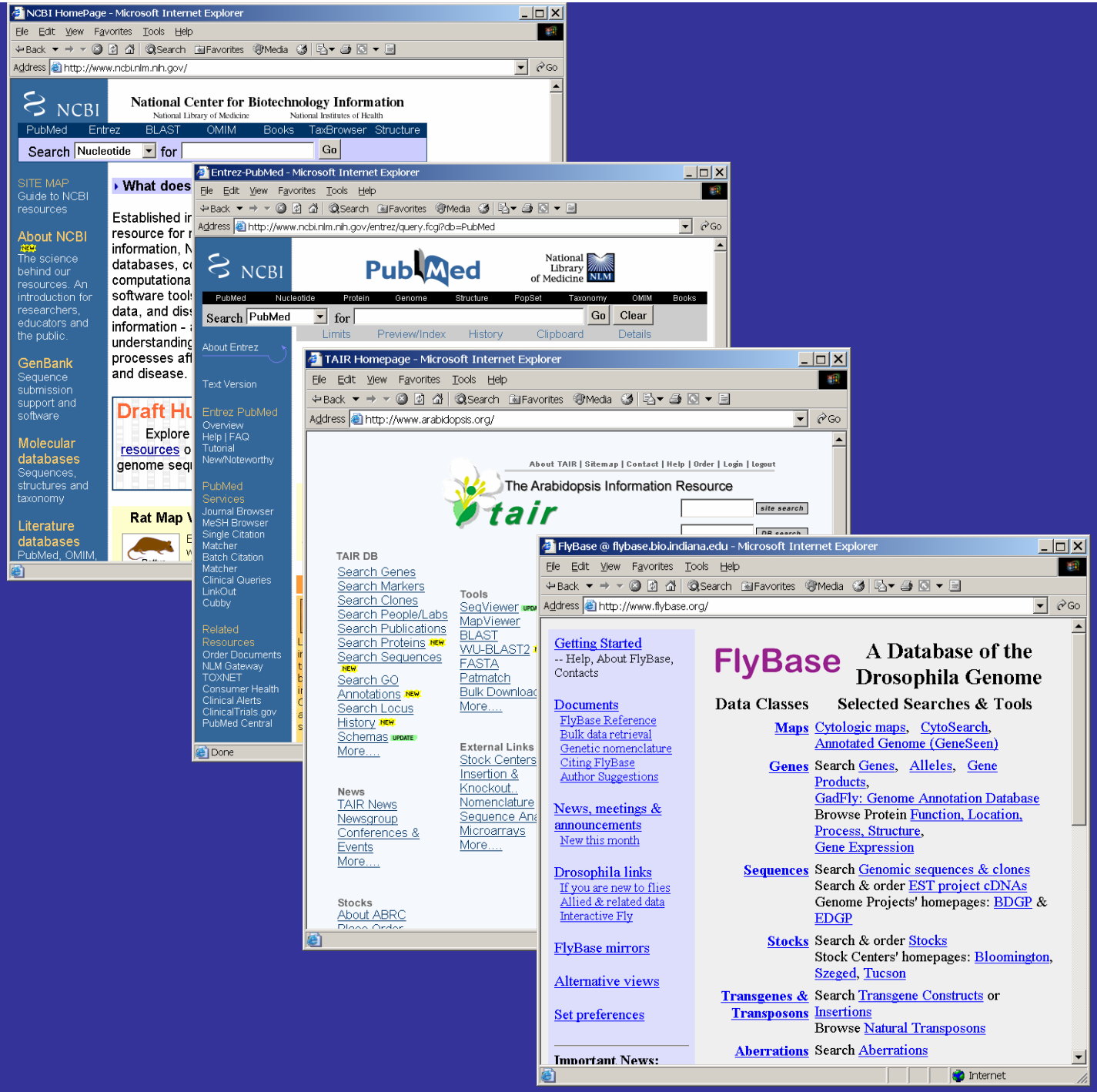
Overview of current biological databases

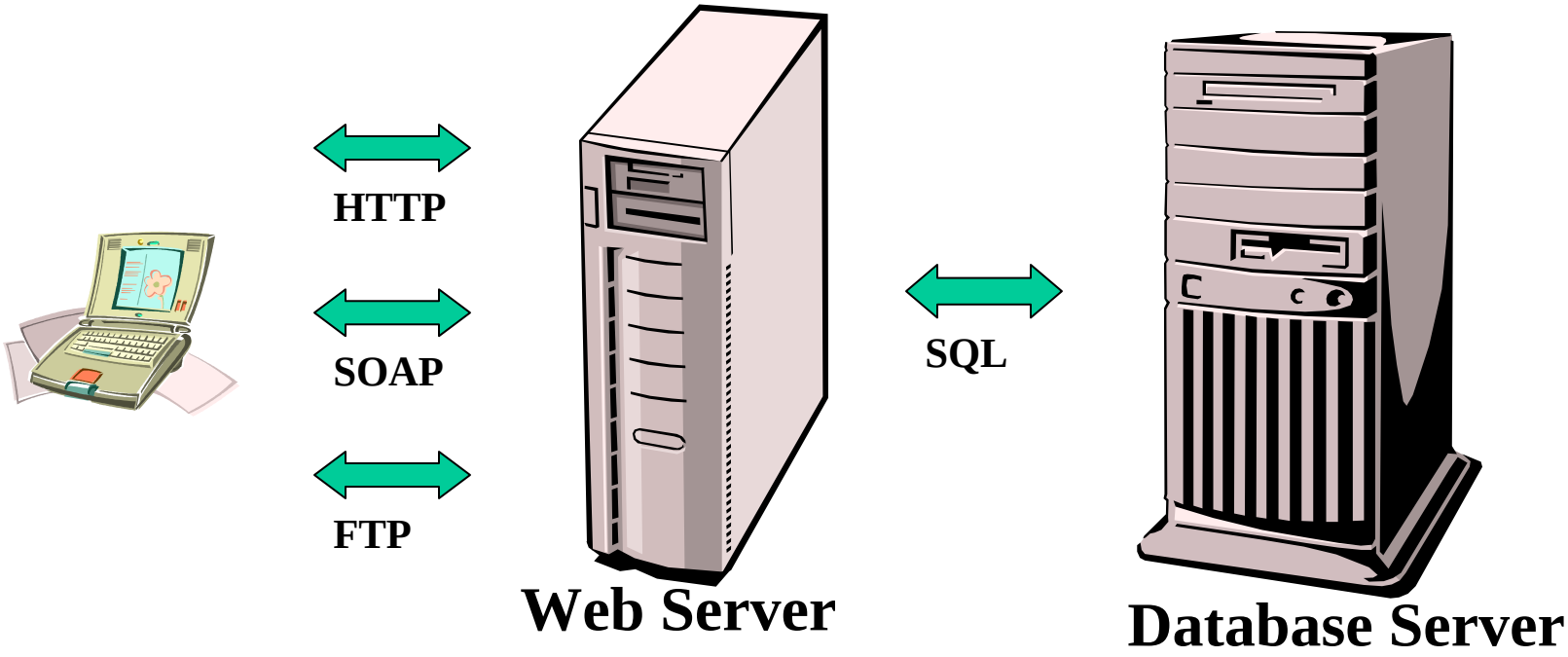
Qi Sun

Computational Biology Service Unit

Cornell University


NCBI HomePage - Microsoft Internet Explorer
http://www.ncbi.nlm.nih.gov/
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
PubMed Entrez BLAST OMIM Books TaxBrowser Structure
Search Nucleotide for Go
SITE MAP
Guide to NCBI resources
About NCBI
The science behind our resources. An introduction for researchers, educators and the public.
GenBank
Sequence submission support and software
Molecular databases
Sequences, structures and taxonomy
Literature databases
PubMed, OMIM,
What does
Established in
resource for
information,
databases,
computationa
software tool
data, and dis
information -
understanding
processes af
and disease.
Draft Hu
Explore
resources o
genome seq
Rat Map
Entrez-PubMed - Microsoft Internet Explorer
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
NCBI
PubMed
National Library of Medicine
PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books
Search PubMed for Go Clear
Limits Preview/Index History Clipboard Details
About Entrez
Text Version
Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
PubMed Services
Journal Browser
MeSH Browser
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
LinkOut
Cubby
Related Resources
Order Documents
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central
Done
TAIR Homepage - Microsoft Internet Explorer
http://www.arabidopsis.org/
About TAIR | Sitemap | Contact | Help | Order | Login | Logout
The Arabidopsis Information Resource
tair
site search
TAIR DB
Search Genes
Search Markers
Search Clones
Search People/Labs
Search Publications
Search Proteins
Search Sequences
Search GO Annotations
Search Locus History
Schemas
More....
Tools
SeqViewer
MapViewer
BLAST
WU-BLAST2
FASTA
Patmatch
Bulk Download
More....
External Links
Stock Centers
Insertion & Knockout..
Nomenclature
Sequence Ana
Microarrays
More....
News
TAIR News
Newsgroup
Conferences & Events
More....
Stocks
About ABRC
FlyBase @ flybase.bio.indiana.edu - Microsoft Internet Explorer
http://www.flybase.org/
Getting Started
-- Help, About FlyBase, Contacts
Documents
FlyBase Reference
Bulk data retrieval
Genetic nomenclature
Citing FlyBase
Author Suggestions
News, meetings & announcements
New this month
Drosophila links
If you are new to flies
Allied & related data
Interactive Fly
FlyBase mirrors
Alternative views
Set preferences
Important News:
FlyBase
A Database of the Drosophila Genome
Data Classes
Selected Searches & Tools
Maps Cytologic maps, CytoSearch, Annotated Genome (GeneSeen)
Genes Search Genes, Alleles, Gene Products, GadFly: Genome Annotation Database
Browse Protein Function, Location, Process, Structure, Gene Expression
Sequences Search Genomic sequences & clones
Search & order EST project cDNAs
Genome Projects' homepages: BDGP & EDGP
Stocks Search & order Stocks
Stock Centers' homepages: Bloomington, Szeged, Tucson
Transgenes & Transposons Search Transgene Constructs or Insertions
Browse Natural Transposons
Aberrations Search Aberrations
Internet

Platforms for Bioinformatics



Platforms for Bioinformatics

Open source

- Linux
- Apache
- Mysql
- Perl/Python/PHP

Micorsoft

- Windows
- ASP.NET
- SQL Server
- C#

NCBI Sequence Data Model

Archival database (GenBank, GenPept)

vs

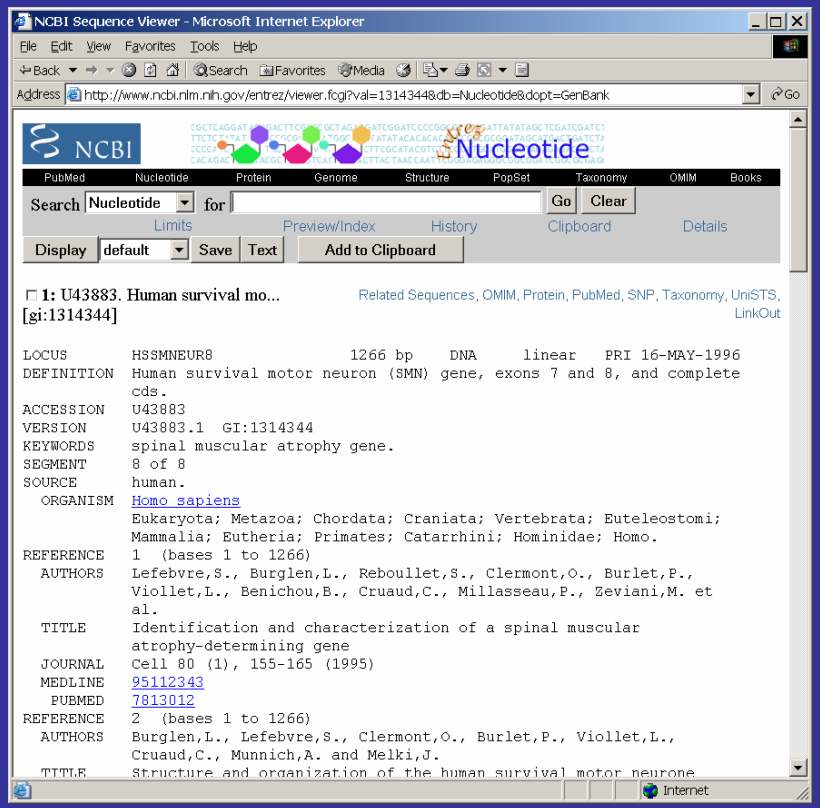
Computer algorithm generated database (Unigene)

vs

Manually curated database (RefSeq)

The NCBI Data Model

Genbank- A DNA centered database



Identifier:

1. **LOCUS (obsolete)**
2. **Accession (version)**
3. **GI**

1: U43883. Human survival mo... [gi:1314344]

Related Sequences, OMIM, Protein, PubMed, SNP, Taxonomy, UniSTS, LinkOut

```
LOCUS       HSSMNEUR8                1266 bp    DNA     linear   PRI 16-MAY-1996
DEFINITION  Human survival motor neuron (SMN) gene, exons 7 and 8, and complete
            cds.
ACCESSION   U43883
VERSION     U43883.1  GI:1314344
KEYWORDS    spinal muscular atrophy gene.
SEGMENT     8 of 8
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1266)
```

Features

```
mRNA            join(U43876.1:575..688,U43877.1:104..175,
                U43878.1:118..237,U43879.1:84..284,U43880.1:69..221,
                U43881.1:103..198,U43882.1:53..163,209..262,707..>1266)
                /gene="SMN"
                /product="survival motor neuron"
                /note="spinal muscular atrophy gene; cDNA sequence found
                in GenBank Accession Number U18423"
CDS             join(U43876.1:608..688,U43877.1:104..175,
                U43878.1:118..237,U43879.1:84..284,U43880.1:69..221,
                U43881.1:103..198,U43882.1:53..163,209..259)
                /gene="SMN"
                /note="spinal muscular atrophy gene"
                /codon_start=1
                /product="survival motor neuron"
                /protein_id="AAC50473.1"
                /db_xref="GI:1314346"
                /translation="MAMSSGGSGGGVPEQEDSVLFRRGTGQSDDSDIWDDTALIKAYD
                KAVASFKHALKNGDICETSGKPKTTPKRKPAKKNKSQKKNTAASLQQWKVGDKCSAIW
                SEDGCIYPATIASIDFKRETCVVVYTGYGNREEQNLSDLLSPICEVANNIEQNAQENE
                NESQVSTDESENSRSPGNKSDNIKPKSAPWNSFLPPPPPMPGPRLGPGKPGLKFNGPP
                PPPPPPPPHLLSCWLPPFPSGPPIIPPPPPICPDSLDDADALGSMLISWYMSGYHTGY
                YMGFRQNQKEGRCSHSLN"
```

GenPept- A protein centered database

☐ **1:** AAC50473. survival motor ne...[gi:1314346] BLink, Nucleotide, OMIM, Related Sequences, PubMed, SNP, Taxon

```
LOCUS       AAC50473                294 aa            linear   PRI 16-MAY-1996
DEFINITION  survival motor neuron.
ACCESSION   AAC50473
PID         g1314346
VERSION     AAC50473.1  GI:1314346
DBSOURCE    locus HSSMNEUR1 accession U43876.1
            locus HSSMNEUR2 accession U43877.1
            locus HSSMNEUR3 accession U43878.1
            locus HSSMNEUR4 accession U43879.1
            locus HSSMNEUR5 accession U43880.1
            locus HSSMNEUR6 accession U43881.1
            locus HSSMNEUR7 accession U43882.1
            locus HSSMNEUR8 accession U43883.1
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 294)
  AUTHORS   Burglen,L., Lefebvre,S., Clermont,O., Burlet,P., Viollet,L.,
            Cruaud,C., Munnich,A. and Melki,J.
  TITLE     Structure and organization of the human survival motor neurone
            (SMN) gene
  JOURNAL   Genomics 32 (3), 479-482 (1996)
  MEDLINE   96435930
   PUBMED   8838816
REFERENCE   2  (residues 1 to 294)
  AUTHORS   Burglen,L.
```

FTP sites:

GenBank: ftp://ftp.ncbi.nih.gov/genbank/

GenPept: ftp://ftp.ncifcrf.gov/pub/genpept/

Problems with Genbank and Genpept

- It does not distinguish the sequence categories.
- Lot of redundancy.
  - Same gene could be deposited into the database many times with different names
  - Different version of the same gene could be submitted many times with different accession number.
- The features of genbank record could be chaotic.

NCBI Sequence Databases

Archival database (GenBank, GenPept)

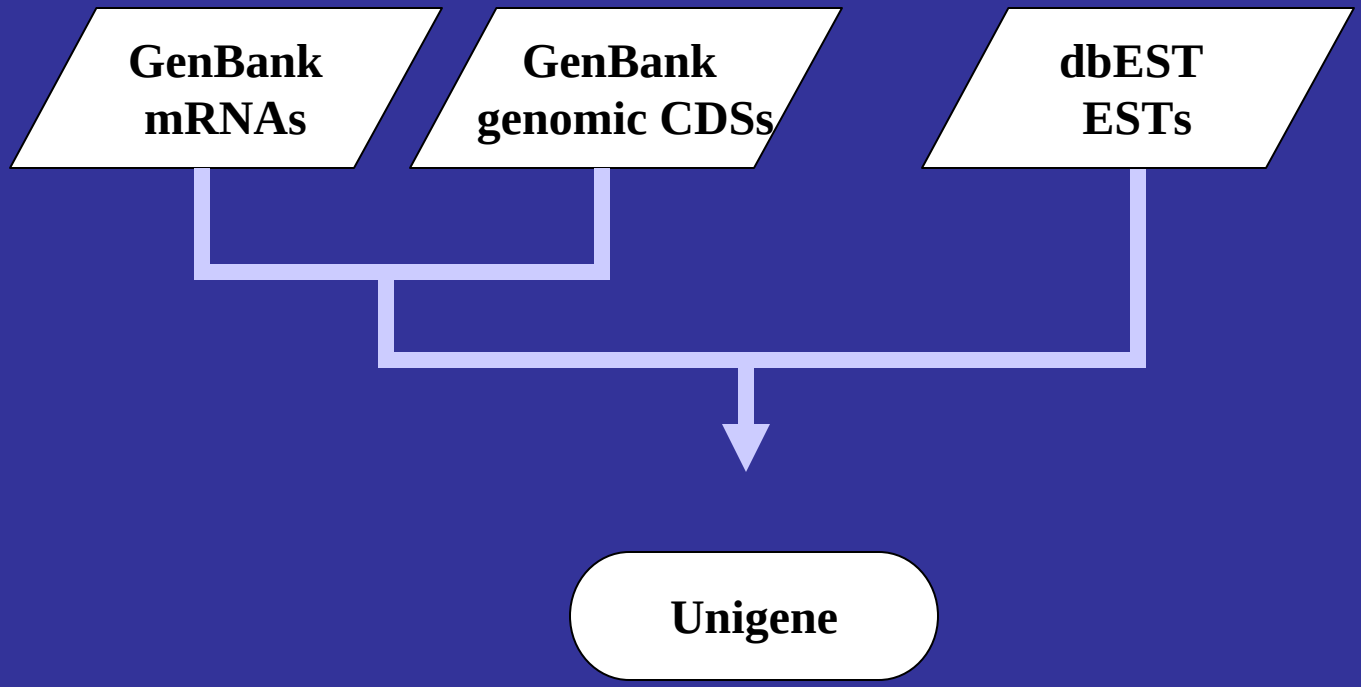
vs

Computer algorithm generated database (Unigene)

vs

Curated database (RefSeq, Locuslink ...)

UniGene
a non-redundant set of gene-oriented clusters



Unigene identifier

Hs for human
Mm for mouse
Rn for rat
Bt for cow
Dr for zebrafish
Dm for fruitfly
Aga for mosquito
Xl for frog
At for cress
Hv for barley
Os for rice
Ta for wheats
Zm for maize

Examples:

Mm.213407

Hs.13303

At.138

NCBI Sequence Databases

Archival database (GenBank, GenPept)

vs

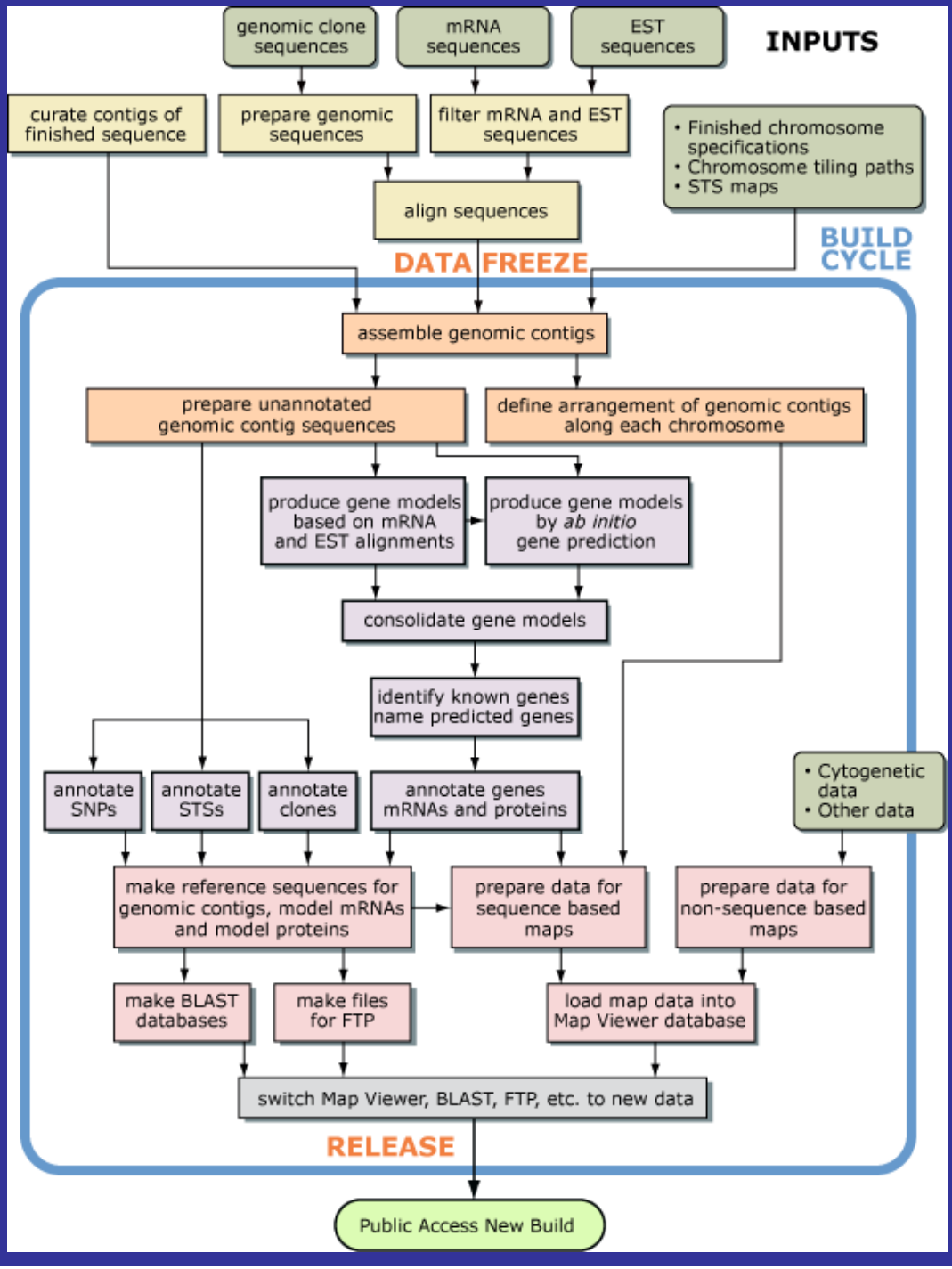
Computer generated database (Unigene)

vs

Curated database (RefSeq, Gene ...)

NCBI human genome annotation pipeline

The refseq incorporate the predicted transcript and protein sequences, experimentally identified mRNA sequences, EST sequences.



Refseq Accession Numbers:

NT_123456 constructed genomic contigs

NM_123456 mRNAs

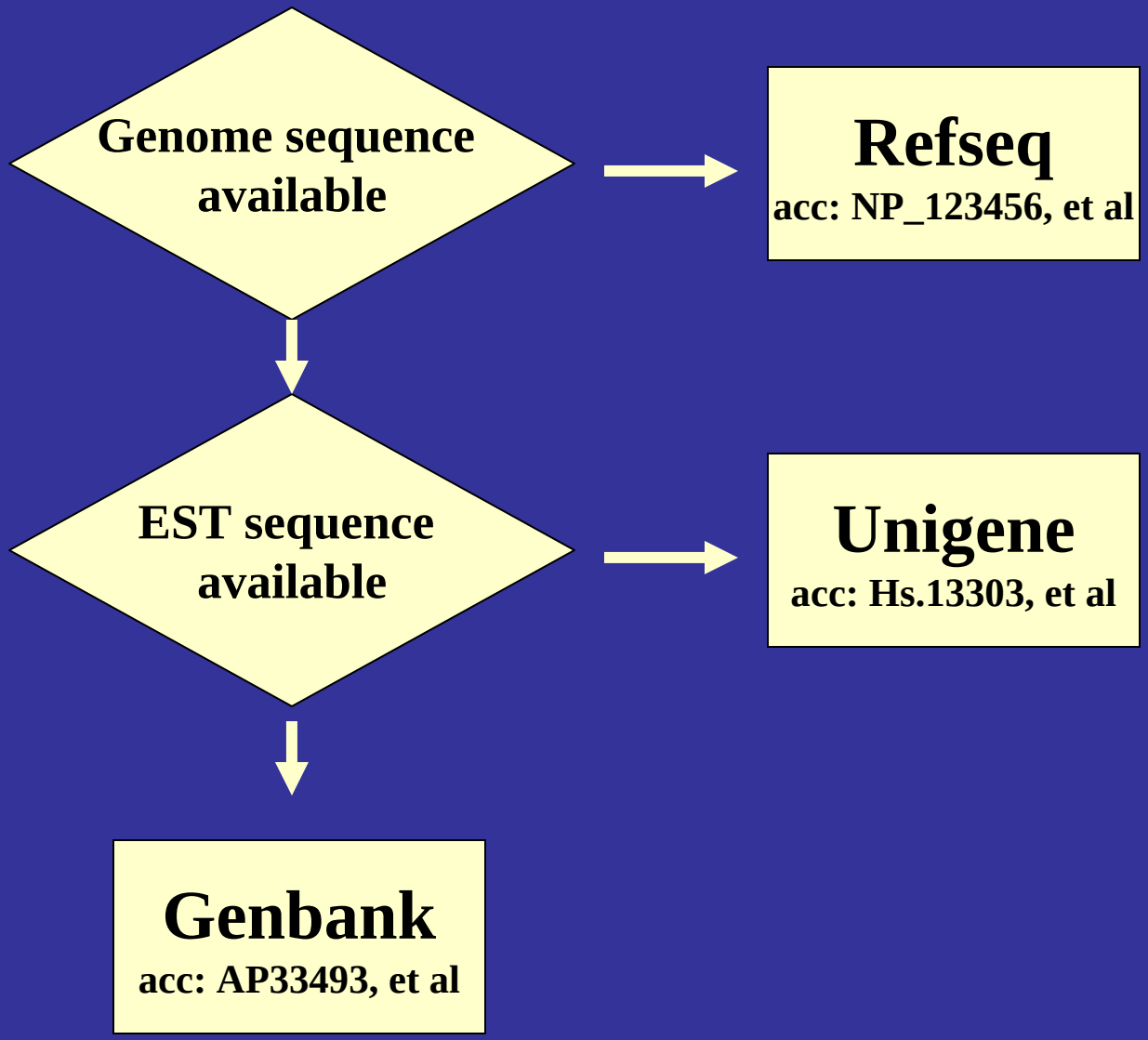
NP_123456 proteins

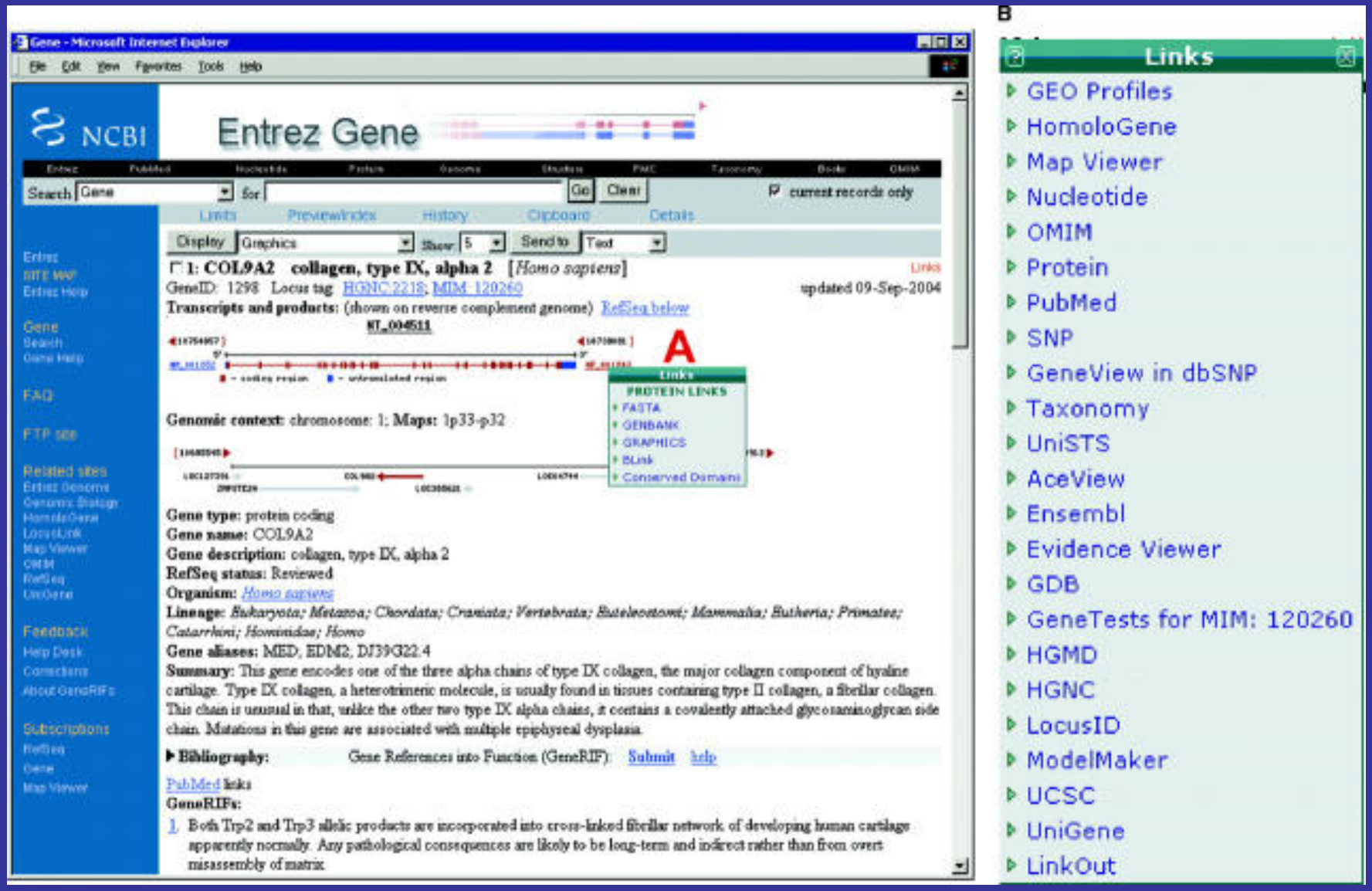
NC_123456 chromosomes

XM_123456 predicted mRNA

XP_123456 predicted protein

Refseq? Unigene? Genbank?




Gene - Microsoft Internet Explorer
NCBI
Entrez Gene
Search Gene for Go Clear current records only
Limits PreviewIndex History Clipboard Details
Display Graphics Show 5 Send to Text
1: COL9A2 collagen, type IX, alpha 2 [Homo sapiens]
GeneID: 1298 Locus tag HGNC:2218, MIM:120260
updated 09-Sep-2004
Transcripts and products: (shown on reverse complement genome) RefSeq below
NT_004511
A
Links
PROTEIN LINKS
FASTA
GENBANK
GRAPHICS
BLink
Conserved Domains
Genomic context: chromosome: 1; Maps: 1p33-p32
Gene type: protein coding
Gene name: COL9A2
Gene description: collagen, type IX, alpha 2
RefSeq status: Reviewed
Organism: Homo sapiens
Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo
Gene aliases: MED, EDM2, DJ39G22.4
Summary: This gene encodes one of the three alpha chains of type IX collagen, the major collagen component of hyaline cartilage. Type IX collagen, a heterotrimeric molecule, is usually found in tissues containing type II collagen, a fibrillar collagen. This chain is unusual in that, unlike the other two type IX alpha chains, it contains a covalently attached glycosaminoglycan side chain. Mutations in this gene are associated with multiple epiphyseal dysplasia.
Bibliography: Gene References into Function (GeneRIF): Submit help
PubMed links
GeneRIFs:
1. Both Trp2 and Trp3 allelic products are incorporated into cross-linked fibrillar network of developing human cartilage apparently normally. Any pathological consequences are likely to be long-term and indirect rather than from overt misassembly of matrix
B
Links
GEO Profiles
HomoloGene
Map Viewer
Nucleotide
OMIM
Protein
PubMed
SNP
GeneView in dbSNP
Taxonomy
UniSTS
AceView
Ensembl
Evidence Viewer
GDB
GeneTests for MIM: 120260
HGMD
HGNC
LocusID
ModelMaker
UCSC
UniGene
LinkOut

Go to the web

Files that you can download from the NCBI gene database

gene_info
gene2refseq
gene2go

NCBI Search engine

Entrez

- boolean operators “AND” “OR” “NOT”
- entrez tags
- using limits
- MeSH terms

Batch Entrez

search by accession list

Other Sequence Databases:

Genomic DNA: Ensembl Genome annotation database

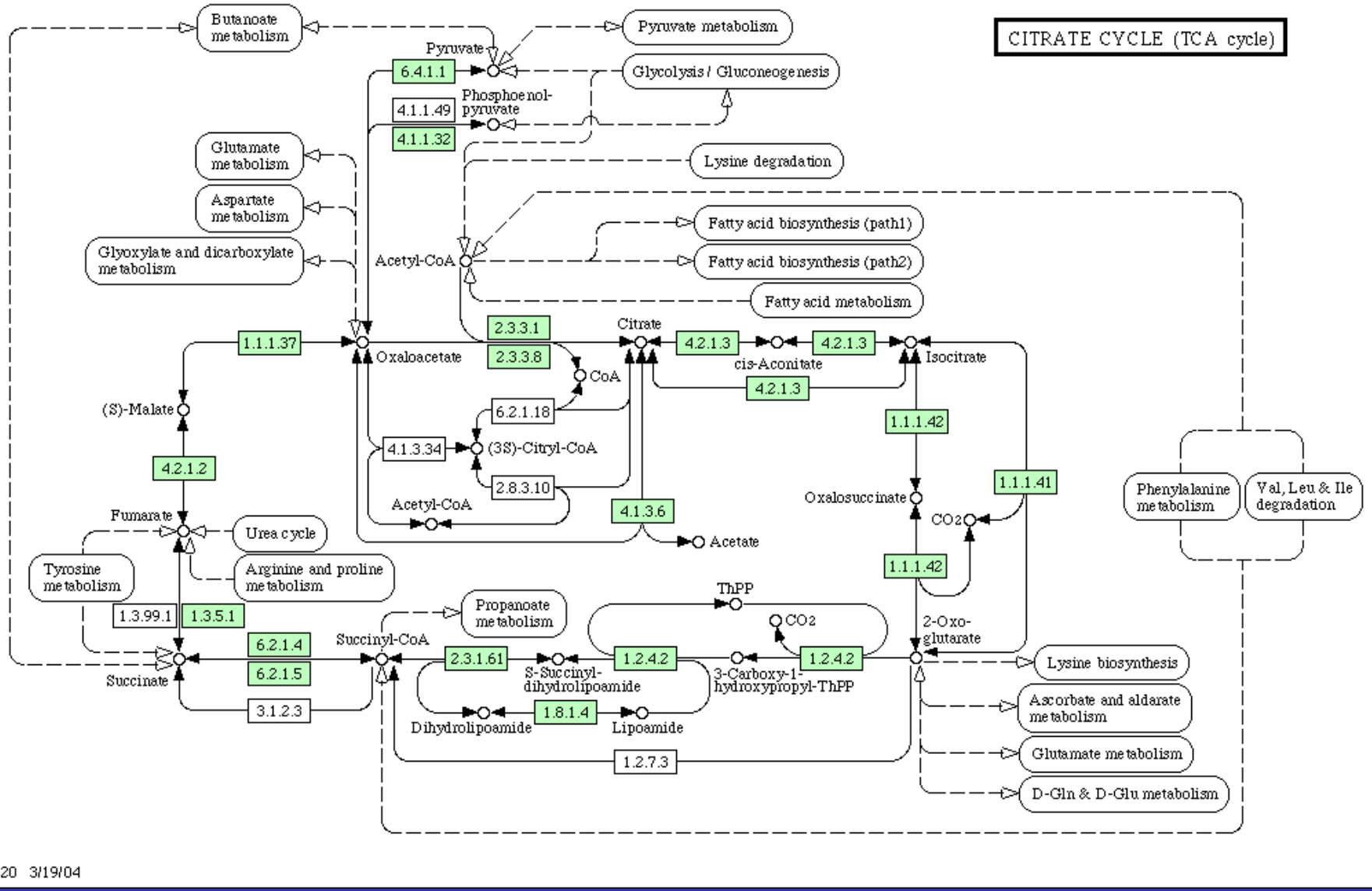
(http://www.ensembl.org, HTTP, FTP, MySQL interface)

Protein: Uniprot

(http://www.pir.uniprot.org/)

KEGG database

go to the web



GO

Gene Ontology

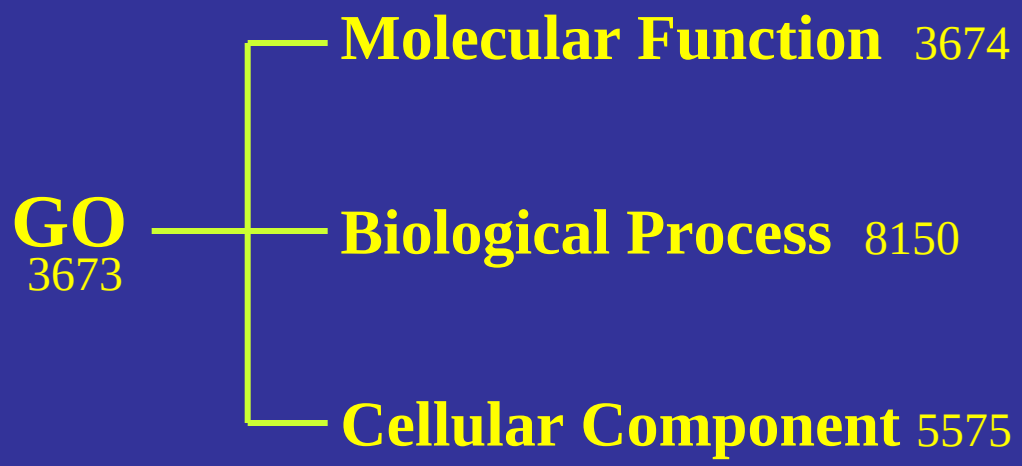
1. **Molecular Function**
2. **Biological Process**
3. **Cellular Component**

http://www.geneontology.org

Public Database - 2

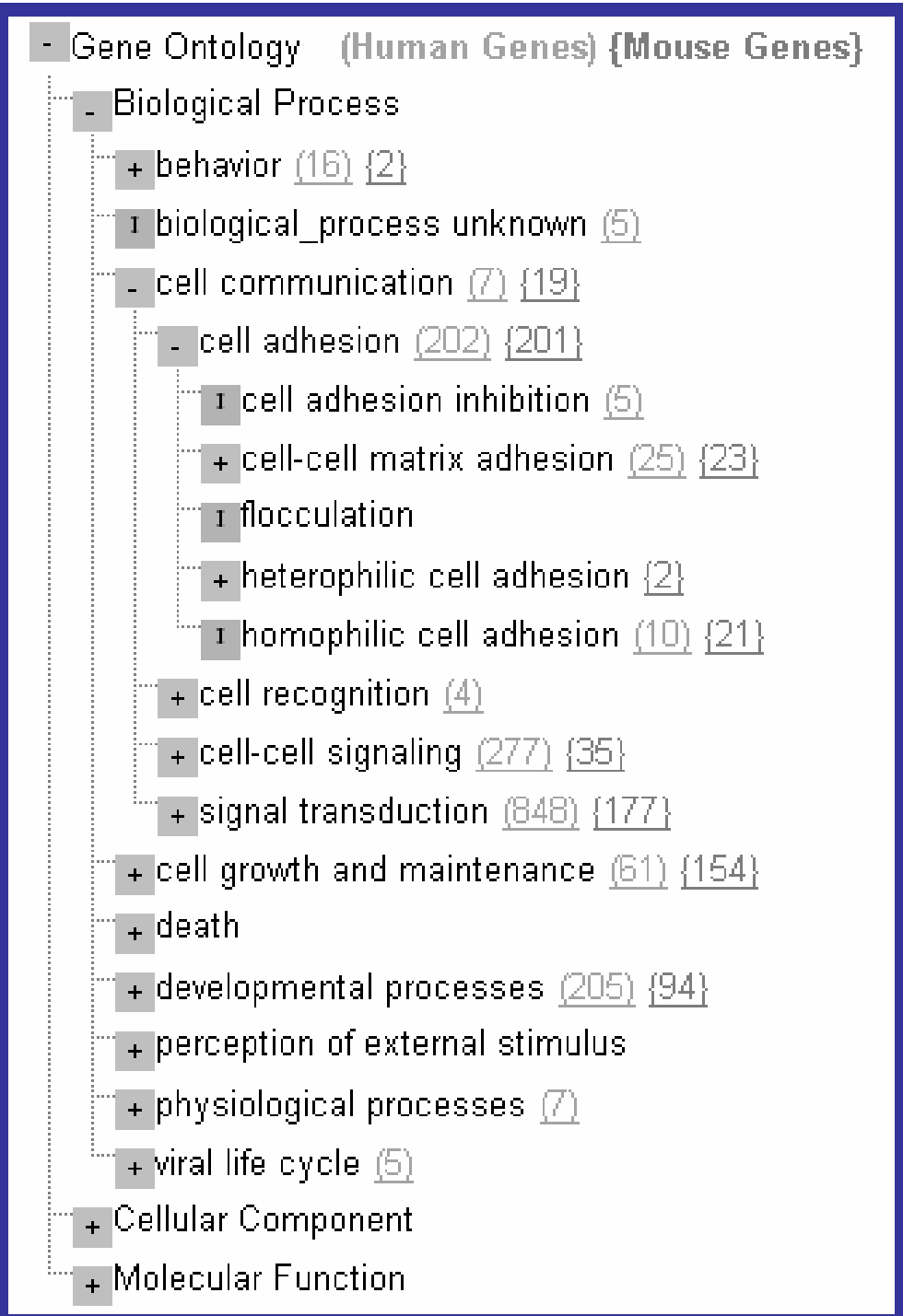
- Eubacteria
- **Eukaryotes**
  - **Animals**
    - Echinoderms (sea urchins, starfish, sea cucumbers, etc)
    - **Vertebrates (fish etc.)**
      - **Terrestrial Vertebrates**
        - Frogs
        - Salamanders
        - Turtles
        - **Dinosaurs**
          - Modern Birds
        - Mammals
      - Teleost fish
    - Cnidaria (jellyfish, anemones, corals, etc.)
    - Annelida (segmented worms)
    - Cephalopoda (octopods, squids, etc.)
    - **Arthropoda**
      - **Insects**
        - Dragonflies and Damselflies
        - Lice
        - True Bugs
        - Beetles
        - Wasps, Bees, and Ants
        - Flies
        - Butterflies and Moths
        - Crickets, Katydids, and Grasshoppers
      - **Arachnids**
        - Spiders
        - Mites
        - Scorpions
  - Fungi
  - **Green Plants**
    - Ferns
    - Flowering Plants

Public Database - 2



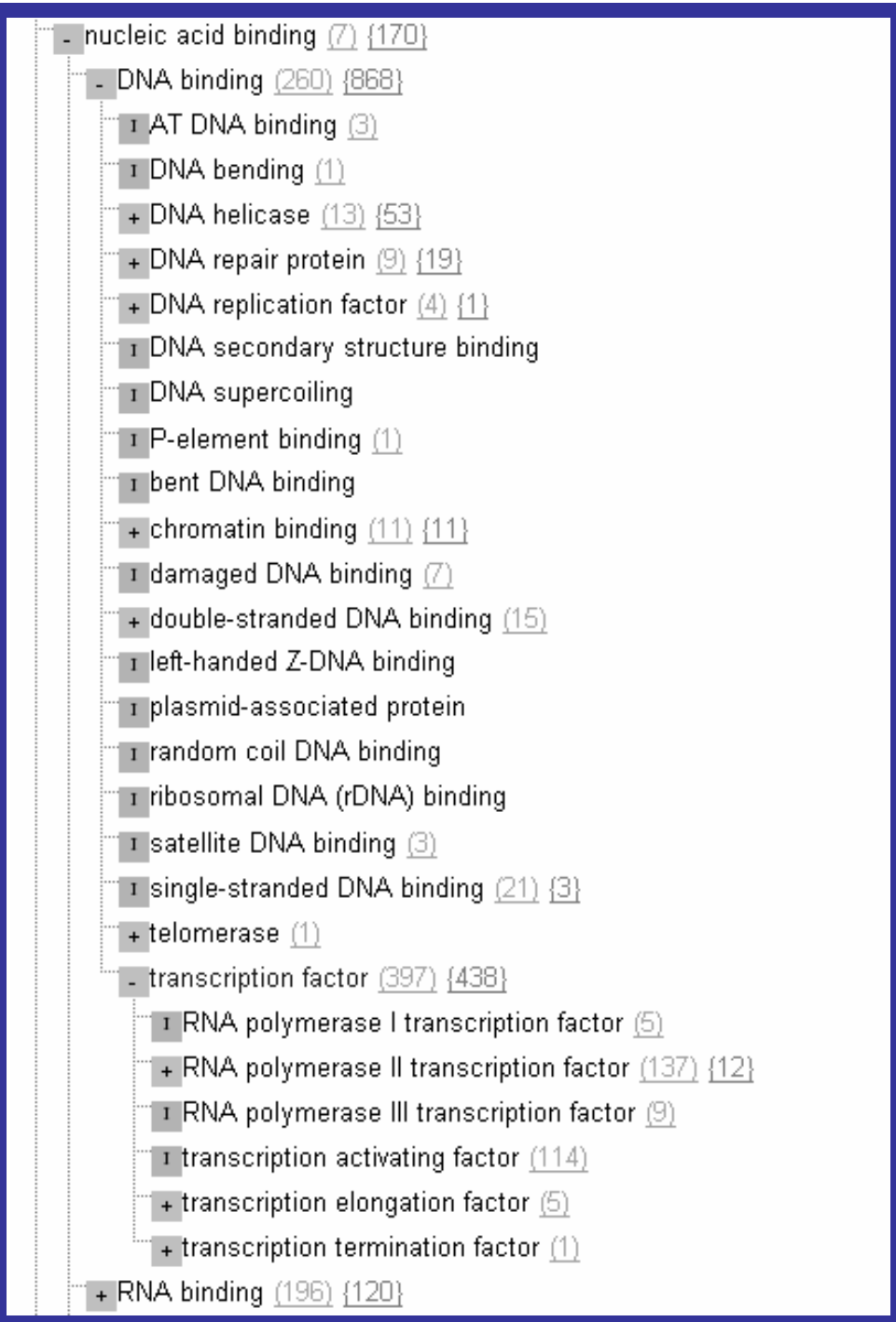
GO Example 1:

Biological Process



GO Example 2:

Molecular Function



Gene Ontology Annotation

Smn: survival motor neuron
Gene ID: 39844

Gene Ontology™:

Term	Evidence	Source	Pub
• nucleus	IDA	MGD	pm
• nucleus	IEA	MGD	pm
• cytoplasm	IDA	MGD	pm
• RNA binding	IEA	MGD	pm
• mRNA processing	IEA	MGD	pm
• nucleic acid binding	IEA	MGD	pm

Species Specific Databases

- Arabidopsis – TAIR
- Yeast – SGD
- Fly – FLYBASE
- Worm – WORMBASE
- Mouse – MGD