchanged daily
- Best way to exchange new data, and updates

# Which Tool?

- **BankIt**: Web based tool which is simple, easy to use, great for simple submissions, but not ideal for complicated ones.
  - Sakura (DDBJ)
  - WebIn (EMBL)
- **Sequin**: Client that you need to d/l to your computer, a little harder to learn, but has great documentation, and ideal for complicated, large, multiple submissions.
- **tbl2asn**: ideal for batch records, command line, scriptable, can work with sequin

# Which tool?



# In closing ...

- Often only use FASTA files (eg for BLAST)
- GBFF are simply human readable versions of these records
- GBFF have become a vehicle for a lot more information than they were meant to do
- Keep in mind that GenBank is DNA centric and is a poor vehicle for protein and mRNA expression/interaction information

# In closing (cont'd) ...

- Able to recognize various data formats, and know what their primary use is.
- Know, understand and utilize all types of sequence identifiers.
- Know and understand various feature types present in the GenBank flat files.
- Know and understand the various GenBank divisions.

## In closing (cont'd) ...

- Open access to sequences is not only essential for all of the work we do, if it was not there, there would be no bioinformatics, no BLAST, no CBW
- As critical as open access to sequence information is the open access to the literature.

# Closing (part4)

- I urge you all to only publish in open access journals
- I urge you to convince your colleagues and mentors to do the same
- PLoS Biology, BMC genetics, Genome Biology and so forth – great journals!
- More journals are going open access: be part of what caused this wave!

# Resources

- W W W:
  - <http://www.ncbi.nlm.nih.gov>
  - <http://www.ddbj.nig.ac.jp/>
  - <http://www.ebi.ac.uk/>
  - <http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html>
  - <http://www.ebi.ac.uk/embl/>
  - <http://www.pir.uniprot.org/>
  - <http://www.expasy.ch/sprot/>
  - <http://www.rcsb.org/pdb/>
  - <http://www.ncbi.nlm.nih.gov/Genbank/> (submission info)
  - <http://genome-www.stanford.edu/Saccharomyces/>

# Resources

- W W W:

- <http://nar.oupjournals.org/content/vol30/issue1/>
- <http://nar.oupjournals.org/content/vol31/issue1/>
- <http://www.ncbi.nlm.nih.gov/HTGS/>
- <http://www.ncbi.nlm.nih.gov/dbEST/>
- <http://www.ncbi.nlm.nih.gov/Genbank/wgs.html>
- <http://www.ncbi.nlm.nih.gov/dbSTS/>
- <http://www.ncbi.nlm.nih.gov/dbGSS/>
- <http://www.ncbi.nlm.nih.gov/genome/guide/>