NCBI Field Guide

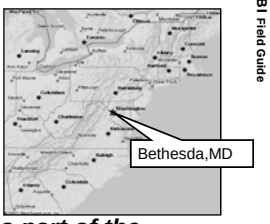
NCBI Molecular Biology Resources

NCBI Databases

March 2007

NCBI Field Guide

The National Center for Biotechnology Information



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

NCBI Field Guide

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

NCBI Field Guide

NCBI Databases and Services

- GenBank largest sequence database
- Free public access to biomedical literature
  - PubMed free Medline
  - PubMed Central full text online access
- Entrez integrated molecular and literature databases
- BLAST highest volume sequence search service
- VAST structure similarity searches
- Software and Databases

Types of Databases

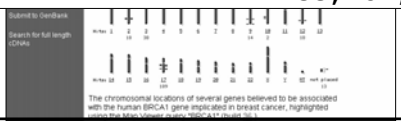
NCBI Field Guide

- Primary Databases
  - Original submissions by experimentalists
  - Content controlled by the submitter
    - Examples: GenBank, SNP, GEO
- Derivative Databases
  - Built from primary data
  - Content controlled by third party (NCBI)
    - Examples: Refseq, TPA, RefSNP, UniGene, NCBI Protein, Structure, Conserved Domain

Entrez Nucleotides

NCBI Field Guide

Primary	
• GenBank / EMBL / DDBJ	**86,766,287**
Derivative	
• RefSeq	**1,715,255**
• Third Party Annotation	**5,312**
• PDB	**7,334**
Total	**88,494,392**



What is GenBank?

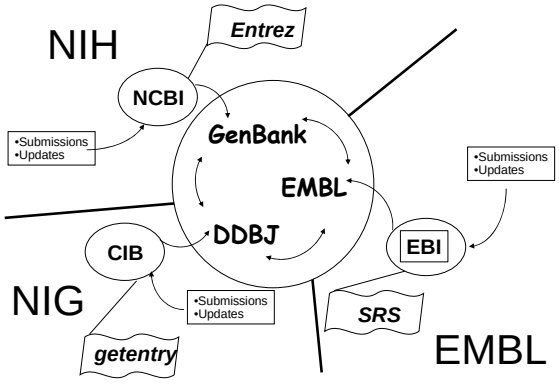
NCBI's Primary Sequence Database

NCBI Field Guide

- Nucleotide only sequence database
- Archival in nature
  - Historical
  - Reflective of submitter point of view (subjective)
  - Redundant
- GenBank Data
  - Direct submissions (traditional records)
  - Batch submissions (EST, GSS, STS)
  - ftp accounts (genome data)
- Three collaborating databases
  - GenBank
  - DNA Database of Japan (DDBJ)
  - European Molecular Biology Laboratory (EMBL) Database

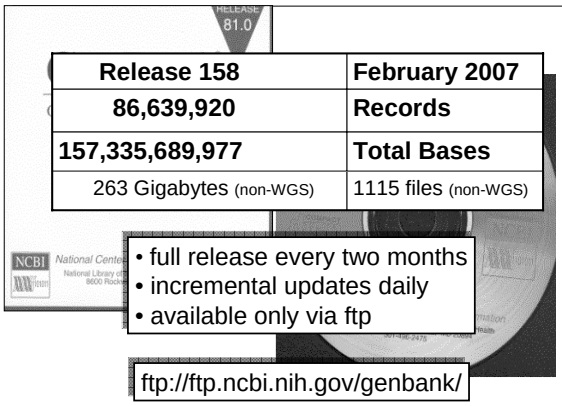
International Sequence Database Collaboration

NCBI Field Guide



GenBank: NCBI's Primary Sequence Database

NCBI Field Guide



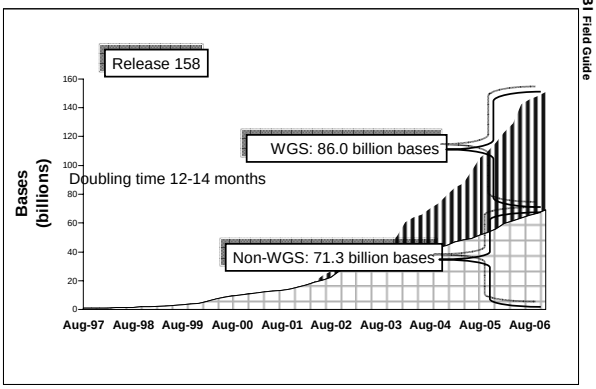
Release 158	February 2007
86,639,920	Records
157,335,689,977	Total Bases
263 Gigabytes (non-WGS)	1115 files (non-WGS)

- full release every two months
- incremental updates daily
- available only via ftp

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

The Growth of GenBank

NCBI Field Guide



Organization of GenBank: Traditional Divisions

NCBI Field Guide

Records are divided into 18 Divisions.

- 12 Traditional
- 6 Bulk

Traditional Divisions:
- Direct Submissions (Sequin and BankIt)
- Accurate
- Well characterized

PRI Primate
PLN Plant and Fungal
BCT Bacterial and Archeal
INV Invertebrate
ROD Rodent
VRL Viral
VRT Other Vertebrate
MAM Mammalian
PHG Phage
SYN Synthetic (cloning vectors)
ENV Environmental Samples
UNA Unannotated

Entrez query: `gbdiv_xxx[Properties]`

Organization of GenBank: Bulk Divisions

NCBI Field Guide

Records are divided into 18 Divisions.

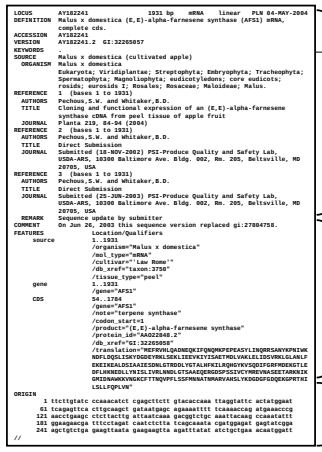
- 12 Traditional
- 6 Bulk

BULK Divisions:
- Batch Submission (Email and FTP)
- Inaccurate
- Poorly characterized

EST Expressed Sequence Tag
GSS Genome Survey Sequence
HTG High Throughput Genomic
STS Sequence Tagged Site
HTC High Throughput cDNA
PAT Patent

Entrez query: `gbdiv_xxx[Properties]`

A Traditional GenBank Record



Header

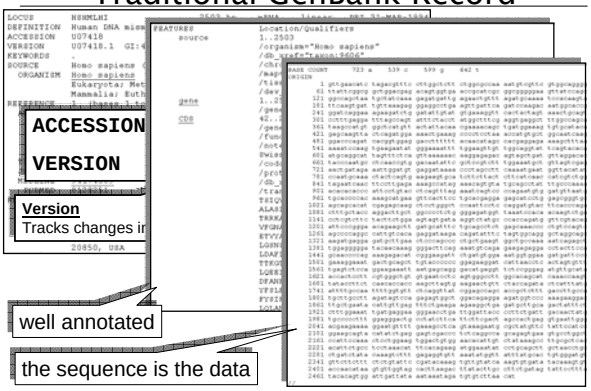
The Flatfile Format

Feature Table

Sequence

NCBI Field Guide

Traditional GenBank Record



NCBI Field Guide

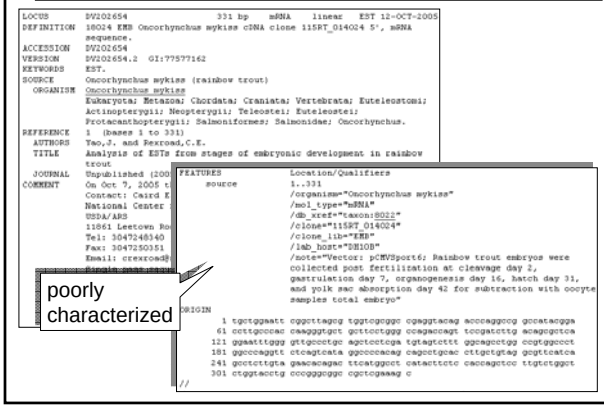
Bulk Divisions

- Batch Submission and htg (email and ftp)
- Inaccurate
- Poorly Characterized

- Expressed Sequence Tag
  - 1st pass single read cDNA
- Genome Survey Sequence
  - 1st pass single read gDNA
- High Throughput Genomic
  - incomplete sequences of genomic clones
- Sequence Tagged Site
  - PCR-based mapping reagents

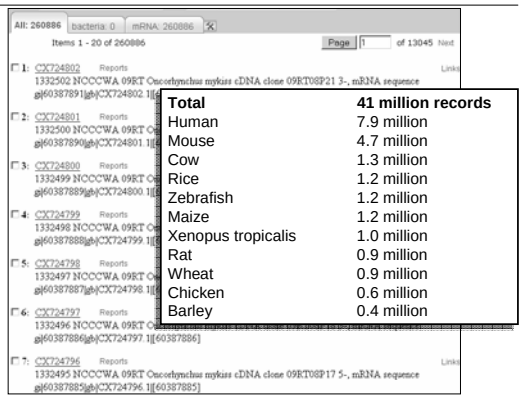
NCBI Field Guide

GenBank Bulk Sequence: EST



NCBI Field Guide

ESTs in Entrez



Total	41 million records
Human	7.9 million
Mouse	4.7 million
Cow	1.3 million
Rice	1.2 million
Zebrafish	1.2 million
Maize	1.2 million
Xenopus tropicalis	1.0 million
Rat	0.9 million
Wheat	0.9 million
Chicken	0.6 million
Barley	0.4 million

NCBI Field Guide

Derivative Databases

NCBI Field Guide

Entrez Protein: Derivative Database

Data Source	Sequences
GenPept	**6,937,176**
RefSeq	**3,359,561**
Third Party Annotation	**5,136**
Swiss Prot	**255,159**
PIR	**29,996**
PRF	**12,079**
PDB	**91,116**
PAT Division	**669,035**
Total	**10,690,223**
BLAST nr total (no patents or env)	**4,545,310**

NCBI Field Guide

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...

NCBI Field Guide

Redundant Proteins

NCBI Field Guide

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

```
>gi|13905126|gb|AAH06850.1| MutL protein homolog 1 ...
MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIQVIV...
EDLDIVCERFTTSKLQSFEDLASISTYGFRGEALASISHVAHVTITTKTAD...
```

```
>gi|1079787|gb|AAA82079.1| DNA mismatch repair prot...
MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIQVIV...
EDLDIVCERFTTSKLQSFEDLASISTYGFRGEALASISHVAHVTITTKTAD...
```

GenPept

```
>gi|4557757|ref|NP_000240.1| MutL protein homolog 1...
MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIQVIV...
EDLDIVCERFTTSKLQSFEDLASISTYGFRGEALASISHVAHVTITTKTAD...
```

NCBI RefSeq

```
>gi|730028|sp|P40692|MLH1_HUMAN DNA mismatch repair...
MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIQVIV...
EDLDIVCERFTTSKLQSFEDLASISTYGFRGEALASISHVAHVTITTKTAD...
```

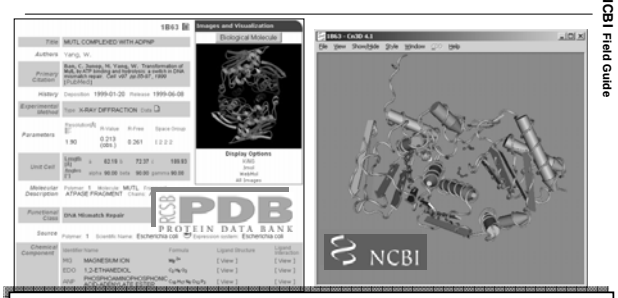
Swiss-Prot

```
>gi|741682|prf||2007430A DNA mismatch repair protei...
MSFVAGVIRRLDETVVNRIAAGEVIQRPANAIKEMIENCLDAKSTSIQVIV...
EDLDIVCERFTTSKLQSFEDLASISTYGFRGEALASISHVAHVTITTKTAD...
```

PRF

Protein Sequences from Structures

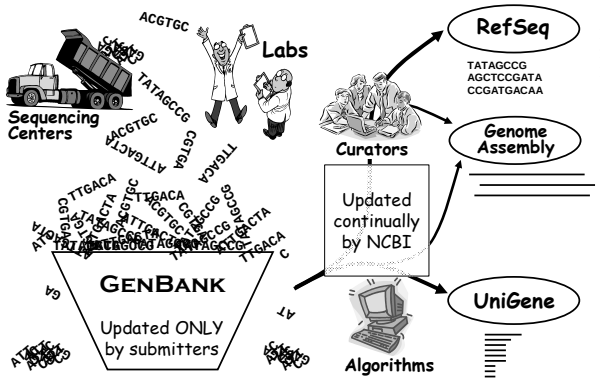
NCBI Field Guide




```
>gi|5542073|pdb|1B63|A Chain A, Mutl Complexed With Adpnp
SHMPIQVLPPQLANQIAAGEVVERPASVVKELVENSLDAGATRIDIDIERGGAKLIRIRDNGCGIKKDEL
ALALARHATSKIASLDDLEAIISLGFRGEALASISSVSRLTLTSRTAEQQEAWQAYAEGRDMNVTVKPAA
HPVGTTLEVLDLFYNTPARRKFLRTEKTEFNHIDEIIRRIALARFDVTINLSHNGKIVRQYRAVPEGGQK
ERRLGAICGTAFLEQALAIEWQHGDLTLRGWVADPNHTTPALAEIQYCYVNGRMMRDRLINHAIRQACED
KLGADQQPAFVLYLEIDPHQVDVNVHPAKHEVRFHQSRLVHDFIYQGVLSVLQ
```

Primary vs. Derivative Sequence Databases

NCBI Field Guide



RefSeq: NCBI's Derivative Sequence Database

NCBI Field Guide

- **Curated transcripts and proteins**
  - reviewed
  - human, mouse, rat, fruit fly, zebrafish, arabidopsis microbial genomes (proteins), and more
- **Model transcripts and proteins**
- **Assembled Genomic Regions (contigs)**
  - human genome
  - mouse genome
  - rat genome
  - chicken
  - honeybee
  - sea urchin
- **Chromosome records**
  - Human genome
  - microbial
  - organelle

`srcdb_refseq[Properties]`

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

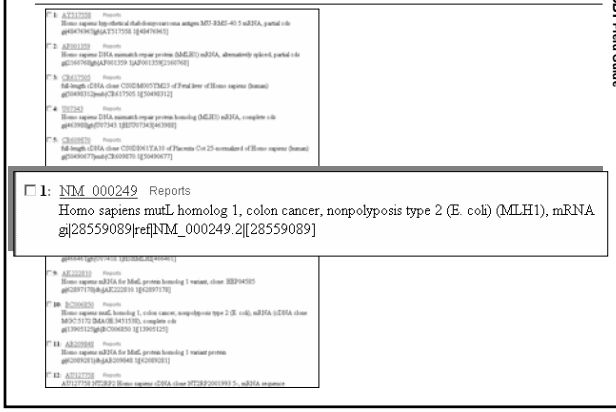
Selected RefSeq Accession Numbers

NCBI Field Guide

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
Assemblies	
NT_123456	Contig
NW_123456	WGS Supercontig

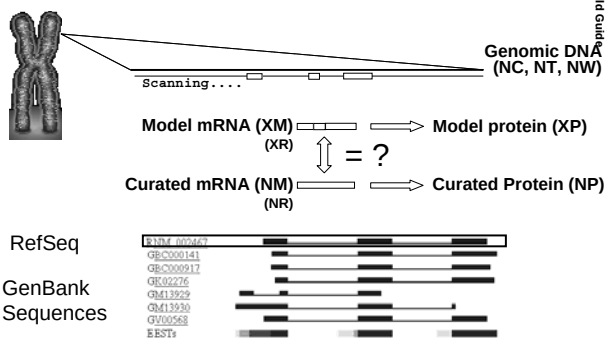
GenBank to RefSeq

NCBI Field Guide



RefSeqs: Annotation Reagents

NCBI Field Guide

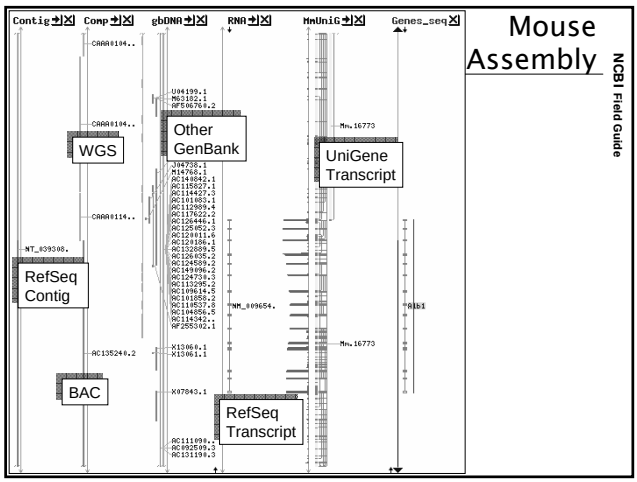


RefSeq Benefits

NCBI Field Guide

- non-redundancy
- explicitly linked nucleotide and protein sequences
- updates to reflect current sequence data and biology
- data validation
- format consistency
- distinct accession series
- stewardship by NCBI staff and collaborators

Mouse Assembly



NCBI Field Guide

Expressed Sequences

UniGene
GEO

NCBI Field Guide

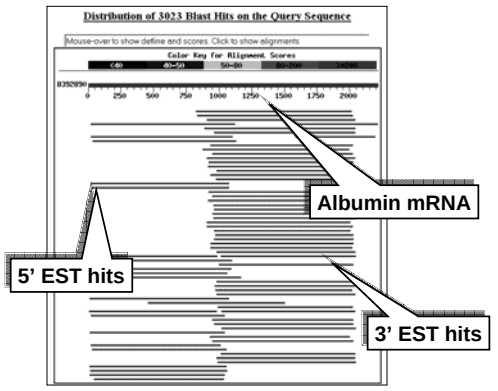
What is UniGene?

A gene-oriented view of sequence entries

- **MegaBlast based automated sequence clustering**
- **Now informed by genome hits *New!***
- **Nonredundant set of gene oriented clusters**
- **Each cluster a unique gene**
- **Information on tissue types and map locations**
- **Includes known genes and uncharacterized ESTs**
- **Useful for gene discovery and selection of mapping reagents**

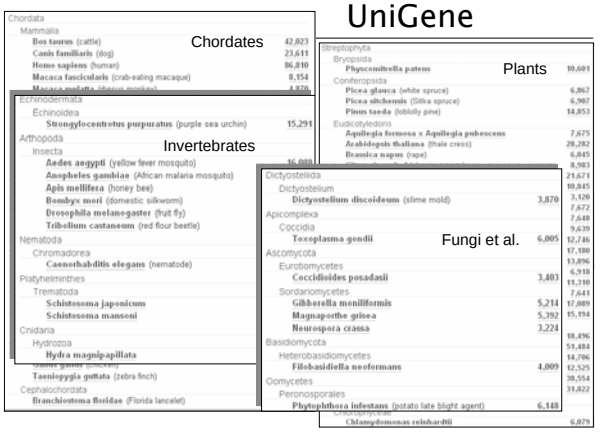
NCBI Field Guide

EST hits: Human mRNA



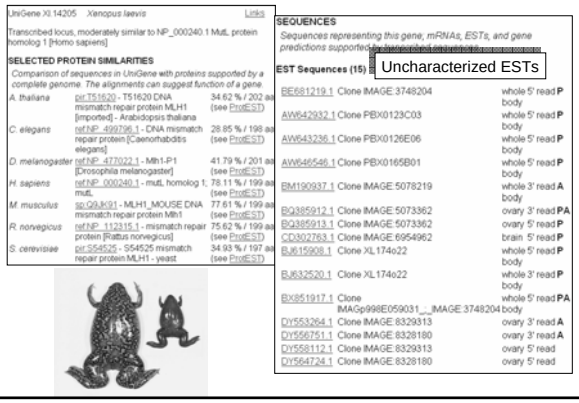
NCBI Field Guide

UniGene



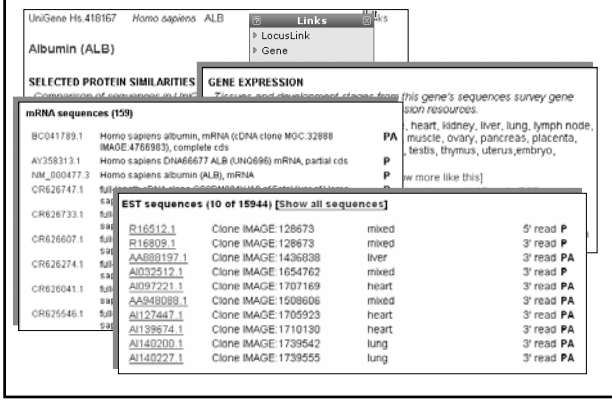
NCBI Field Guide

Xenopus laevis MLH1 Cluster



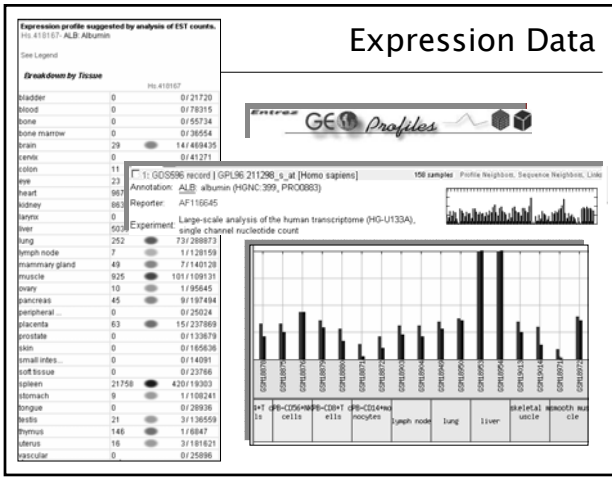
NCBI Field Guide

Human ALB Cluster



NCBI Field Guide

Expression Data



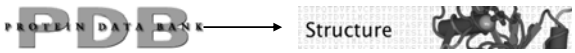
NCBI Field Guide

Other NCBI Databases

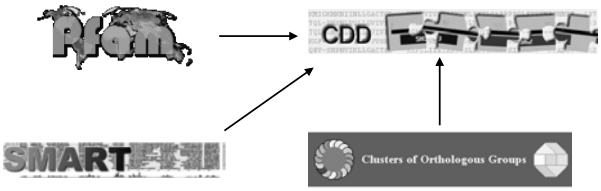
NCBI Field Guide

•Structure:	imported structures (PDB) Cn3D viewer, NCBI curation
•CDD:	conserved domain database Protein families (COGs and KOGs) Single domains (PFAM, SMART, CD)
•dbSNP:	nucleotide polymorphism
•Gene:	gene records Unifies LocusLink and Microbial Genomes

NCBI Field Guide



NCBI Structures and Domains



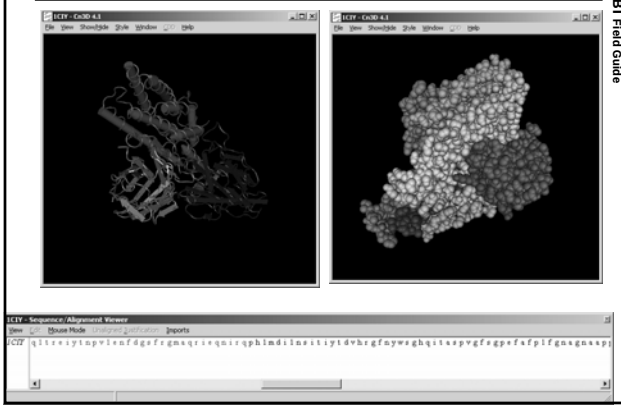
MMDB: Molecular Modeling Data Base

NCBI Field Guide

- Derived from experimentally determined PDB records
- Value added to PDB records including:
  - Addition of explicit chemical graph information
  - Validation (secondary structure elements)
  - Inclusion of Taxonomy, Citation
  - Conversion to ASN.1 data description language
- Structure neighbors determined by
  Vector Alignment Search Tool (VAST)

Cn3D 4.1: *Bacillus thuringiensis* Toxin

NCBI Field Guide



VAST: Structure Neighbors

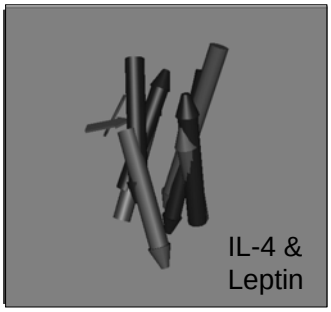
Vector Alignment Search Tool

For each protein chain,

locate SSEs (secondary structure elements),

and represent them as individual vectors.

align the vectors



Human IL-4

NCBI Field Guide

Protein Domains

- Structural Domain
  - Discrete independently folding unit of a protein
- Conserved Domain (sequence-based)
  - Protein region with recognizable position-specific pattern of sequence conservation
- Sequence-based domains often roughly correspond to structural domains
- Domains often have distinct, identifiable functions

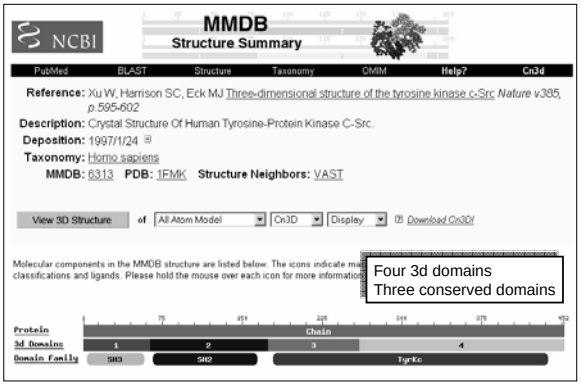
NCBI Field Guide

NCBI's Conserved Domain Database

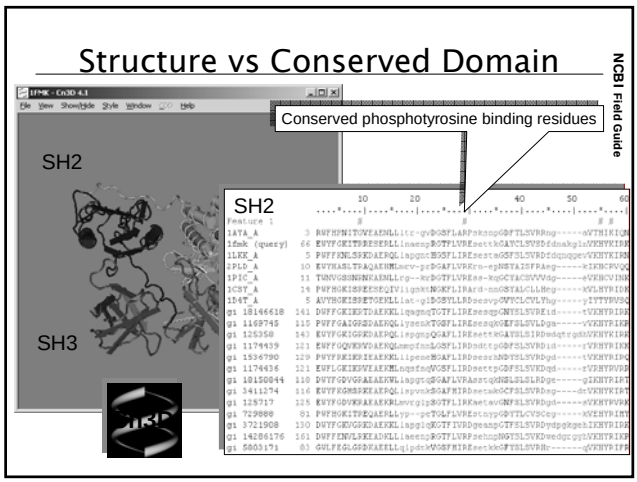
- PSI-BLAST –based score matrices
- Searchable with RPS-BLAST
- Sources
  - SMART
  - PFAM
  - COGs
  - NCBI curated domains
    - structure informed alignments

NCBI Field Guide

Src Domains



NCBI Field Guide


Structure vs Conserved Domain
NCBI Field Guide
Conserved phosphotyrosine binding residues
SH2
SH3
SH2