National Center for Biotechnology Information

By,

Kavisa Ghosh,

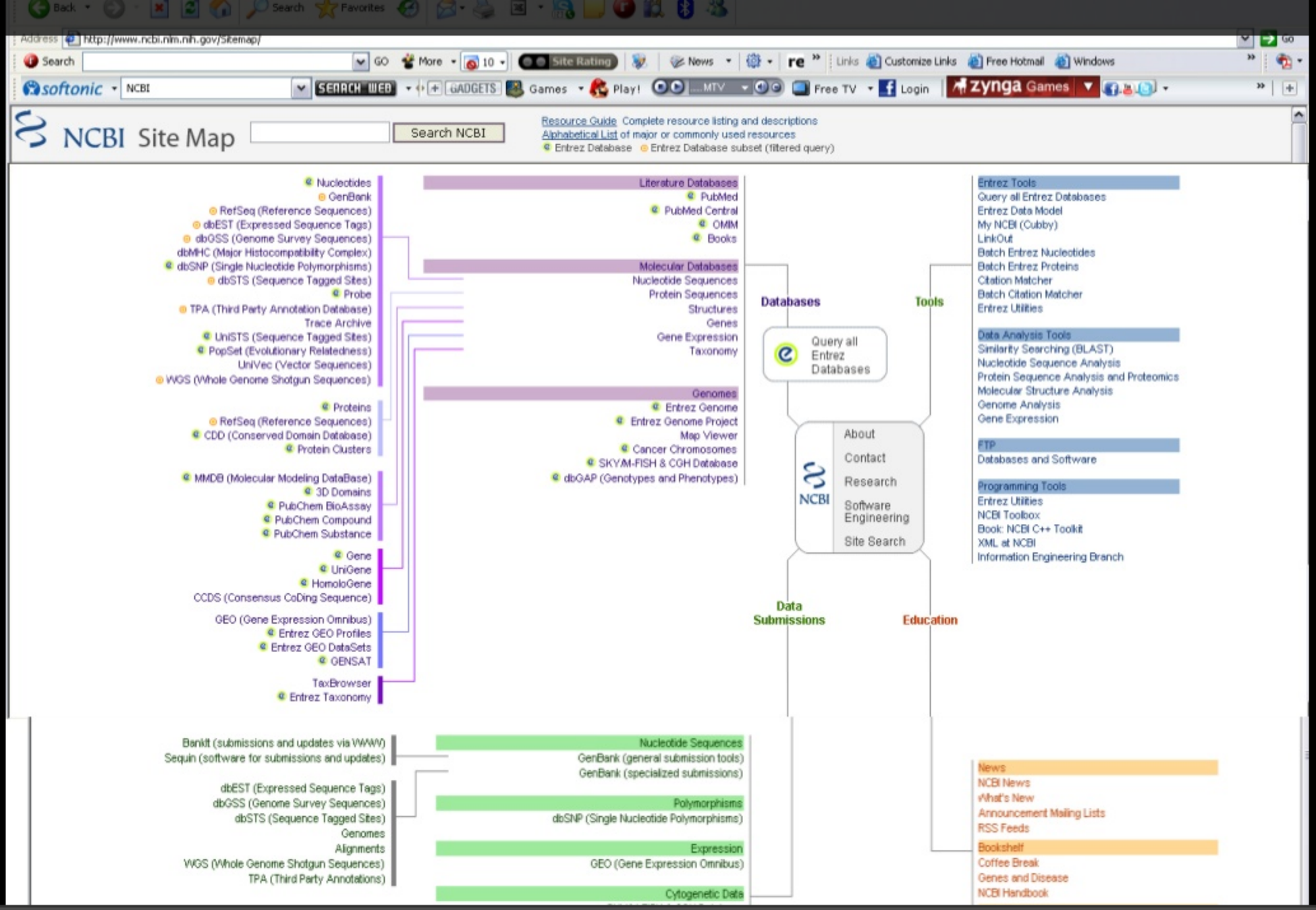
V M.Sc.Biotechnology(Int.)

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

NCBI HOME PAGE


http://www.ncbi.nlm.nih.gov/Sitemap/
NCBI Site Map
Search NCBI
Resource Guide Complete resource listing and descriptions
Alphabetical List of major or commonly used resources
Entrez Database
Entrez Database subset (filtered query)
Nucleotides
GenBank
RefSeq (Reference Sequences)
dbEST (Expressed Sequence Tags)
dbGSS (Genome Survey Sequences)
dbMHC (Major Histocompatibility Complex)
dbSNP (Single Nucleotide Polymorphisms)
dbSTS (Sequence Tagged Sites)
Probe
TPA (Third Party Annotation Database)
Trace Archive
UniSTS (Sequence Tagged Sites)
PopSet (Evolutionary Relatedness)
UniVec (Vector Sequences)
WGS (Whole Genome Shotgun Sequences)
Proteins
RefSeq (Reference Sequences)
CDD (Conserved Domain Database)
Protein Clusters
MMDB (Molecular Modeling DataBase)
3D Domains
PubChem BioAssay
PubChem Compound
PubChem Substance
Gene
UniGene
HomoloGene
CCDS (Consensus CoDing Sequence)
GEO (Gene Expression Omnibus)
Entrez GEO Profiles
Entrez GEO DataSets
GENSAT
TaxBrowser
Entrez Taxonomy
Literature Databases
PubMed
PubMed Central
OMIM
Books
Molecular Databases
Nucleotide Sequences
Protein Sequences
Structures
Genes
Gene Expression
Taxonomy
Genomes
Entrez Genome
Entrez Genome Project
Map Viewer
Cancer Chromosomes
SKY/M-FISH & CGH Database
dbGAP (Genotypes and Phenotypes)
Databases
Query all Entrez Databases
Tools
Entrez Tools
Query all Entrez Databases
Entrez Data Model
My NCBI (Cubby)
LinkOut
Batch Entrez Nucleotides
Batch Entrez Proteins
Citation Matcher
Batch Citation Matcher
Entrez Utilities
Data Analysis Tools
Similarity Searching (BLAST)
Nucleotide Sequence Analysis
Protein Sequence Analysis and Proteomics
Molecular Structure Analysis
Genome Analysis
Gene Expression
FTP
Databases and Software
Programming Tools
Entrez Utilities
NCBI Toolbox
Book: NCBI C++ Toolkit
XML at NCBI
Information Engineering Branch
NCBI
About
Contact
Research
Software Engineering
Site Search
Data Submissions
Education
Bankit (submissions and updates via WWW)
Sequin (software for submissions and updates)
dbEST (Expressed Sequence Tags)
dbGSS (Genome Survey Sequences)
dbSTS (Sequence Tagged Sites)
Genomes
Alignments
WGS (Whole Genome Shotgun Sequences)
TPA (Third Party Annotations)
Nucleotide Sequences
GenBank (general submission tools)
GenBank (specialized submissions)
Polymorphisms
dbSNP (Single Nucleotide Polymorphisms)
Expression
GEO (Gene Expression Omnibus)
Cytogenetic Data
News
NCBI News
What's New
Announcement Mailing Lists
RSS Feeds
Bookshelf
Coffee Break
Genes and Disease
NCBI Handbook

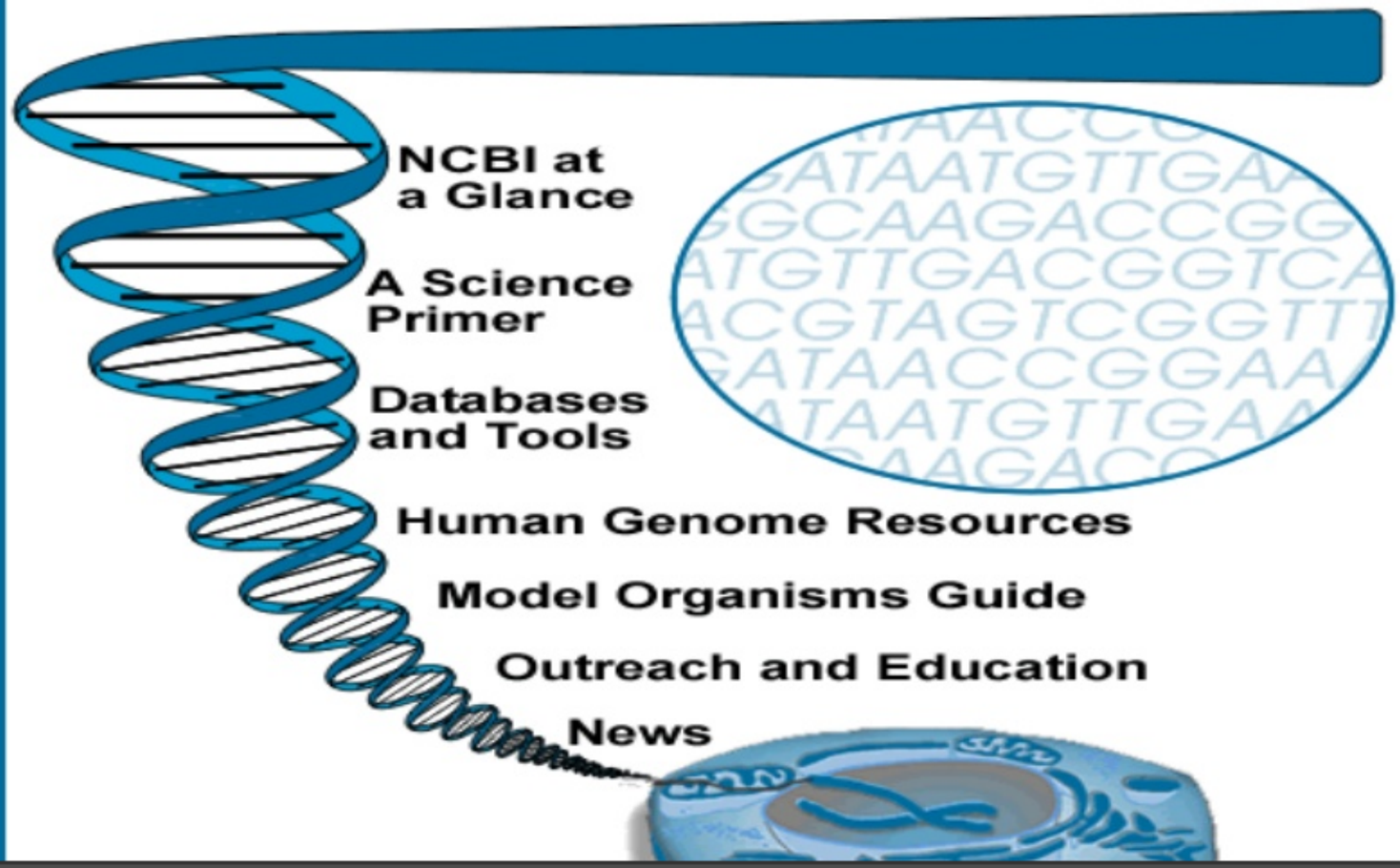
NCBI

About NCBI

National Center for Biotechnology Information

About NCBI Site Map

About NCBI	NCBI at a Glance	A Science Primer	Databases and Tools
Human Genome Resources	Model Organisms Guide	Outreach and Education	News

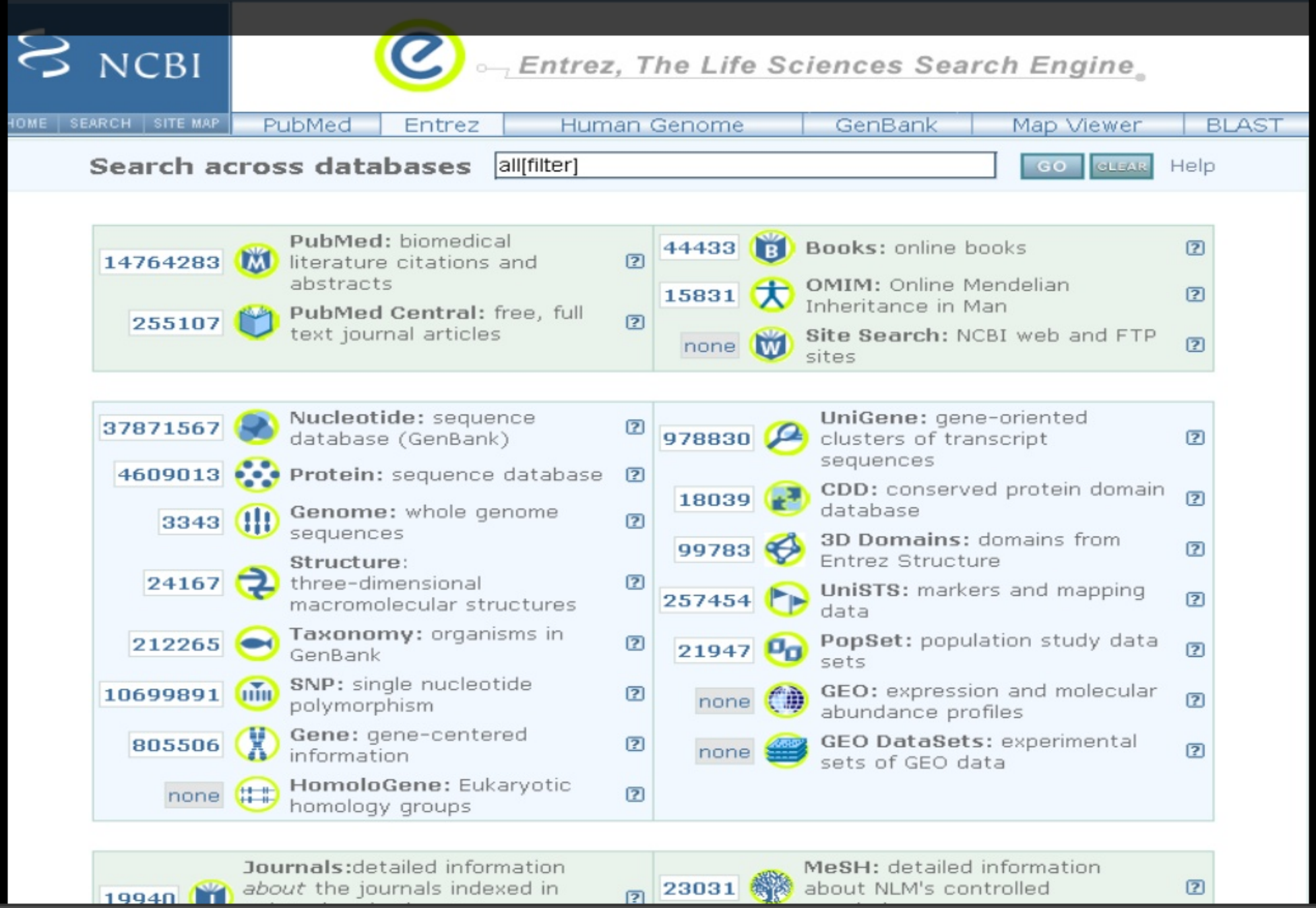


Entrez: An Integrated Database Search and Retrieval System

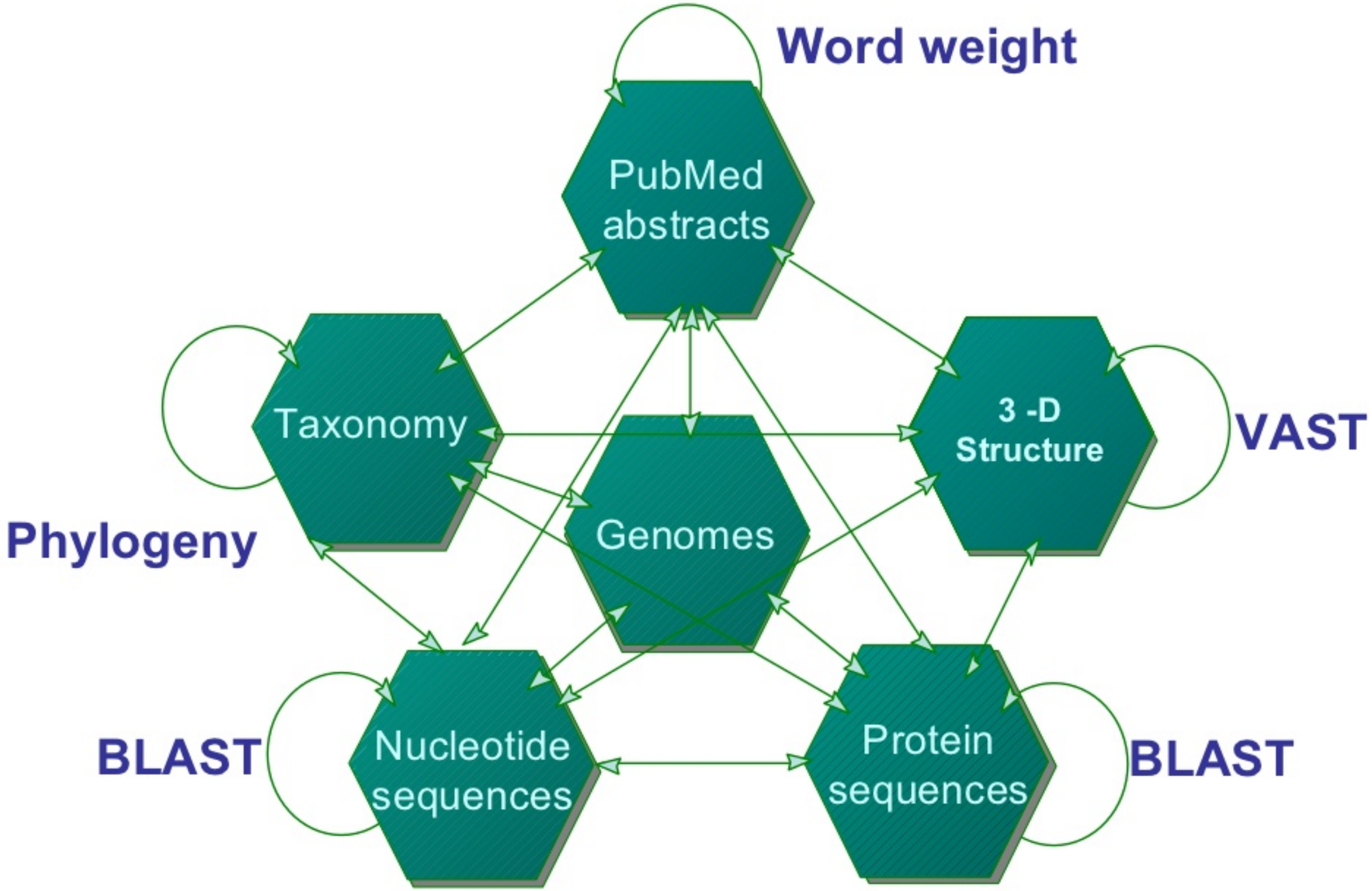
Entrez
search and retrieval system

Entrez is a retrieval system for searching several linked databases. It provides access to:

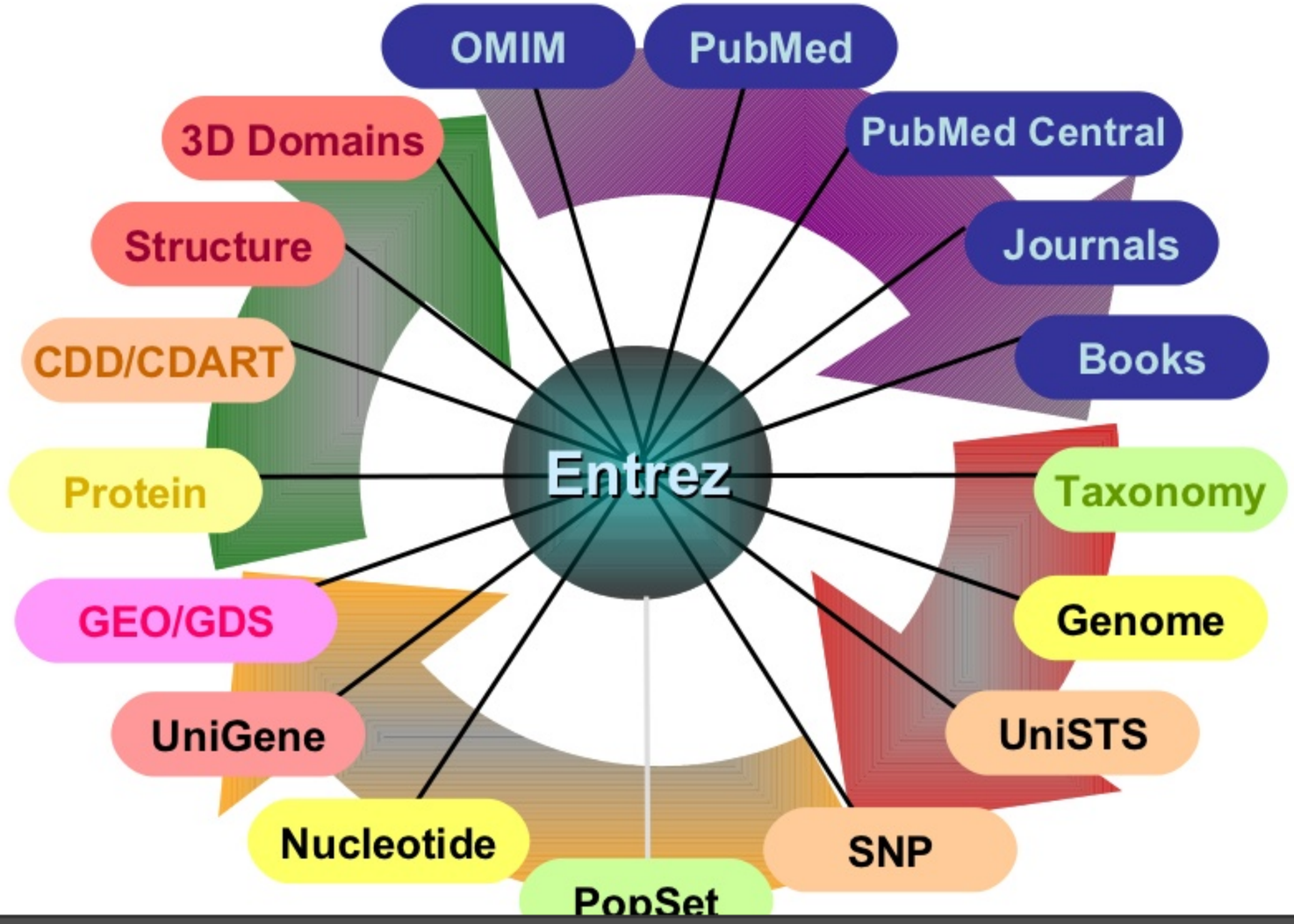
- PubMed: The biomedical literature (PubMed)
- Nucleotide: sequence database (GenBank)
- Protein: sequence database
- Structure: three-dimensional macromolecular structures
- Genome: complete genome assemblies
- PopSet: population study data sets
- OMIM: Online Mendelian Inheritance in Man
- Taxonomy: organisms in GenBank
- Books: BookShelf online books
- 3D Domains: domains from Entrez Structure
- UniSTS: markers and mapping data
- SNP: single nucleotide polymorphisms
- CDD: conserved domains
- Journals: journals in Entrez
- UniGene: gene-oriented clusters of transcript sequences
- PMC: full-text digital archive of life sciences journal literature


NCBI
Entrez, The Life Sciences Search Engine
HOME
SEARCH
SITE MAP
PubMed
Entrez
Human Genome
GenBank
Map Viewer
BLAST
Search across databases
all[filter]
GO
CLEAR
Help
14764283
PubMed: biomedical literature citations and abstracts
255107
PubMed Central: free, full text journal articles
44433
Books: online books
15831
OMIM: Online Mendelian Inheritance in Man
none
Site Search: NCBI web and FTP sites
37871567
Nucleotide: sequence database (GenBank)
4609013
Protein: sequence database
3343
Genome: whole genome sequences
24167
Structure: three-dimensional macromolecular structures
212265
Taxonomy: organisms in GenBank
10699891
SNP: single nucleotide polymorphism
805506
Gene: gene-centered information
none
HomoloGene: Eukaryotic homology groups
978830
UniGene: gene-oriented clusters of transcript sequences
18039
CDD: conserved protein domain database
99783
3D Domains: domains from Entrez Structure
257454
UniSTS: markers and mapping data
21947
PopSet: population study data sets
none
GEO: expression and molecular abundance profiles
none
GEO DataSets: experimental sets of GEO data
19940
Journals: detailed information about the journals indexed in
23031
MeSH: detailed information about NLM's controlled

Entrez: Database Integration



The *(ever expanding)* Entrez System



Entrez Databases

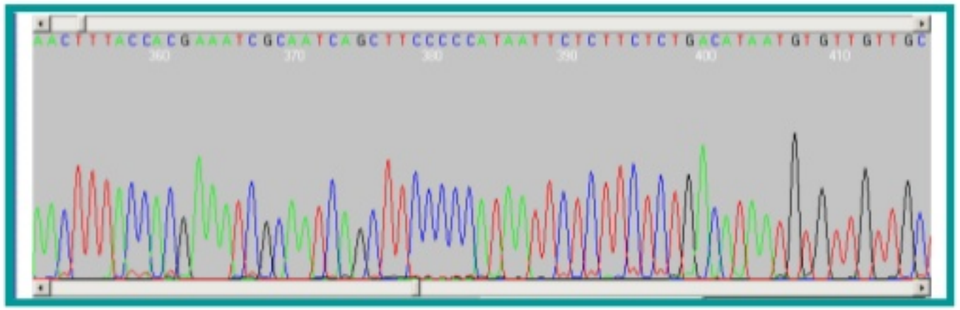
- All Molecular Database entries are organized by organism ***(Taxonomy Database)***.
- Each record is assigned a UID.
  - A “unique integer identifier” for internal tracking
- Each record is indexed by data fields.
  - [author], [title], [organism], and many others
- Each record is given a Document Summary.
  - a summary of the record’s content (DocSum)
- Each record is manually or computationally assigned links to biologically related UIDs in and across databases.

Literature Databases

- PubMed
- Books
- PubMed Central
- Journals
- On-Line Mendelian Inheritance in Man (OMIM)

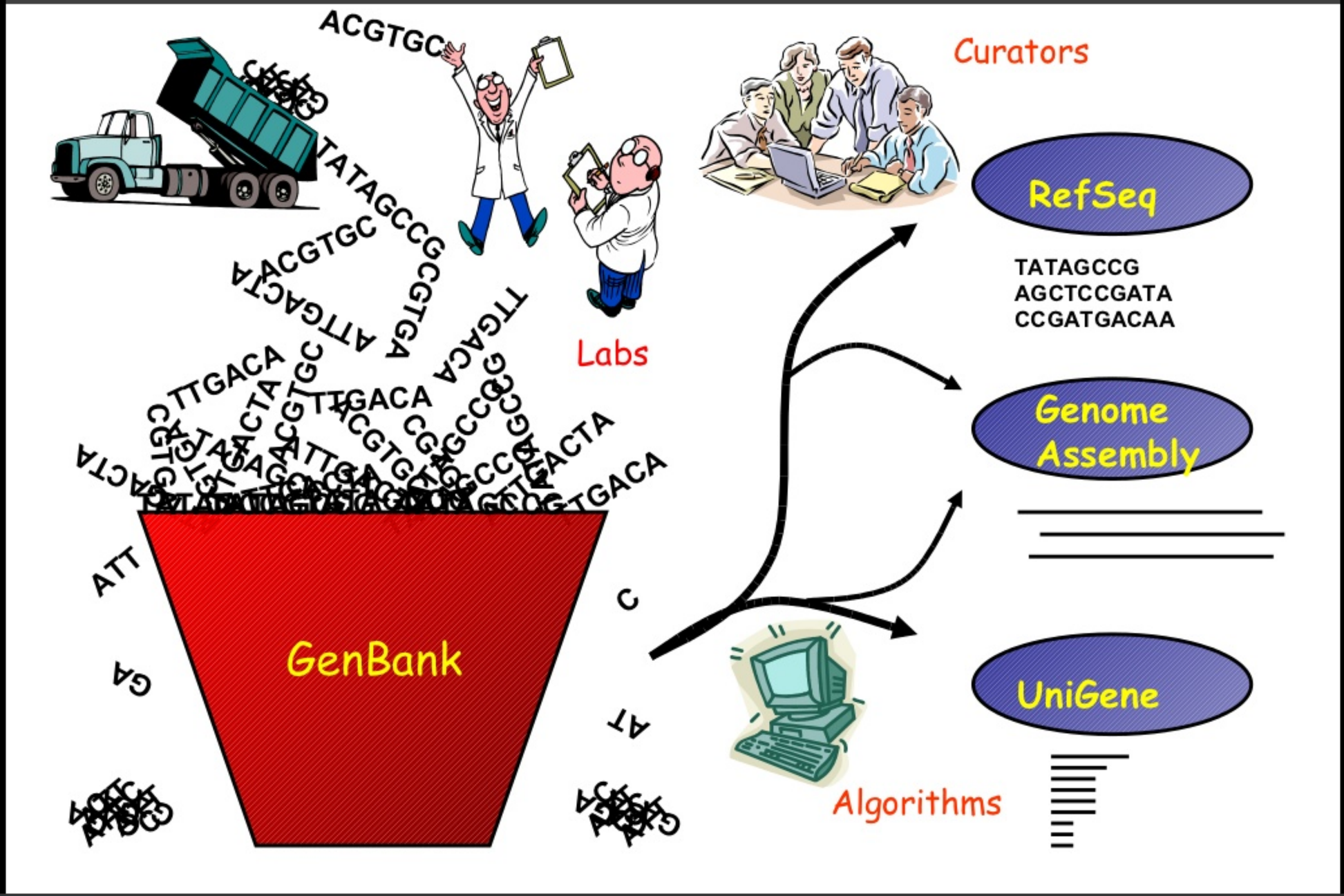
Molecular Sequence Databases

- Sequence Databases
  - Nucleotide (GenBank)
    - Taxonomy
    - PopSet
  - Protein
- Marker Databases
  - Single Nucleotide Polymorphisms (SNP's, dbSNP)
  - Sequence Tagged Sites (STS's, dbSTS)
  - Expressed Sequence Tags (EST's, dbEST)
    - UniGene



Molecular Databases

- Primary Databases
  - Original submissions by experimentalists
  - Database staff organize but don't add additional information
    - **Example: GenBank**
- Derivative Databases
  - Human curated
    - compilation and correction of data
    - **Example: SWISS-PROT, NCBI RefSeq mRNA**
  - Computationally Derived
    - **Example: UniGene**
  - Combinations
    - **Example: NCBI Genome Assembly**


ACGTGC
TATAGCCG
ACGTGC
ATTGACTA
TTGACA
Labs
Curators
RefSeq
TATAGCCG
AGCTCCGATA
CCGATGACAA
Genome Assembly
GenBank
UniGene
Algorithms

Derivative Databases

Sequencing Centers

GenBank

Updated ONLY by submitters

EST

STS

HTG

GSS

INV VRT PHG VRL

PRI ROD PLN MAM BCT

Updated by NCBI

UniGene

UniSTS

RefSeq

What is GenBank?

- 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

The Old Way


NUCLEOTIDE SEQUENCES 1984 PART 1
NUCLEOTIDE SEQUENCES 1984 PART 2
NUCLEOTIDE SEQUENCES 1985 VOLUME I
NUCLEOTIDE SEQUENCES 1985 VOLUME II
NUCLEOTIDE SEQUENCES 1985 VOLUME III
NUCLEOTIDE SEQUENCES 1985 VOLUME IV
NUCLEOTIDE SEQUENCES 1986/1987 VOLUME I
NUCLEOTIDE SEQUENCES 1986/1987
NUCLEOTIDE SEQUENCES 1986/1987
NUCLEOTIDE SEQUENCES 1986/1987
NUCLEOTIDE SEQUENCES 1986/1987
NUCLEOTIDE SEQUENCES 1986/1987 VOLUME VI
NUCLEOTIDE SEQUENCES 1986/1987
NUCLEOTIDE SEQUENCES 1986/1987 VOLUME VIII

GenBank: NCBI's Primary Sequence Database

Release 136	June 2003
25,592,865 *18,197,119(June 2002)*	Records
32,528,249,295 *22,616,937,182(June 2002)*	Nucleotides
110,000 +	Species

- full release every two months
- incremental and cumulative updates daily
- available only through internet

ftp://ftp.ncbi.nih.gov/genbank/
ftp://genbank.sdsc.edu/pub
ftp://bio-mirror.net/biomirror/genbank/

GenBank Continued…

- 106,533,156,756 bases in 108,431,692 sequence records in the traditional GenBank divisions
- 148,165,117,763 bases in 48,443,067 sequence records in the WGS division as of August 2009.

GenBank Divisions

PRI	(28)	Primate
ROD	(15)	Rodent
PLN	(20)	Plant and Fungal
BCT	(18)	Bacterial/Archeal
INV	(7)	Invertebrate
VRT	(7)	Other Vertebrate
VRL	(4)	Viral
MAM	(2)	Mammalian
PHG	(1)	Phage
SYN	(1)	Synthetic
ENV	(4)	Envir. samples
UNA	(1)	Unannotated

EST	(570)	Expressed Sequence Tag
GSS	(197)	Genome Survey Sequence
HTG	(88)	High Throughput Genomic
PAT	(27)	Patent
STS	(9)	Sequence Tagged Site

"Organismal"
(Traditional)

- Organized by taxonomy (sort of)
- Direct submissions (Sequin/Bankit)
- Accurate (~1 error per 10,000 bp)
- Well characterized

"Functional"
(Bulk)

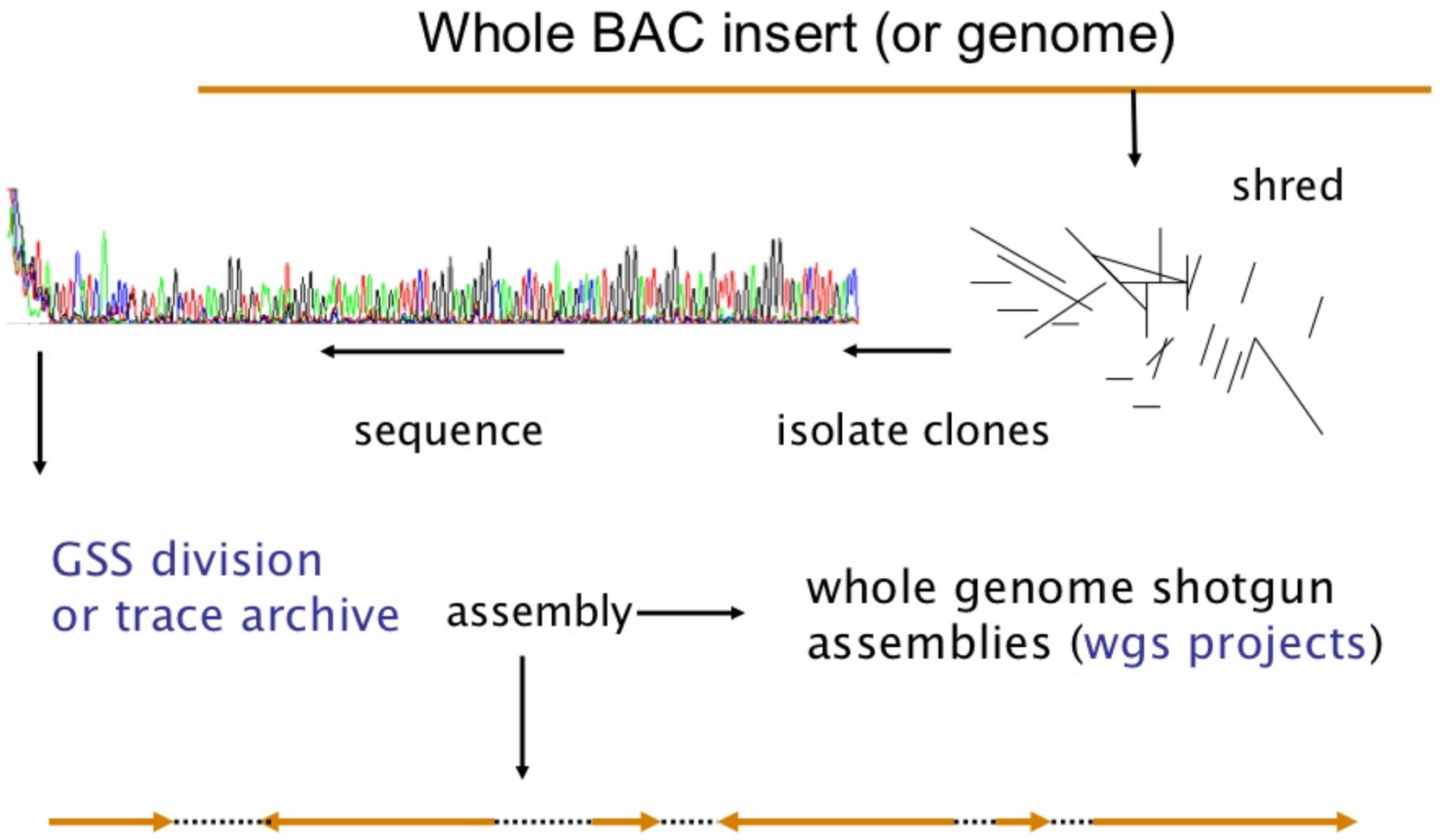
- Organized by sequence type
- Batch submissions (ftp/email)
- Less accurate

GenBank Functional (Bulk) Divisions

- **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

Whole Genome Shotgun

GSS, HTG, WGS


Whole BAC insert (or genome)
shred
sequence
isolate clones
GSS division
or trace archive
assembly
whole genome shotgun
assemblies (wgs projects)

Whole Genome Shotgun Projects

- 685 projects
  - Bacteria (320)
  - Environmental sequences (14)
  - Archaea (8)
  - Eukaryotes (140), including:
    - Chicken, Rat, Mouse, Dog (2), Chimpanzee, Human
    - Pufferfish (2)
    - Honeybee, Anopheles, Fruit Flies (3), Silkworm
    - Nematode (2)
    - Yeasts (8), Aspergillus (2)
    - Rice (2)

Whole Genome Shotgun (WGS) Projects

23: AACC00000000 Links
Homo sapiens chromosome 7, whole genome shotgun sequencing project
gi|50364594|gb|AACC00000000.2|AACC02000000[50364594]

wgs master[properties]

24: AACQ00000000 Links
Candida albicans SC5314, whole genome shotgun sequencing project
gi|46445633|gb|AACQ00000000.1|AACQ01000000[46445633]

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

25: AABT00000000
Aspergillus terreus ATCC 20542, whole genome shotgun sequencing project
gi|27262064|gb|AABT00000000.1|AABT01000000[27262064]

26: NZ_AAEZ00000000 Links
Pseudomonas syringae pv. phaseolicola 1448A, unfinished sequence, whole genome shotgun sequencing project
gi|50591998|ref|NZ_AAEZ00000000.1|NZ_AAEZ01000000[50591998]

27: NZ_AAFA00000000 Links
Streptococcus suis 89/1591, unfinished sequence, whole genome shotgun sequencing project
gi|50591969|ref|NZ_AAFA00000000.1|NZ_AAFA01000000[50591969]

28: AAFA00000000 Links
Streptococcus suis 89/1591, whole genome shotgun sequencing project
gi|50557642|gb|AAFA00000000.1|AAFA01000000[50557642]

NCBI ENTREZ Genome Project connection information discovery

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC

Search Genome Project for [ ] Go Clear

Limits | Preview/Index | History | Clipboard | Details

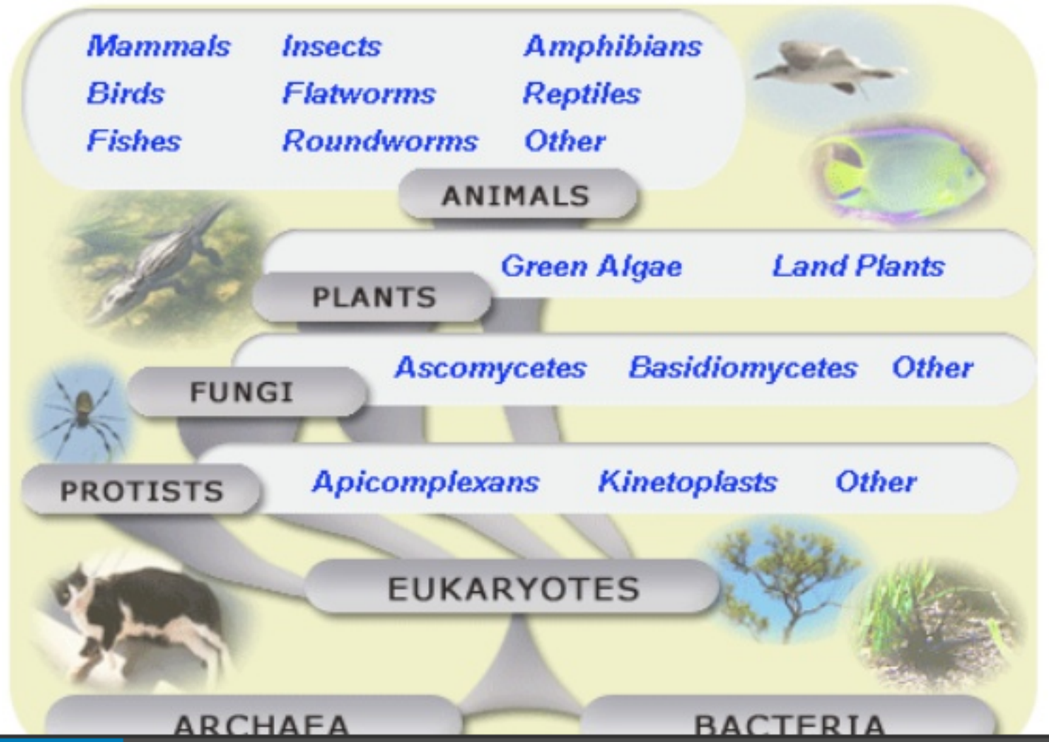
About Entrez

Entrez Genome Project
Home
Overview
Help
Statistics
Sequencing Centers

Submitting Project
Submissions
Project Instructions
General Genome Submissions
Feature Tables
Bacterial Genome Submissions
Whole Genome Shotgun Sequences

Related

Welcome to the NCBI Entrez Genome Project database.
This searchable database is a collection of complete and incomplete large-scale sequencing, assembly, annotation, and mapping projects for cellular organisms. The database is organized into organism-specific overviews that function as portals from which all projects in the database pertaining to that organism can be browsed and retrieved. Read more...



NCBI Resources

Entrez Gene
gene-related information
Entrez Genome
sequence and map data from whole genomes
Eukaryotic Projects
eukaryotic-specific genome projects
Genomic Biology
organism-specific links
Prokaryotic Projects
prokaryotic-specific genome projects
Organellar Genomes
organellar reference sequences and tools
Plant Genomes
major plant genome projects
RefSeq
the reference sequence project
Viral Genomes
viral reference sequences and tools
WGS Sequences

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 well-characterized genes and novel ESTs
- Useful for gene discovery and selection of mapping reagents

NCBI

UniGene
Homepage
Query Tips
FAQs
DDD
Download UniGene

Related Resources
HomoloGene
dbEST
Trace Archive
BLAST
CGAP
MGC cDNA clones

UniGene Dr.12379 *Danio rerio* il17rd

Links

Links
- Gene
- HomoloGene
- Nucleotide
- OMIM
- Protein
- Pubmed
- Unigene_homologous
- ZFIN

Interleukin 17 receptor D (il17rd)

SELECTED PROTEIN SIMILARITIES

Comparison of sequences in UniGene with proteins supported by a complete genome. The alignments can suggest function of a gene.

H. sapiens	pir:T42695 - T42695 hypothetical protein DKFZp434N1928.1 - human	55.56 % / 553 aa (see ProtEST)
M. musculus	ref:NP_032385.1 - interleukin 17 receptor [Mus musculus]	26.71 % / 269 aa (see ProtEST)

GENE EXPRESSION

Tissues and development stages from this gene's sequences survey gene expression. Links to other NCBI expression resources.

cDNA sources:	kidney, olfactory rosettes, ovary, testis, gastrula, larval, adult
Restricted Expression:	gastrula [Show more like this]
Expression Profile:	View expression levels using UniGene's EST ProfileViewer
ZFIN:	Gene Expression provided by the Zebrafish Information Network

MAPPING POSITION

Genomic location specified by transcript mapping, radiation hybrid mapping, genetic mapping or cytogenetic mapping.

UniGene

Species	UniGene Entries
Chordata	
Mammalia	
Bos taurus	23,572
Canis familiaris	3,905
Homo sapiens	105,651
Mus musculus	79,846
Rattus norvegicus	39,874
Sus scrofa	20,426
Aves	
Gallus gallus	12,220
Amphibia	
Xenopus laevis	22,129
Xenopus tropicalis	17,102
Actinopterygii	
Danio rerio	17,951
Oncorhynchus mykiss	13,458
Oryzias latipes	8,127
Salmo salar	1,033
Ascidiacea	
Ciona intestinalis	13,600
Echinodermata	
Echinoidea	
Strongylocentrotus purpuratus	2,592
Arthropoda	
Insecta	
Anopheles gambiae	14,085
Apis mellifera	4,756
Bombyx mori	2,008
Drosophila melanogaster	12,322
Nematoda	
Chromadorea	
Caenorhabditis elegans	15,655
Platyhelminthes	
Embryophyta	
Bryopsida	
Physcomitrella patens	6,946
Coniferopsida	
Pinus taeda	8,669
Eudicotyledons	
Arabidopsis thaliana	19,666
Glycine max	12,567
Helianthus annuus	1,897
Lactuca sativa	9,629
Lycopersicon esculentum	3,658
Medicago truncatula	5,398
Populus tremula x Populus tremuloides	2,711
Solanum tuberosum	5,476
Vitis vinifera	11,832
Liliopsida	
Hordeum vulgare	11,980
Oryza sativa	25,089
Saccharum officinarum	4,771
Sorghum bicolor	7,075
Triticum aestivum	24,543
Zea mays	12,941
Chlorophyta	
Chlorophycaea	
Chlamydomonas reinhardtii	5,874
Mycetozoa	
Dictyosteliida	
Dictyostelium discoideum	3,853
Apicomplexa	
Coccidia	
Toxoplasma gondii	6,001

genomes → transcripts → proteins

- non-redundant; best representative
- updates to reflect current sequence data and biology
- distinct, stable accession series

RefSeq: NCBI's Derivative Sequence Database

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

```
srcdb_refseq[Properties]
```

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

RefSeq Benefits

- 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

RefSeq Accession Numbers

Third Party Annotation (TPA) Database

- Annotations of existing GenBank sequences
- Allows for community annotation of genomes
- Direct submissions
  - BankIt
  - Sequin

`tpa[Properties]`

Other NCBI Databases

- **dbSNP:** nucleotide polymorphism
- **Geo:** Gene Expression Omnibus
  - microarray and other expression data
- **Gene:** gene records
  - Unifies LocusLink and Microbial Genomes
- **Structure:** imported structures (PDB)
  - Cn3D viewer, NCBI curation
- **CDD:** conserved domain database
  - Protein families (COGs and KOGs)
  - Single domains (PFAM, SMART, CD)

NCBI Protein Databases

- GenPept GenBank, EMBL, DDBJ CDS translations
- RefSeq mRNA based (NP_) and genome based (XP_)
- Swiss-Prot curated high quality protein reviews
- PIR protein information resource Georgetown University
- PRF protein resource foundation
- PDB Protein Databank sequences from structures

PDB PROTEIN DATA BANK → Structure

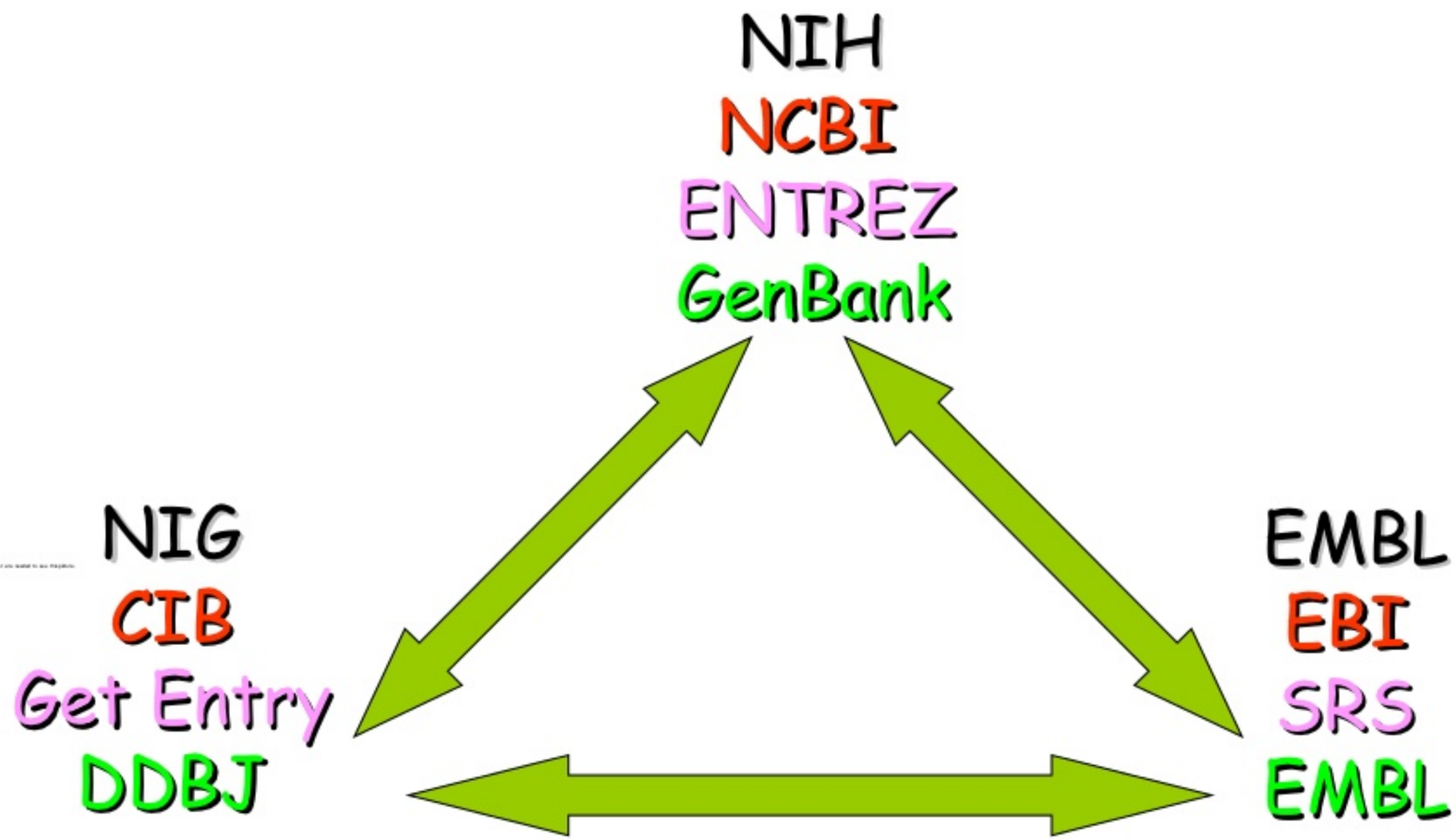
NCBI Structures and Domains

Pfam → CDD

SMART

Clusters of Orthologous Groups

The International Nucleotide Sequence Database Collaboration



Sequence formats

ASN.1
DNAStrider
EMBL
Fitch
GCG
GenBank/GB
IG/Stanford
MSF
NBRF
Olsen
PAUP/NEXUS
Pearson/Fasta
Phylip
PIR/CODATA
Plain/Raw
Pretty
Zuker

NOTE:

- FASTA is a popular sequence format

NCBI Nucleotide

PubMed | Nucleotide | Protein | Genome | Structure | PopSet | Taxonomy

Search Nucleotide for [ ] Go Clear

Limits | Preview/Index | History | Clipboard

Display GenBank Save Text Add to Clipboard Get Subsequence

1: AF467571. Hydrobia acuta ac...[gi:22416476]

```
LOCUS       AF467571                638 bp    DNA     linear   INV 22-AUG-2002
DEFINITION  Hydrobia acuta acuta isolate 1479 cytochrome c oxidase subunit I
            (COI) gene, partial cds; mitochondrial gene for mitochondrial
            product.
ACCESSION   AF467571
VERSION     AF467571.1  GI:22416476
KEYWORDS    .
SOURCE      Hydrobia acuta acuta.
  ORGANISM  Mitochondrion Hydrobia acuta acuta
            Eukaryota; Metazoa; Mollusca; Gastropoda; Caenogastropoda;
            Mesogastropoda; Rissooidea; Hydrobiidae; Hydrobia.
REFERENCE   1  (bases 1 to 638)
  AUTHORS   Wilke,T. and Pfenninger,M.
  TITLE     Separating historic events from recurrent processes in cryptic
            species: phylogeography of mud snails (Hydrobia spp.)
  JOURNAL   Mol. Ecol. 11 (8), 1439-1451 (2002)
   PUBMED   12144664
REFERENCE   2  (bases 1 to 638)
  AUTHORS   Wilke,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (11-JAN-2002) Department of Microbiology and Tropical
            Medicine, The George Washington University, 2300 Eye Street,
            Washington, DC 20037, USA
FEATURES             Location/Qualifiers
     source          1..638
                     /organism="Hydrobia acuta acuta"
                     /organelle="mitochondrion"
                     /isolate="1479"
                     /sub_species="acuta"
```

GenBank format

NCBI Entrez Nucleotide

PubMed | Nucleotide | Protein | Genome | Structure | PopSet | Taxono

Search Nucleotide for [ ] Go Clear

Limits | Preview/Index | History | Clipboard

Display FASTA Save Text Add to Clipboard Get Subsequence

1: AF467571. Hydrobia acuta ac...[gi:22416476]

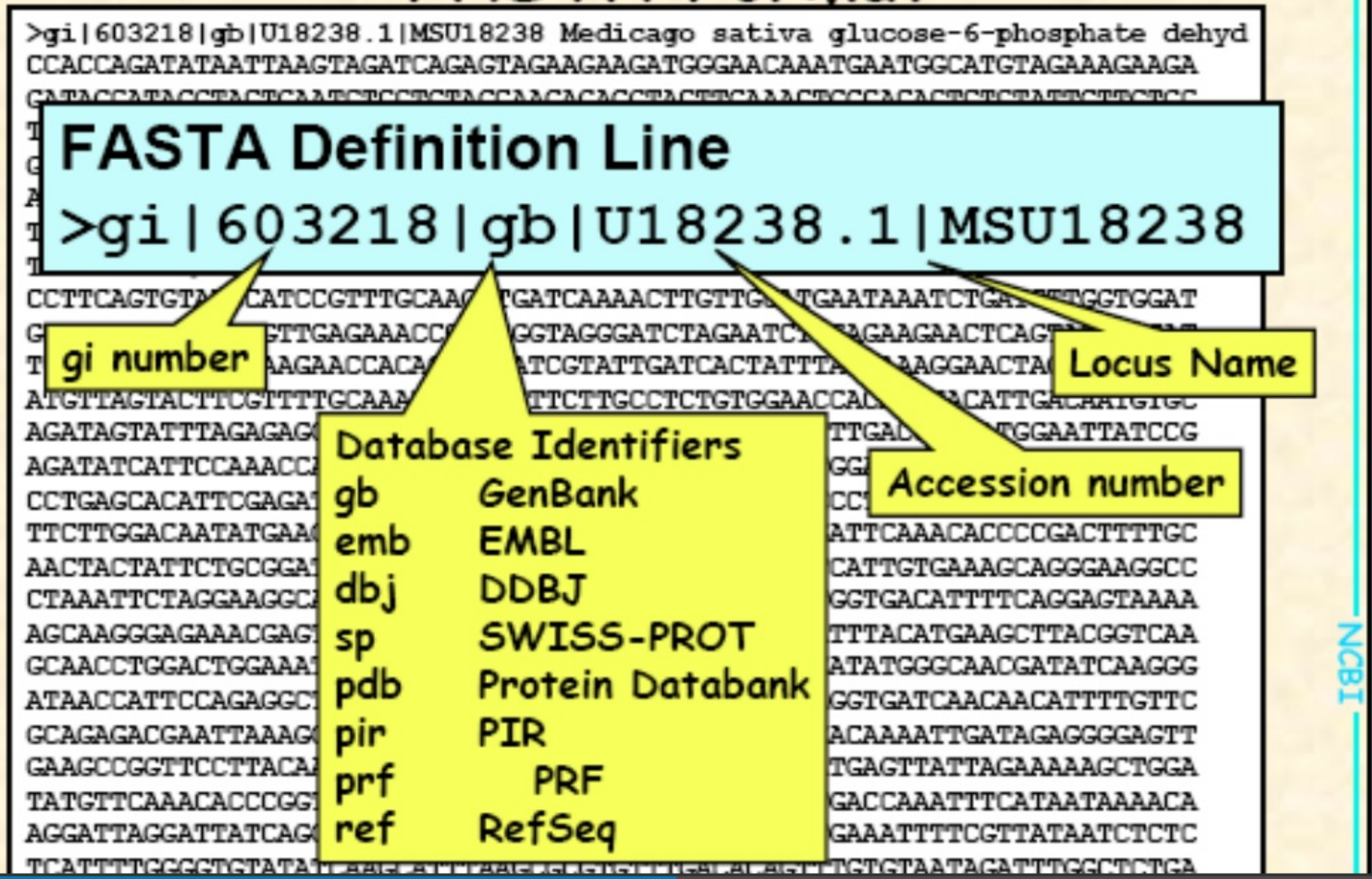
```
>gi|22416476|gb|AF467571.1| Hydrobia acuta acuta isolate 1479 cytochrome c oxidase subunit I (COI) gene, parti
ATTTTATTTGGTATGTGGTCTGGGTTAGTAGGTACAGCACTAAGTTTGTTAATTCGTGCTGAACTAGGTC
AGCCTGGTGCGCTTTTGGGTGATGATCAGCTTTATAACGTAATTGTTACTGCTCATGCCTTTGTTATAAT
TTTTTTTCTTGTAATGCCTATAATAATTGGTGGCTTTGGAAATTGATTAGTGCCTTTAATACTTGGTGCT
CCAGATATAGCTTTTCCTCGGCTTAATAACATAAGTTTCTGACTTTTACCTCCTGCTTTGCTATTATTAC
TTTCTTCGGCAGCTGTAGAGAGAGGAGCGGGGACAGGATGAACCGTGTATCCCCCATTATCTAGTAACAT
TGCTCACGCGGGGGGGTCTGTAGATTTAGCTATTTTTTCTCTCCACTTAGCGGGTGTTTCTTCTATTCTT
GGGGCTGTAAATTTTATTACAACTATCATTAATATACGGTGACGAGGAATGCAGTTTGAGCGGCTTCCGT
TGTTCGTATGATCTGTAAAAATTACTGCCATTCTATTATTACTATCTTTACCTGTCTTAGCTGGTGCTAT
TACTATGCTTTTAACGGATCGAAATTTTAATACTGCATTTTTCGACCCAGCAGGAGGTGGAGACCCTATT
TTATACCA
```

Revised: July 5, 2002.

Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

Fasta format

FASTA Format



Data Analysis Tools

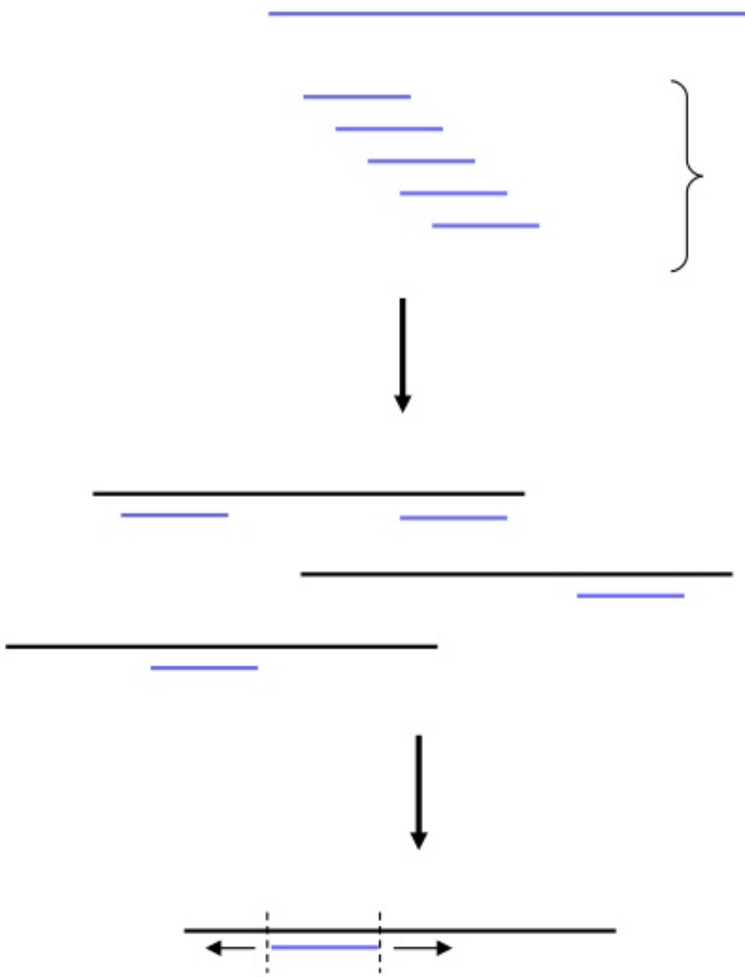
BLAST

*B*asic *L*ocal *A*lignment *S*earch *T*ool

Basic Local Alignment Search Tool

- Widely used similarity search tool
- Heuristic approach based on Smith Waterman algorithm
- Finds best local alignments
- Provides statistical significance
- All combinations (DNA/Protein) query and database.
  - DNA vs DNA
  - DNA translation vs Protein
  - Protein vs Protein
  - Protein vs DNA translation
  - DNA translation vs DNA translation
- www, standalone, and network clients

How BLAST works - pictoral



Query Sequence

"words" (subsequences of the query sequence)

Query words are compared to the database (target sequences) and exact matches identified

For each word match, alignment is extended in both directions to find alignments that score greater than some threshold (maximal segment

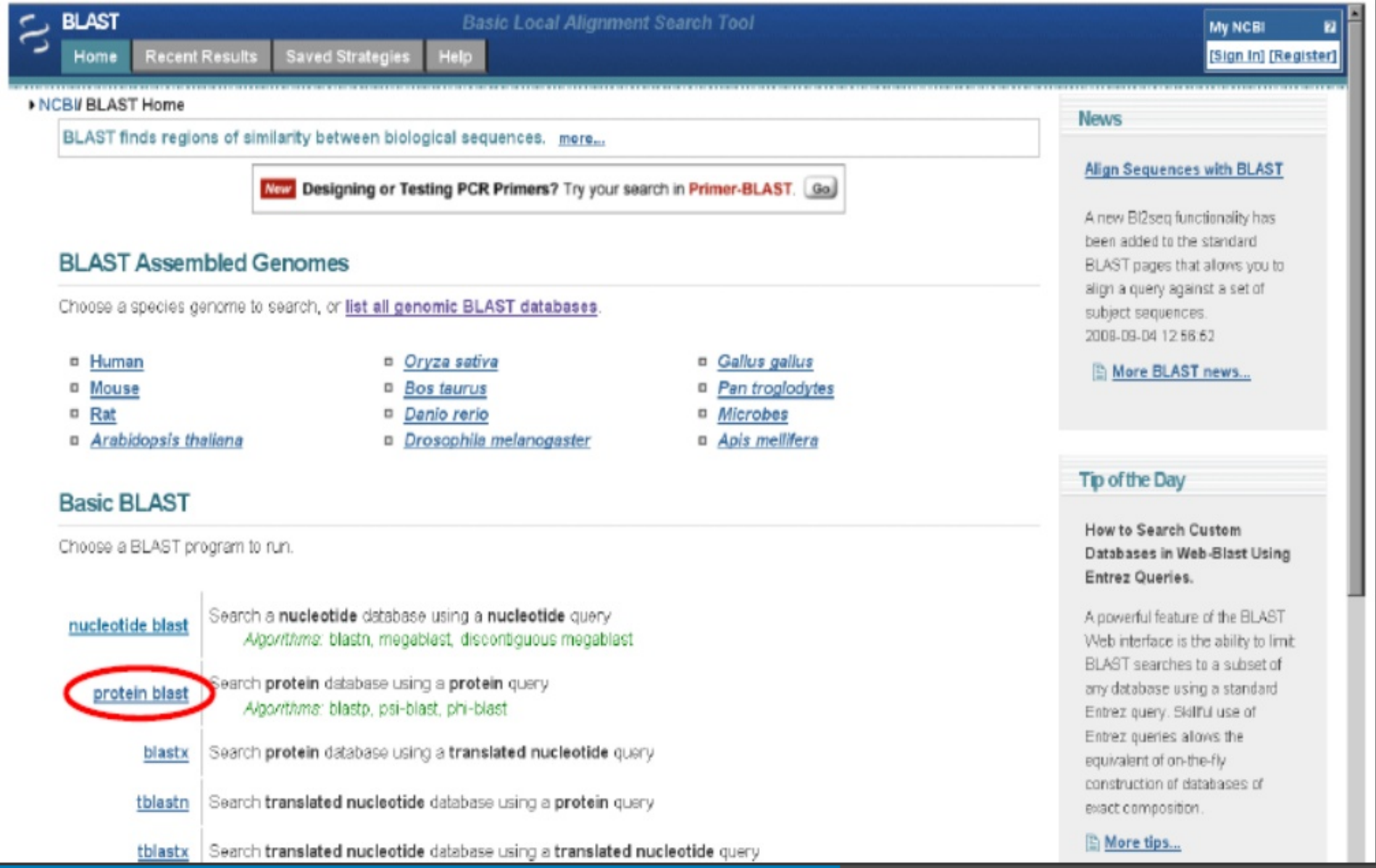
http://blast.ncbi.nlm.nih.gov/Blast.cgi

Basic BLAST

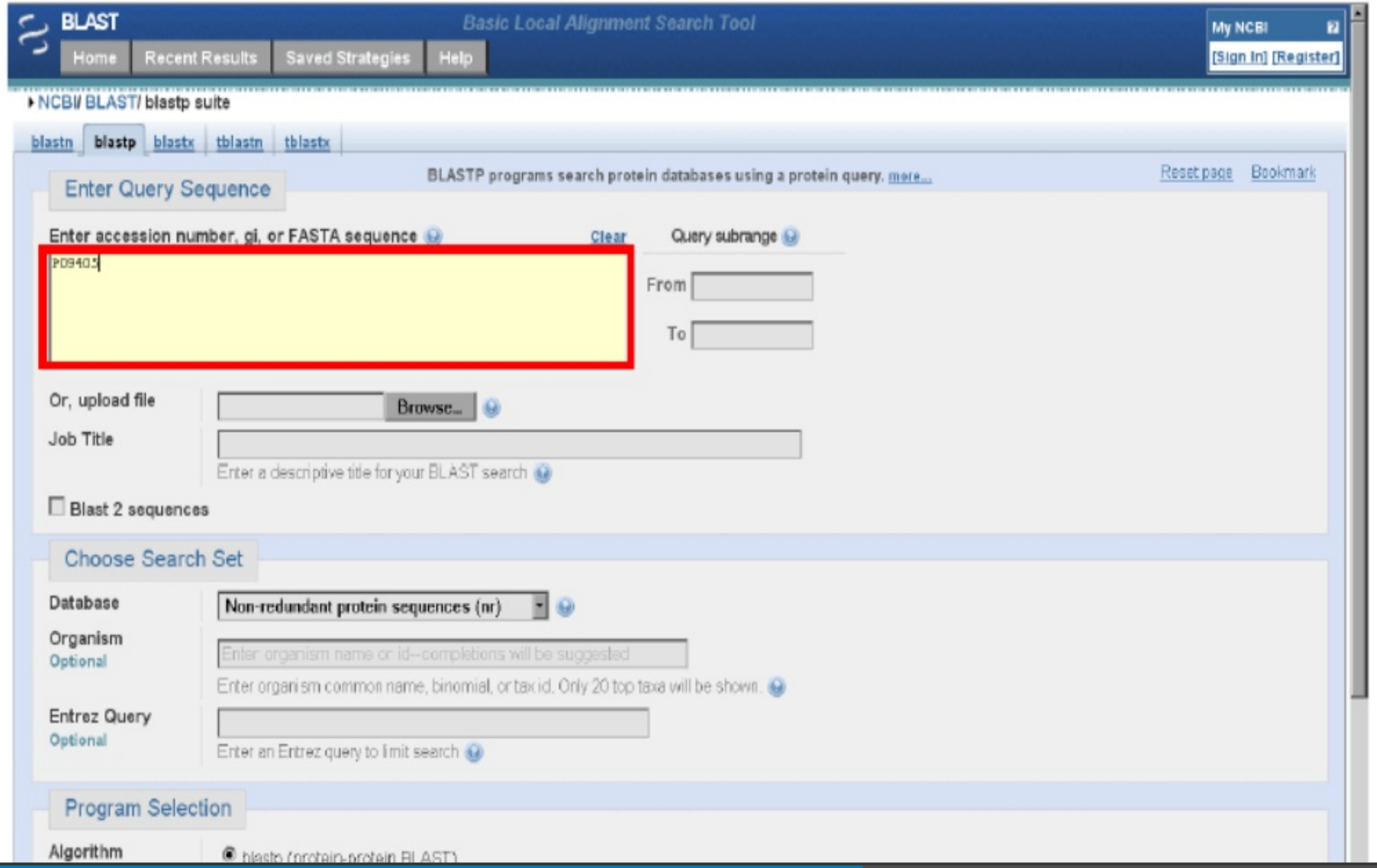
Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms*: blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms*: blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

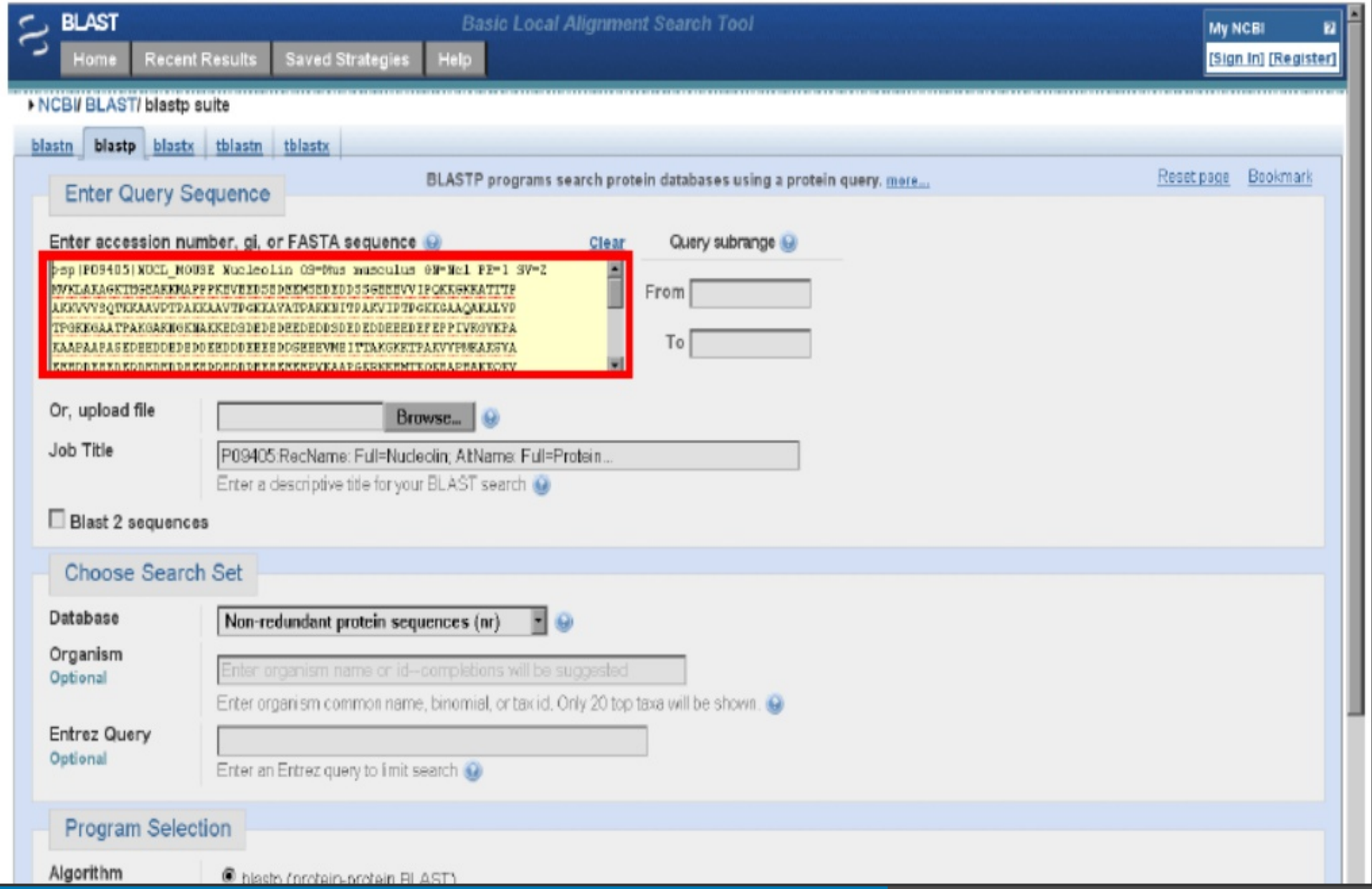
BLASTing a sequence at NCBI – programs



BLASTing a sequence at NCBI – enter accession



BLASTing a sequence at NCBI – enter sequence



BLASTing a sequence at NCBI – job status

BLAST Basic Local Alignment Search Tool

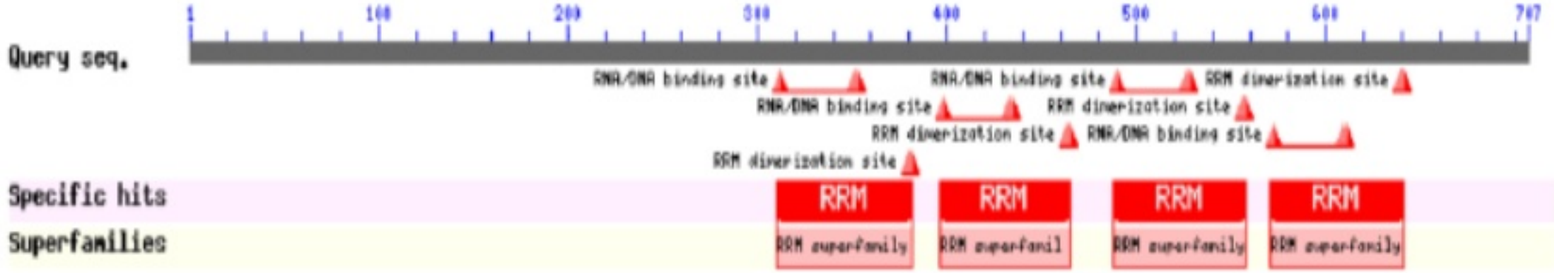
Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite/ Formatting Results - SVTZB9YT011 [Formatting options]

Job Title: gi|128843|sp|P09405.2|NUCL_MOUSE RecName:...

Putative conserved domains have been detected, click on the image below for detailed results.



Request ID	SVTZB9YT011
Status	Searching
Submitted at	Sat Feb 7 14:58:35 2009
Current time	Sat Feb 7 14:58:48 2009
Time since submission	00:00:13

This page will be automatically updated in **6** seconds

BLASTing a sequence at NCBI – blast summary

BLAST Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite/ Formatting Results - SVTZB9YT011

Edit and Resubmit | Save Search Strategies | ▶Formatting options | ▶Download

gi|128843|sp|P09405.2|NUCL_MOUSE RecName:...

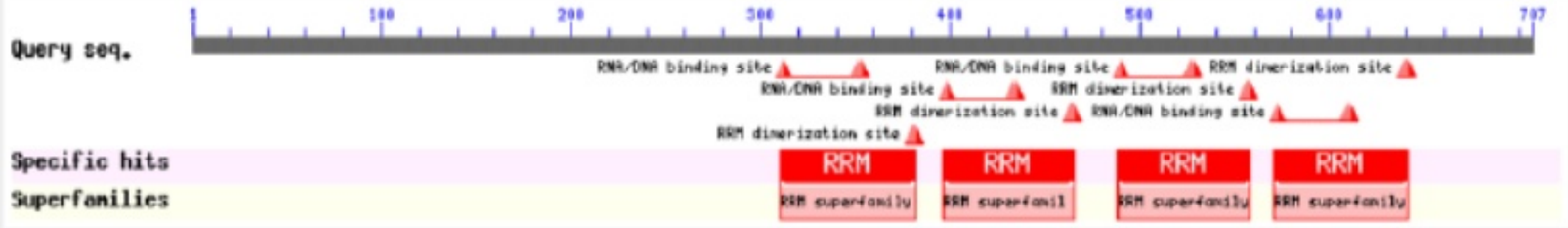
Query ID	lcl|22157	Database Name	nr
Description	gi|128843|sp|P09405.2|NUCL_MOUSE RecName: Full=Nucleolin; AltName: Full=Protein C23	Description	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type	amino acid	Program	BLASTP 2.2.19+ ▶Citation
Query Length	707		

Other reports: ▶Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures]

▼ Graphic Summary

▼ Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of 172 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

Color key for alignment scores

<40 | 40-50 | 50-80 | 80-200 | >=200

BLASTing a sequence at NCBI – used parameters

Other reports: ▼Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures]

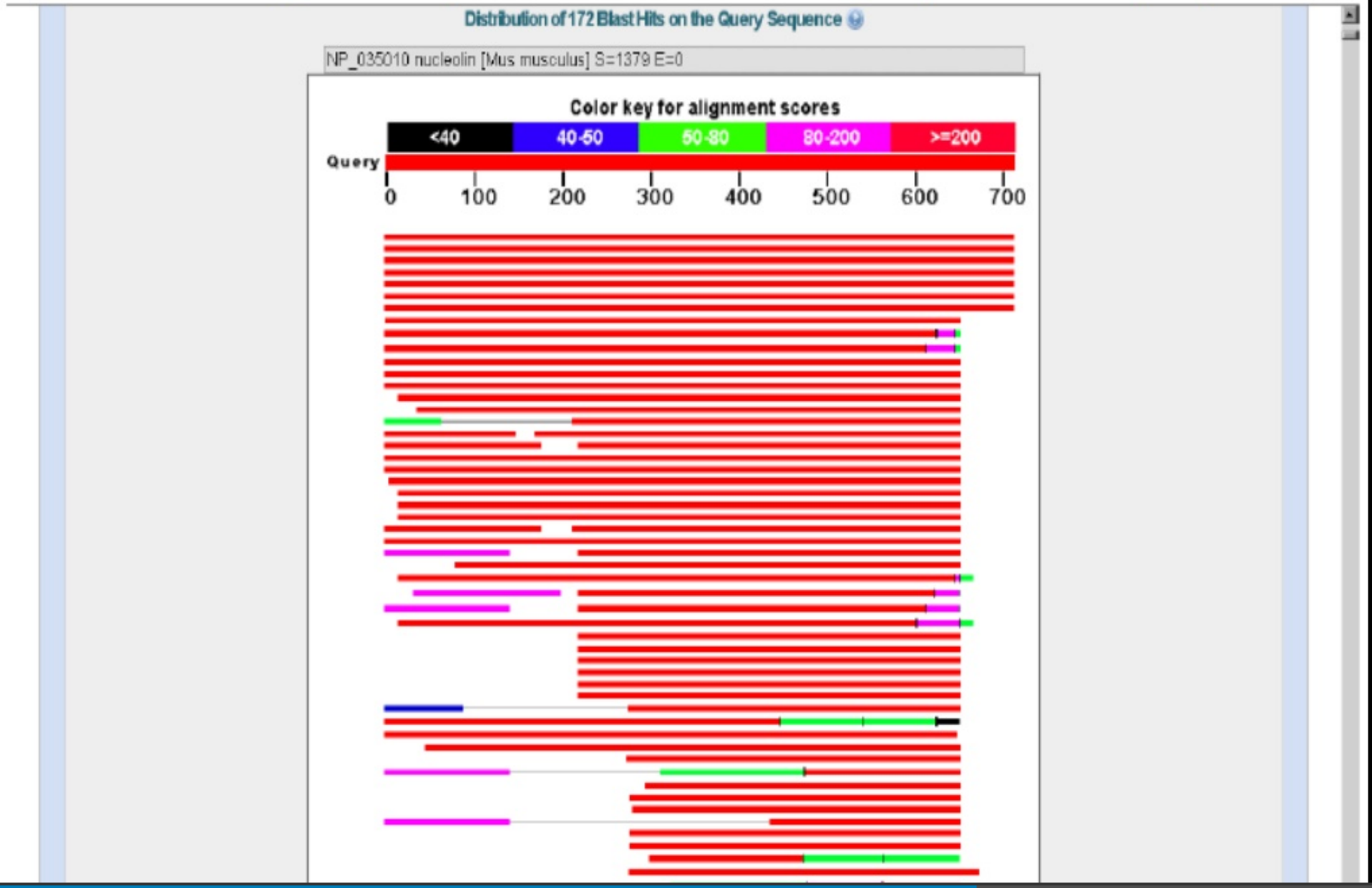
Search Parameters	
Program	blastp
Word size	3
Expect value	10
Hitlist size	100
Gapcosts	11,1
Matrix	BLOSUM62
Threshold	11
Composition-based stats	2
Filter string	F
Genetic Code	1
Window Size	40

Database	
Posted date	Feb 6, 2009 5:53 PM
Number of letters	2,699,408,701
Number of sequences	7,831,890
Entrez query	none

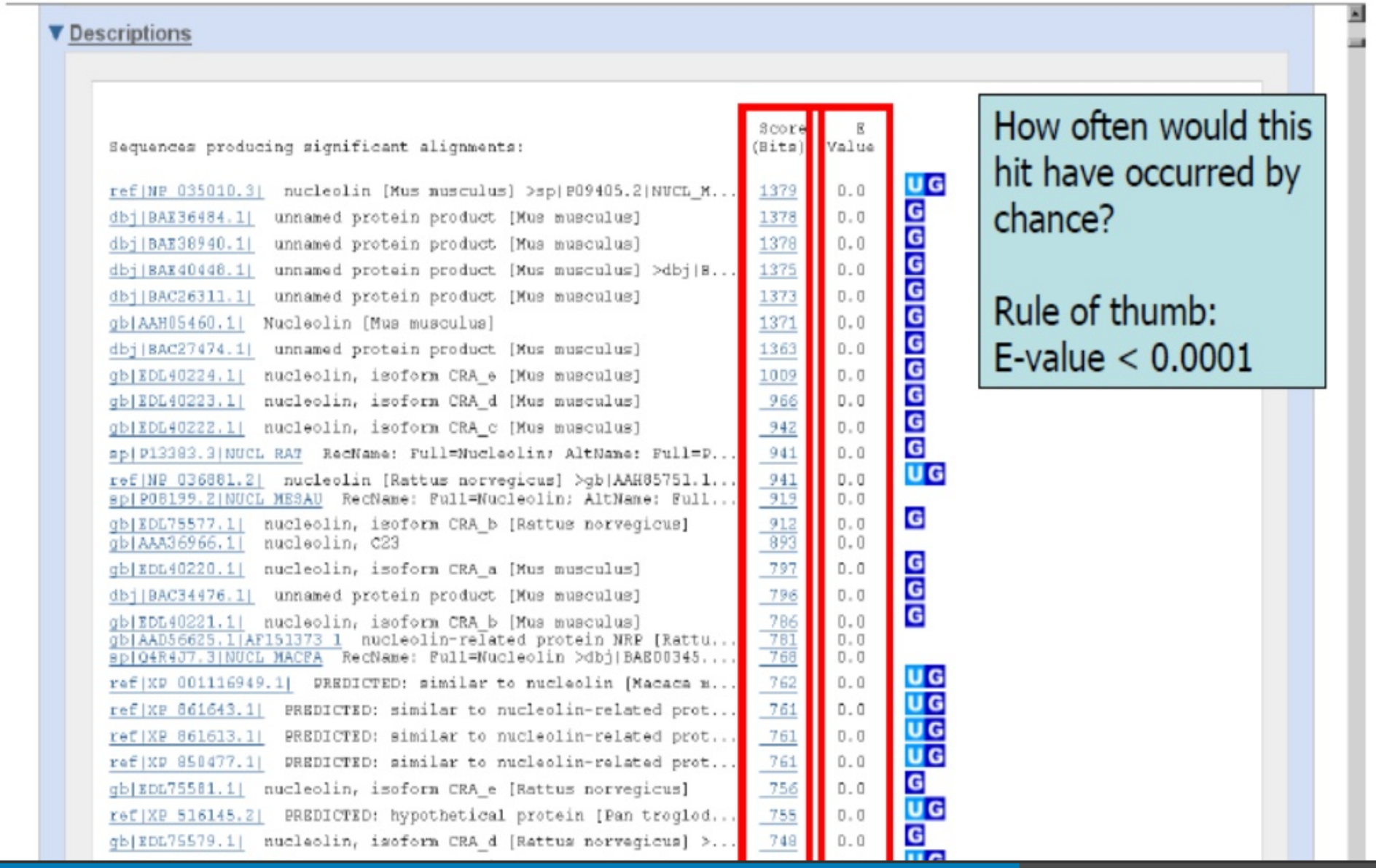
Karlin-Altschul statistics		
Params	Ungapped	Gapped
Lambda	0.302661	0.267
K	0.127079	0.041
H	0.344587	0.14

Results Statistics	
Length adjustment	143
Effective length of query	564
Effective length of database	1579448431

BLASTing a sequence at NCBI – graphical display



BLASTing a sequence at NCBI – hit list



▼ Descriptions

Sequences producing significant alignments:

	Score (Bits)	E Value
ref|NP_035010.3| nucleolin [Mus musculus] >sp|P09405.2|NUCL_M...	1379	0.0
dbj|BAE36484.1| unnamed protein product [Mus musculus]	1378	0.0
dbj|BAE38940.1| unnamed protein product [Mus musculus]	1378	0.0
dbj|BAE40448.1| unnamed protein product [Mus musculus] >dbj|B...	1375	0.0
dbj|BAC26311.1| unnamed protein product [Mus musculus]	1373	0.0
gb|AAH05460.1| Nucleolin [Mus musculus]	1371	0.0
dbj|BAC27474.1| unnamed protein product [Mus musculus]	1363	0.0
gb|EDL40224.1| nucleolin, isoform CRA_e [Mus musculus]	1009	0.0
gb|EDL40223.1| nucleolin, isoform CRA_d [Mus musculus]	966	0.0
gb|EDL40222.1| nucleolin, isoform CRA_c [Mus musculus]	942	0.0
sp|P13383.3|NUCL_RAT RecName: Full=Nucleolin; AltName: Full=P...	941	0.0
ref|NP_036881.2| nucleolin [Rattus norvegicus] >gb|AAH85751.1...	941	0.0
sp|P08199.2|NUCL_MESAU RecName: Full=Nucleolin; AltName: Full...	919	0.0
gb|EDL75577.1| nucleolin, isoform CRA_b [Rattus norvegicus]	912	0.0
gb|AAA36966.1| nucleolin, C23	893	0.0
gb|EDL40220.1| nucleolin, isoform CRA_a [Mus musculus]	797	0.0
dbj|BAC34476.1| unnamed protein product [Mus musculus]	796	0.0
gb|EDL40221.1| nucleolin, isoform CRA_b [Mus musculus]	786	0.0
gb|AAD56625.1|AF151373_1 nucleolin-related protein NRP [Rattu...	781	0.0
sp|Q4R4J7.3|NUCL_MACFA RecName: Full=Nucleolin >dbj|BAE00345....	768	0.0
ref|XP_001116949.1| PREDICTED: similar to nucleolin [Macaca m...	762	0.0
ref|XP_861643.1| PREDICTED: similar to nucleolin-related prot...	761	0.0
ref|XP_861613.1| PREDICTED: similar to nucleolin-related prot...	761	0.0
ref|XP_850477.1| PREDICTED: similar to nucleolin-related prot...	761	0.0
gb|EDL75581.1| nucleolin, isoform CRA_e [Rattus norvegicus]	756	0.0
ref|XP_516145.2| PREDICTED: hypothetical protein [Pan troglod...	755	0.0
gb|EDL75579.1| nucleolin, isoform CRA_d [Rattus norvegicus] >...	748	0.0

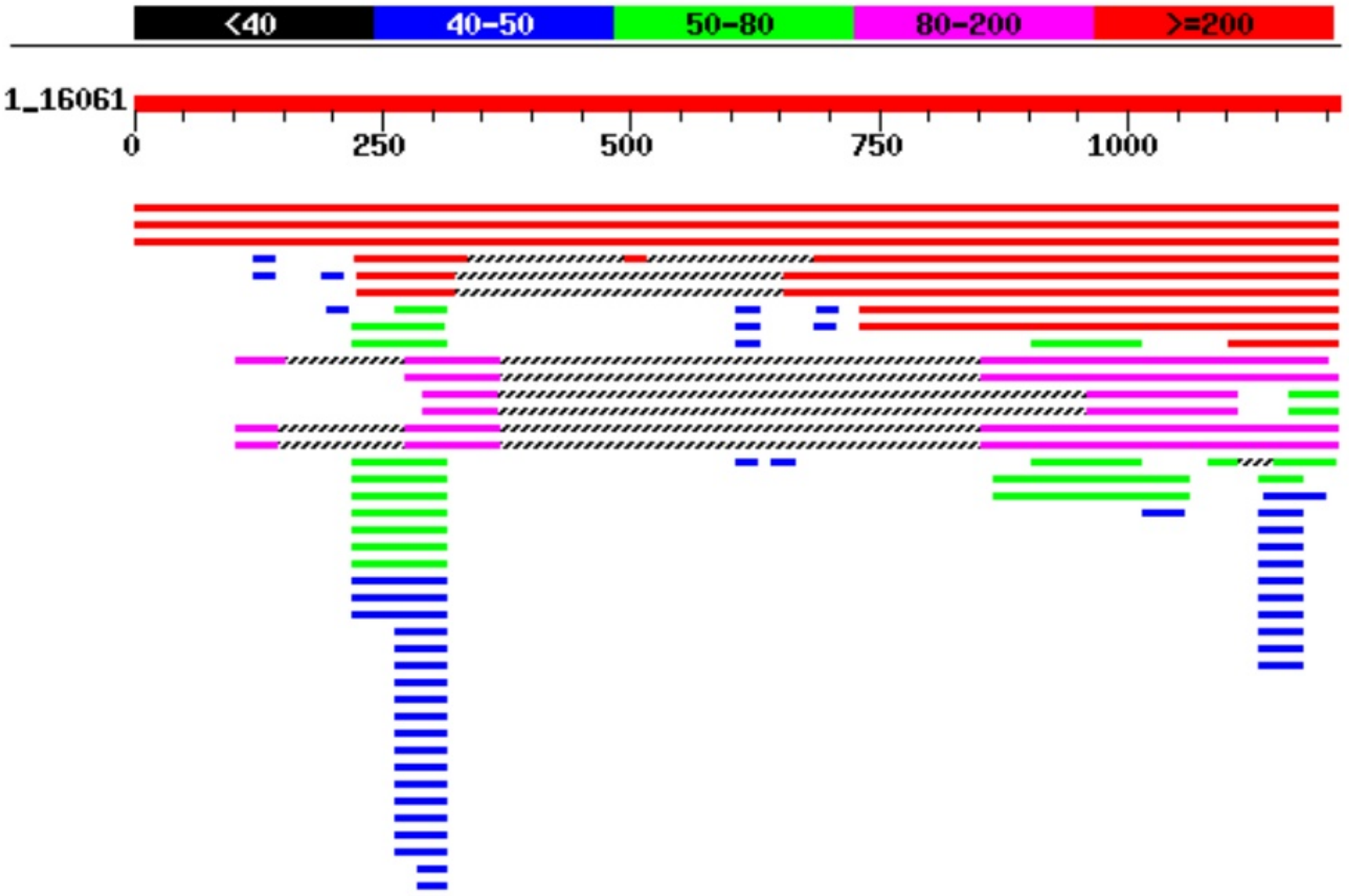
How often would this hit have occurred by chance?

Rule of thumb:
E-value < 0.0001

BLASTing a sequence at NCBI

```
>gb|AAF62554.1|  nucleolin [Oncorhynchus mykiss]
Length=255

 GENE ID: 100135911 LOC100135911 | nucleolin [Oncorhynchus mykiss]

 Score =  239 bits (610),  Expect = 7e-61, Method: Compositional matrix adjust.
 Identities = 133/260 (51%), Positives = 182/260 (70%), Gaps = 11/260 (4%)

Query  283  KKEMTKQKEAPEAKKQKVEGSEPTTPFNLFIGNLNSNKSVNELKFAISELFAKNDLAVVD  342
            K++   +KE P AKK K   SE    F LFIGNLN NK  +E+K A++  F+K +L V D
Sbjct  2    KRKADNKKETPDAKKAK---SESDDTFCLFIGNLNSNKDFDEIKEALAAFFSKKNLEVQD  58

Query  343  VRTGTNRRFGYVDFESAEDLEKALELTGLKVFGNEIKLEKPKGR----DSKKVRAARTLL  398
            VR G ++RFGYV+F SAED+  A+EL G K  G E+K++K + +     + KK R ARTL
Sbjct  59   VRLGASKRFGYVEFASAEDMQTAMELNGKKCMGQELKMDKARSKGNSQEEKKDRDARTLF  118

Query  399  AKNLSFNITEDELKEVFEDAMEIRL-VSQDGKSKGIAYIEFKSEADAEKNLEEKQGAEID  457
             KNL F+ TED+LKEVF +A+EIR+   QDG ++GIAYI FK+EA A+K L E QGA++
Sbjct  119  VKNLPFSATEDDLKEVFANAVEIRIPTGQDGSNRGIAYIAFKTEAMADKMLTEAQGADVQ  178

Query  458  GRSVSLYYTGEKGQRQERTGKTSTWSGESKTLVLSNLSYSATKETLEEVFEKATFIKVPQ  517
            GRS+ + YTG K Q+  R    +  + ESKTL+++NLSYSAT+++L+  FE A  I+VPQ
Sbjct  179  GRSIMVDYTGIKSQKGGRP--PAQAAAESKTLIVNNLSYSATEDSLQSAFEGAVSIRVPQ  236

Query  518  NPHGKPKGYAFIEFASFEDA  537
            N +G+PKG+AF+EF S E A
Sbjct  237  N-NGRPKGPAFVEFESAEXA  255


 Score = 99.8 bits (247),  Expect = 8e-19, Method: Compositional matrix adjust.
 Identities = 76/242 (31%), Positives = 118/242 (48%), Gaps = 29/242 (11%)

Query  396  TLLAKNLSFNITEDELKEVFE--------DAMEIRLVSQDGKSKGIAYIEFKSEADAEKN  447
             L   NL+ N   DE+KE            +  ++RL    G SK   Y+EF S  D +
Sbjct  26   CLPIGNLNSNKDPDEIKEALAAFFSKKNLEVQDVRL----GASKKFGYVEFASAEDMQTA  81

Query  448  LEEKQGAEIDGRSVSLYYTGEKGQRQERTGKTSTWSGESKTLVLSNLSYSATKETLEEVF  507
            +E   G +  G+ + +     KG  QE          +++TL + NL +SAT++ L+EVF
Sbjct  82   MELNGKKCMGQELKMDKARSKGNSQEEKKDR-----DARTLFVKNLPFSATEDDLKEVF  135

Query  508  EKATFIKVPQNPHGKPKGYAFIEFASFEDAKEAINSCNKMEIEGRTIRLELQGSNSR---  564
              A  I++P    G  +G A+I F + E A + +   + L      +++GR+I ++  G  S+
Sbjct  136  ANAVEIRIPTGQDGSNRGIAYIAFKTEAMADKMLTEAQGADVQGRSIMVDYTGIKSQKGG  195

Query  565  ------SQPSKTLFVKGLSEDTTEETLKESFEGSYRARIVTDRETGSSKGFGFYDFNSEE  618
                  + SKTL V  LS   TE++L+ +FEG+V  R+   +  G  KGF FV+F S E
Sbjct  196  RPPAQAAAESKTLIVNNLSYSATEDSLQSAFEGAVSIRV--PQNNGRDKGFAFVEFESAE  253
```

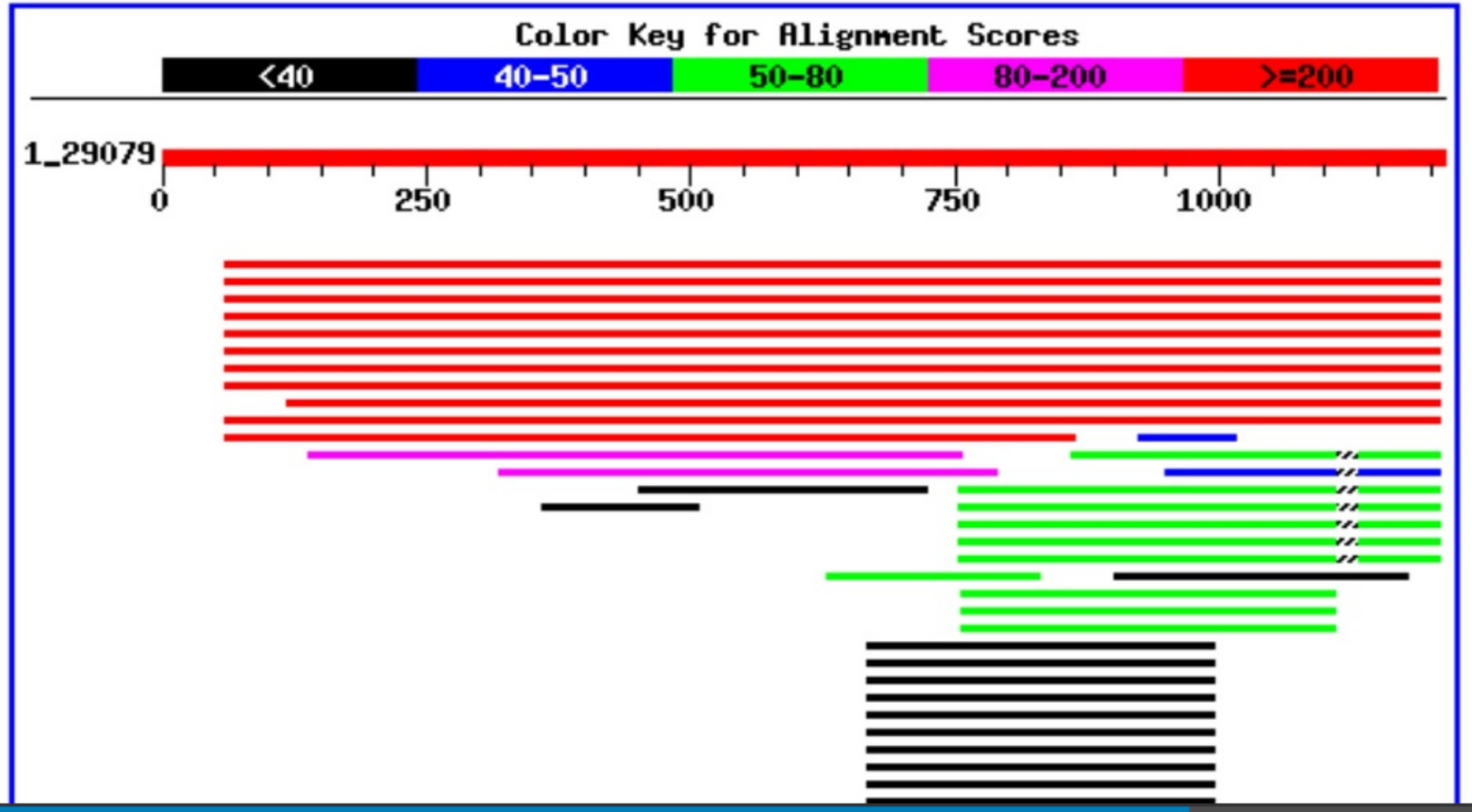
Result Page 1 of BLASTn



Result Page of BLASTx

Distribution of 59 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments



Result Page of BLASTx

Sequences producing significant alignments:	Score (bits)	E Value	
gi|51095067|gb|EAL24310.1| IMP (inosine monophosphate) dehy...	411	e-125	G
gi|57999523|emb|CAI45968.1| hypothetical protein [Homo sapi...	411	e-125	G
gi|51095068|gb|EAL24311.1| IMP (inosine monophosphate) dehy...	411	e-125	G
gi|25014074|sp|P20839|IMD1_HUMAN Inosine-5'-monophosphate d...	411	e-125	G
gi|106722|pir||A35566 IMP dehydrogenase (EC 1.1.1.205) I - ...	411	e-125	S
gi|13543973|gb|AAH06124.1| IMP (inosine monophosphate) dehy...	407	e-124	G S
gi|4504689|ref|NP_000875.1| IMP (inosine monophosphate) deh...	406	e-123	G
gi|307067|gb|AAA36114.1| IMP dehydrogenase type 1 (EC 1.1.1...	407	e-123	G
gi|47077068|dbj|BAD18464.1| unnamed protein product [Homo s...	397	e-120	G
gi|16549223|dbj|BAB70780.1| unnamed protein product [Homo s...	357	e-108	G
gi|51467033|ref|XP_496992.1| PREDICTED: similar to inosine ...	246	8e-65	G
gi|51467471|ref|XP_497019.1| PREDICTED: similar to Inosine-...	197	3e-50	G
gi|44979607|gb|AAS50155.1| IMP dehydrogenase 2 [Homo sapiens]	147	3e-35	G
gi|45708411|gb|AAH03053.1| GMPR2 protein [Homo sapiens]	76	3e-20	G

>gi|51095067|gb|EAL24310.1| IMP (inosine monophosphate) dehydrogenase 1 [Homo sapiens]
gi|34328930|ref|NP_000874.2| inosine monophosphate dehydrogenase 1 isoform a [Homo sapiens]
Length = 599

Score = 411 bits (1057), Expect(2) = e-125
Identities = 207/356 (58%), Positives = 271/356 (76%), Gaps = 4/356 (1%)
Frame = +1

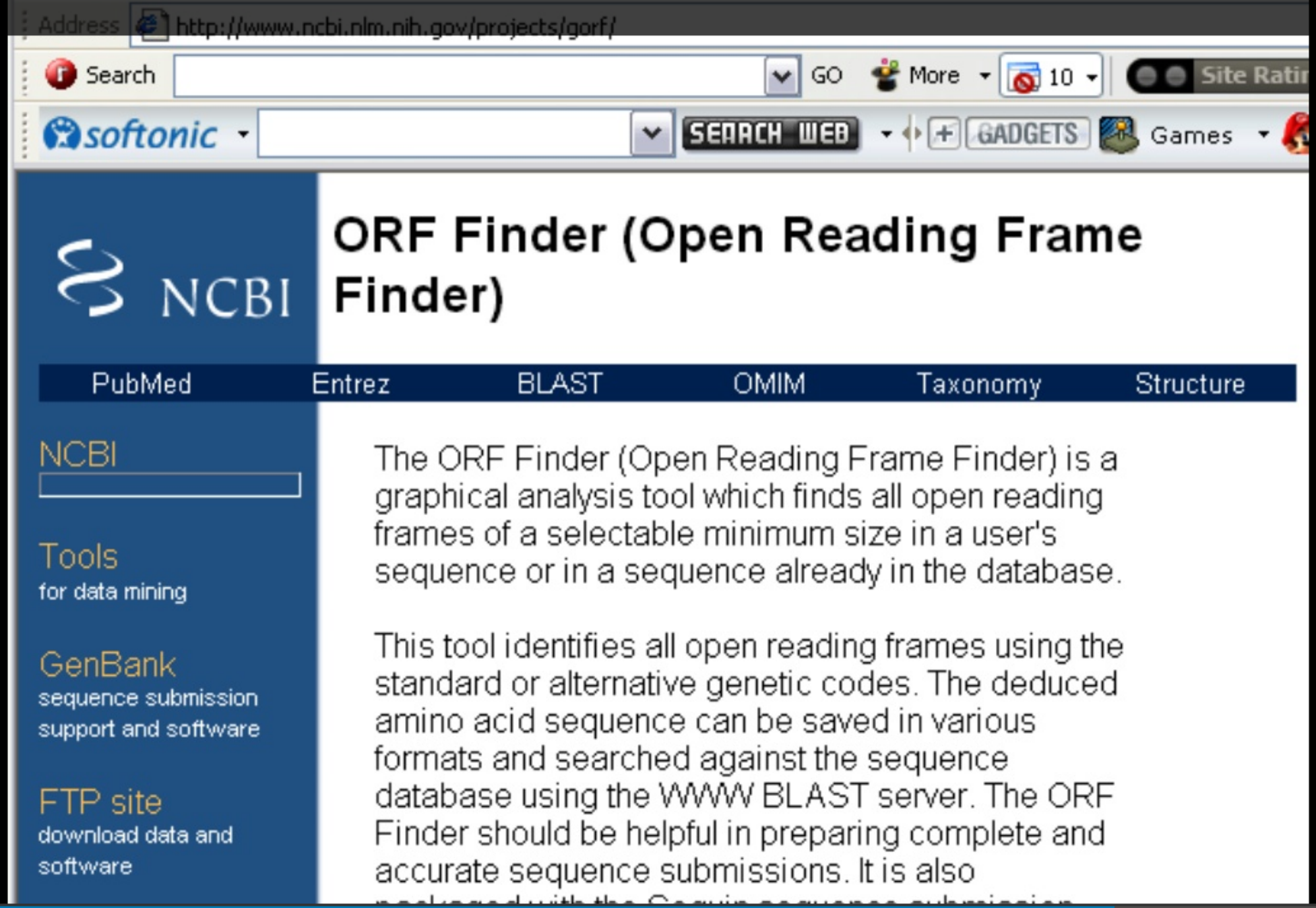
```
Query: 61   DGLSVQELMDSKIRGGLAYNDFLILPGLVDFASSEVSLQTKLTRNITLNIPLVSSPMDTV 240
            DGL+ Q+L  S   GL YNDFLILPG +DF + EV L + LTR ITL  PL+SSPMDTV
Sbjct: 101  DGLTAQQLFASA--DGLTYNDFLILPGFIDFIADEVDLTSALTRKITLKTPLISSPMDTV 158

Query: 241  TESEMATFMALLDGIGFIHHNCTPEDQADMVRRVKNYENGFINNPIVISPTTTVGEAKSM 420
            TE++MA  MAL+ GIGFIHHNCTPE QA+ VR+VK +E GFI +P+V+SP+ TVG+
Sbjct: 159  TEADMAIAMALMGGIGFIHHNCTPEFQANEVRKVKKFEQGFITDPVVLSPSHTVGDVLEA 218

Query: 421  KEKYGFAGFPVTADGKRNAKLVGAITSRDIQFV--EDNSLLVQDVMTK--NPVTGAQGIT 588
            K ++GF+G P+T  G   +KLVG +TSRDI F+  +D++ L+ +VMT     V   G+T
Sbjct: 219  KMRHGFSGIPITETGTMGSKLVGIVTSRDIDFLAEKDHTTLLSEVMTPRIELVVAPAGVT 278

Query: 589  LSEGNEILKKIKKGRLLVVDEKGNLVSMLSRTDLMKNQKYPLASKSANTKQLLWGASIGT 768
            L E NEIL++ KKG+L +V++   LV++++RTDL KN+ YPLASK +  KQLL GA++GT
Sbjct: 279  LKEANEILQRSKKGKLPIVNDCDELVAIIARTDLKKNRDYPLASKDSQ-KQLLCGAAVGT 337

Query: 769  MDADKERLRLLVKAGLDVVILDSSQGNSIFQLNMIKWIKETFPDLEIIAGNVVTKEQAAN 948
            + DK RL LL +AG+DV++LDSSQGNS++Q+ M+ +IK+ +P L++I GNVVT  QA N
Sbjct: 338  REDDKYRLDLLTQAGVDVIVLDSSQGNSVYQIAMVHYIKQKYPHLQVIGGNVVTAAQAKN 397

Query: 949  LIAAGADGLRIGMGTGSICITQKVMACGRPQGTAVYNVCEFANQFGVPCMADGGVQ 1116
            LI AG DGLR+GMG GSICITQ+VMACGRPQGTAVY V E+A +FGVP +ADGG+Q
```

Score = 46.2 bits (108), Expect(2) = e-125
Identities = 21/32 (65%), Positives = 28/32 (87%)
Frame = +2

Query: 1115 KNIGHIITKALALGSSTVMMGGMLAGTTESPG 1210
+ +GH++ KALALG+STVMMG +LA TTE+PG

Sbjct: 453 QTVGHVV-KALALGASTVMMGSLLAATTEAPG 483

Fig 5.14: Result Page of Pairwise ALignment


Address
http://www.ncbi.nlm.nih.gov/projects/gorf/
Search
GO
More
10
softonic
SEARCH WEB
GADGETS
Games
NCBI
ORF Finder (Open Reading Frame Finder)
PubMed
Entrez
BLAST
OMIM
Taxonomy
Structure
NCBI
Tools
for data mining
GenBank
sequence submission
support and software
FTP site
download data and
software
The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database.
This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also

Enter GI or ACCESSION

OrfFind Clear

or sequence in FASTA format

FROM: **TO:**

Genetic codes

1 Standard

Comments and suggestions to:
info@ncbi.nlm.nih.gov

Enter GI or ACCESSION 31343070

OrfFind Clear

or sequence in FASTA format

FROM: **TO:**

Genetic codes

1 Standard

Comments and suggestions to:
info@ncbi.nlm.nih.gov

Enter GI or ACCESSION

OrfFind Clear

or sequence in FASTA format

```
GC
CACTTCTGAGTTGGGGCAGCGGGTTCTAGCTCAG
CTCATGCTGAGAATGTAAGAACTACAAACAAAAT
TT
CTATTAAATTAAGTTTTGTGTCTTAAAAAAAAAA
AAAAAA
```

FROM: **TO:**

Genetic codes

1 Standard

Comments and suggestions to:

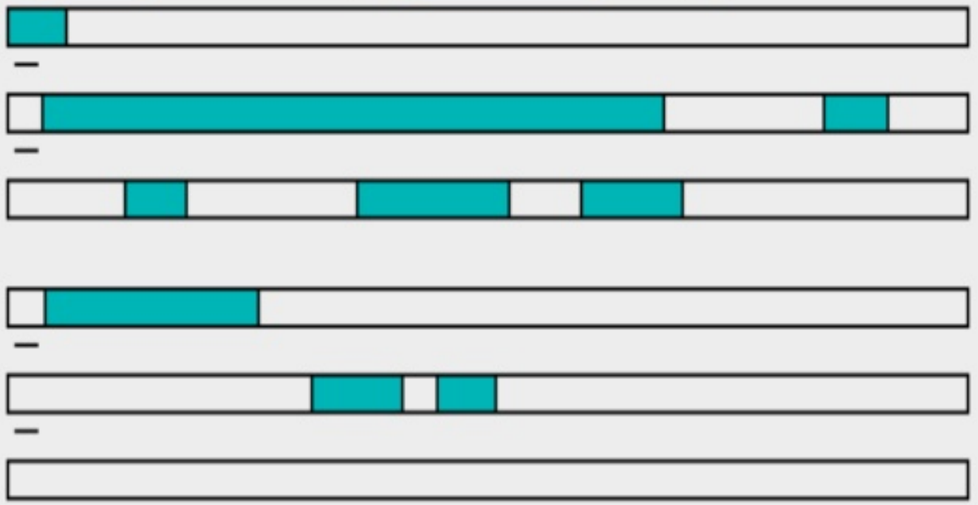
info@ncbi.nlm.nih.gov

Credits to: Tatiana Tatusov and Eugene Tatusov

ORF Finder (Open Reading Frame Finder)

gi|31343070|ref|NM_174000.2| Bos taurus calreticulin (CALR), mRNA

View 1 GenBank Redraw 100 SixFrames



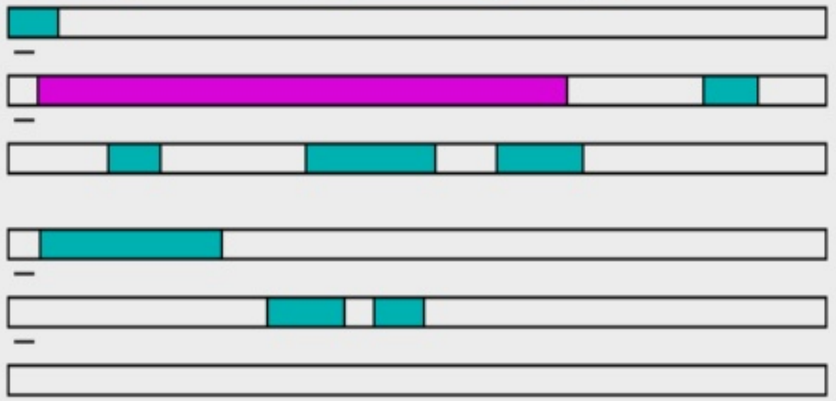
Frame	from	to	Length
+2	71	1324	1254
-1	77	505	429
+3	705	1013	309
+3	1155	1361	207
-2	613	795	183
+2	1646	1771	126
+3	240	362	123
+1	1	123	123
-2	868	984	117

ORF Finder (Open Reading Frame Finder)

gi|31343070|ref|NM_174000.2| Bos taurus calreticulin (CALR), mRNA

Program blastp Database nr BLAST with parameters Cognitor

View 1 GenBank Redraw 100 SixFrames

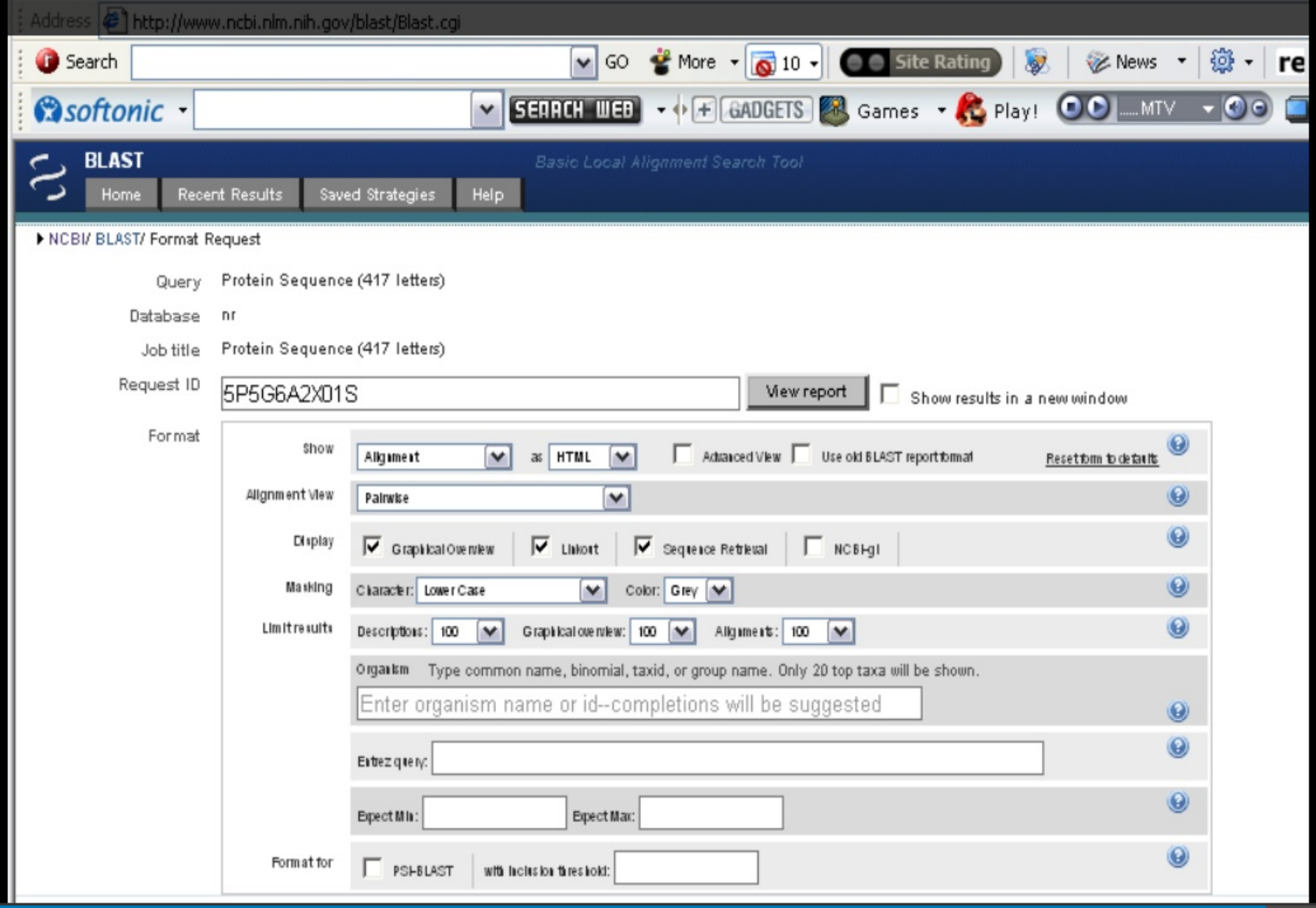


Length: 417 aa

Frame	from	to	Length
+2	71	1324	1254
-1	77	505	429
+3	705	1013	309
+3	1155	1361	207
-2	613	795	183
+2	1646	1771	126
+3	240	362	123
+1	1	123	123
-2	868	984	117

NCBI

```
  71 atgctgctacccgtgccgctgctgctcggccttctcggcctggcc
     M  L  L  P  V  P  L  L  L  G  L  L  G  L  A
 116 gccgctgatcccaccgtttactttaaggagcagtttctggacgga
     A  A  D  P  T  V  Y  F  K  E  Q  F  L  D  G
 161 gacgggtggaccgagcgctggatccaagcacaaaccggat
     D  G  W  T  E  R  W  I  E  S  K  H  K  P  D
 206 tttggcaaattcgttctcagttccggcaagttctatggtgaccaa
     F  G  K  F  V  L  S  S  G  K  F  Y  G  D  Q
 251 gagaaagataaaggcctgcagactagccaggatgcccggttctac
     E  K  D  K  G  L  Q  T  S  Q  D  A  R  F  Y
 296 gctctgtcggccagatttgagcccttcagcaacaagggtcagacg
     A  L  S  A  R  F  E  P  F  S  N  K  G  Q  T
 341 ctggtggtgcagttcacagtgaaacacgagcagaacatcgactgt
     L  V  V  Q  F  T  V  K  H  E  Q  N  I  D  C
 386 gggggcggctatgtgaagctgtttccagctggtttggatcagaca
     G  G  G  Y  V  K  L  F  P  A  G  L  D  Q  T
 431 gacatgcacggagactccgaatacaacatatgtttggcccggac
     D  M  H  G  D  S  E  Y  N  I  M  F  G  P  D
 476 atctgtggccctggcaccaagaaggttcatgtcatcttcaactac
     I  C  G  P  G  T  K  K  V  H  V  I  F  N  Y
 521 aagggcaagaatgtgctgatcaacaaggatatccgctgcaaggac
     K  G  K  N  V  L  I  N  K  D  I  R  C  K  D
 566 gatgaattcacccacctgtacacgctgattgtgcggcctaataat
     D  E  F  T  H  L  Y  T  L  I  V  R  P  N  N
 611 acctatgaggtgaagattgacaacagccaggtggagtcaggctcc
     T  Y  E  V  K  I  D  N  S  Q  V  E  S  G  S
 656 ttggaggacgattgggatttcttgccacccaagaagataaaggat
     L  E  D  D  W  D  F  L  P  P  K  K  I  K  D
 701 cctgatgccgctaagcctgaagactgggacgatcgcgccaagatc
     P  D  A  A  K  P  E  D  W  D  D  R  A  K  I
 746 gatgaccctacagactccaagcctgaggattgggacaagcctgag
     D  D  P  T  D  S  K  P  E  D  W  D  K  P  E
 791 cacattcctgaccctgatgctaagaaacctgaggactgggatgaa
     H  I  P  D  P  D  A  K  K  P  E  D  W  D  E
 836 gagatggacggagagtgggaaccacctgttattcagaacccagag
     E  M  D  G  E  W  E  P  P  V  I  Q  N  P  E
 881 tacaaggggagtggaaacccaggcagatcgacaacccagagtac
     Y  K  G  E  W  K  P  R  Q  I  D  N  P  E  Y

 926 aagggcatttggatccacccagagattgacaaccctgagtattcc
     K  G  I  W  I  H  P  E  I  D  N  P  E  Y  S
 971 cctgacagcaacatctatgcctatgaaaacttcgctgttctaggc
     P  D  S  N  I  Y  A  Y  E  N  F  A  V  L  G
1016 ttggatctttggcaggtcaagtctggcaccatctttgacaacttc
     L  D  L  W  Q  V  K  S  G  T  I  F  D  N  F
1061 ctcatcaccaacgatgaagcgtatgctgaggagtttggcaacgag
     L  I  T  N  D  E  A  Y  A  E  E  F  G  N  E
1106 acgtggggtgttacaaaggcagcagaaaagcaaatgaaggacaag
     T  W  G  V  T  K  A  A  E  K  Q  M  K  D  K
1151 caggatgaagagcagaggctacatgaggaggaggaggagaaagaaa
     Q  D  E  E  Q  R  L  H  E  E  E  E  E  K  K
1196 ggcaaggaggaggaagaggcagacaaagatgatgacgaagacaag
     G  K  E  E  E  E  A  D  K  D  D  D  E  D  K
1241 gatgaggatgaggaggatgaagatgagaaggaagaggaggaggaa
     D  E  D  E  E  D  E  D  E  K  E  E  E  E  E
```


Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Search GO More 10 Site Rating News
softonic SEARCH WEB GADGETS Games Play! MTV
BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ Format Request
Query Protein Sequence (417 letters)
Database nr
Job title Protein Sequence (417 letters)
Request ID 5P5G6A2X01S
View report
Show results in a new window
Format
Show Alignment as HTML Advanced View Use old BLAST report format Reset form to defaults
Alignment View Pairwise
Display Graphical Overview Linkout Sequence Retrieval NCBI-gi
Masking Character: Lower Case Color: Grey
Limit results Descriptions: 100 Graphical overview: 100 Alignments: 100
Organism Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.
Enter organism name or id--completions will be suggested
Entrez query:
Expect Min: Expect Max:
Format for PSI-BLAST with inclusion threshold:

NCBI

BLAST Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ blastp suite/ Formatting Results - 5P5G6A2X01S

Edit and Resubmit | Save Search Strategies | ▷Formatting options | ▷Download

Protein Sequence (417 letters)

Query ID lcl|31691
Description None
Molecule type amino acid
Query Length 417

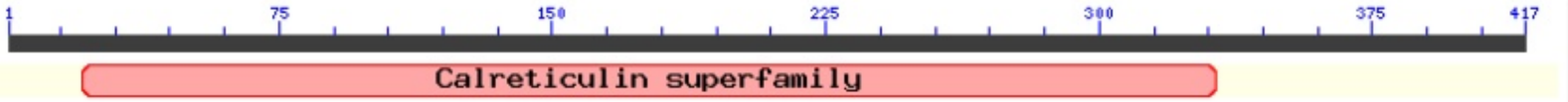
Database Name nr
Description All non-redundant GenBank CDS tr
samples from WGS projects
Program BLASTP 2.2.24+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

Graphic Summary

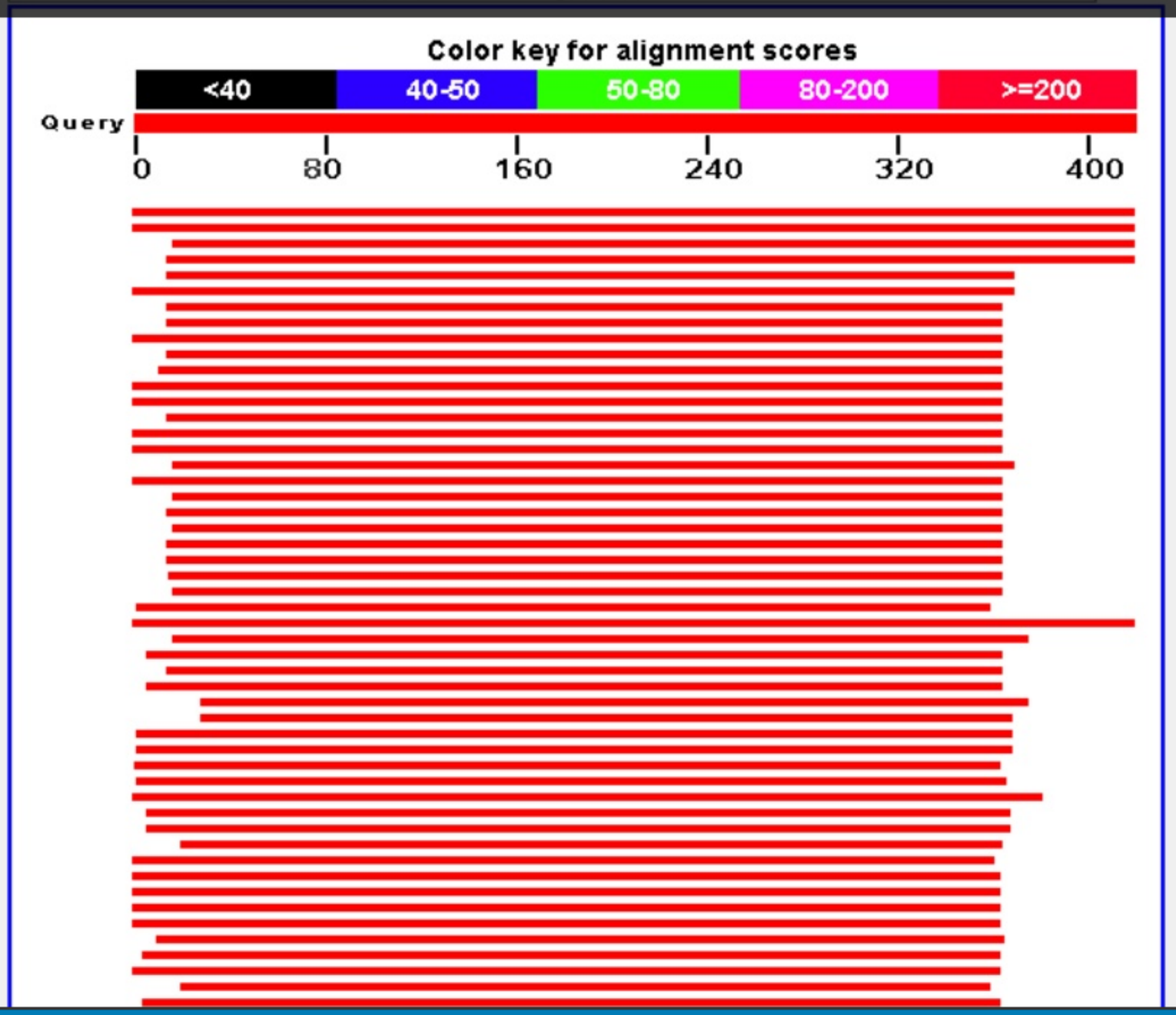
Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of 100 Blast Hits on

Mouse over to see the defline; click to show alignments



Descriptions

Sequences producing significant alignments:

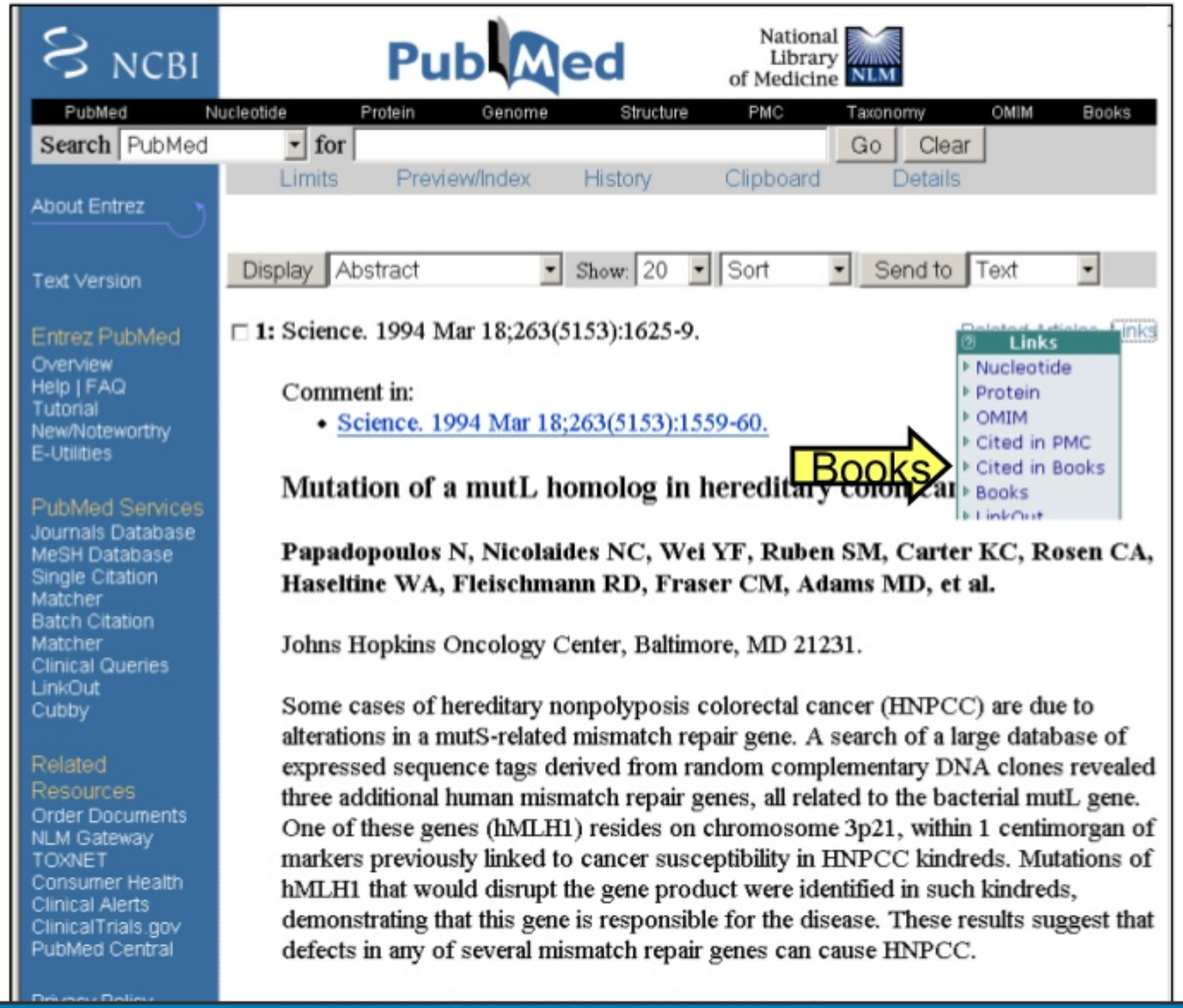
	Score (Bits)	E Value
ref|NP_776425.1| calreticulin precursor [Bos taurus] >sp|P521...	843	0.0
gb|AAI40583.1| CALR protein [Bos taurus] >gb|DAA28020.1| calr...	840	0.0
gb|AAB30209.1| calreticulin [cattle, brain, Peptide, 400 aa]	807	0.0
ref|NP_001167604.1| calreticulin [Sus scrofa] >gb|ADD52600.1|...	719	0.0
ref|NP_001075704.1| calreticulin precursor [Oryctolagus cunic...	704	0.0
dbj|BAE22866.1| unnamed protein product [Mus musculus]	701	0.0
ref|XP_853393.1| PREDICTED: similar to calreticulin isoform 2...	698	0.0
ref|XP_002921056.1| PREDICTED: calreticulin-like [Ailuropoda ...	697	0.0
ref|NP_071794.1| calreticulin precursor [Rattus norvegicus] >...	695	0.0
ref|XP_001504932.1| PREDICTED: similar to calreticulin [Equus...	694	0.0
dbj|BAD96780.1| calreticulin precursor variant [Homo sapiens]	692	0.0
ref|NP_031617.1| calreticulin precursor [Mus musculus] >sp|P1...	692	0.0
dbj|BAE35687.1| unnamed protein product [Mus musculus]	691	0.0
ref|XP_867320.1| PREDICTED: similar to calreticulin isoform 5...	690	0.0
sp|Q8K3H7.1|CALR_CRIGR RecName: Full=Calreticulin; AltName: F...	690	0.0
ref|XP_001110217.1| PREDICTED: calreticulin isoform 2 [Macaca...	688	0.0
gb|AAB20096.1| calreticulin [rabbits, sketetal muscle, Peptid...	688	0.0
ref|XP_002761834.1| PREDICTED: calreticulin-like [Callithrix ...	687	0.0
ref|NP_004334.1| calreticulin precursor [Homo sapiens] >ref|X...	682	0.0
ref|XP_867302.1| PREDICTED: similar to calreticulin isoform 3...	678	0.0
gb|EAW84330.1| calreticulin, isoform CRA_a [Homo sapiens]	673	0.0
ref|XP_533899.2| PREDICTED: similar to calreticulin isoform 1...	669	0.0
ref|XP_867310.1| PREDICTED: similar to calreticulin isoform 4...	650	0.0
ref|XP_001377711.1| PREDICTED: similar to precursor (AA -17 t...	645	0.0
ref|XP_001514084.1| PREDICTED: similar to calreticulin [Ornit...	645	0.0
gb|AAS49610.1| calreticulin [Gallus gallus]	645	0.0

Literature Links

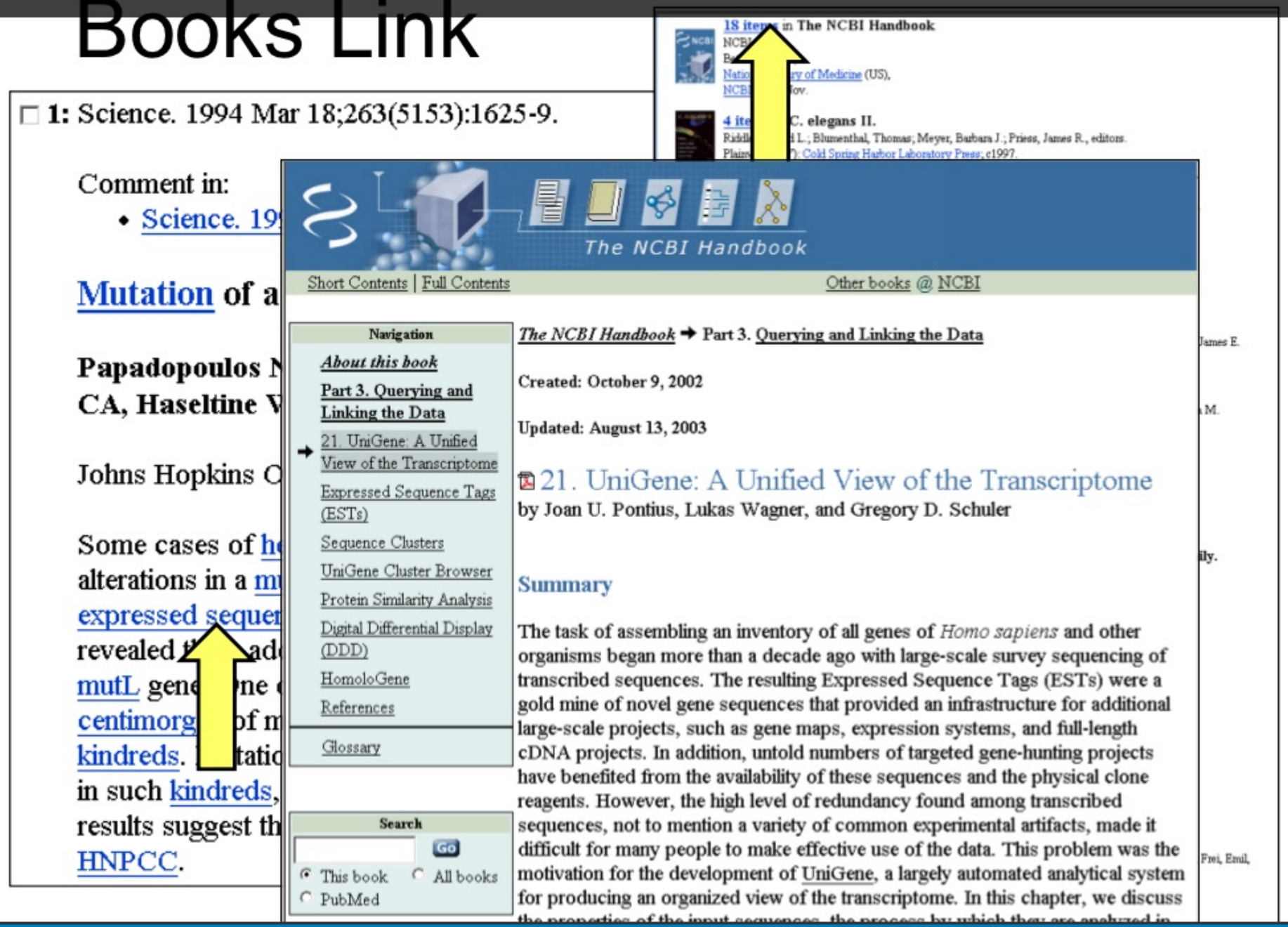
PubMed

OMIM

NM_000249: PubMed



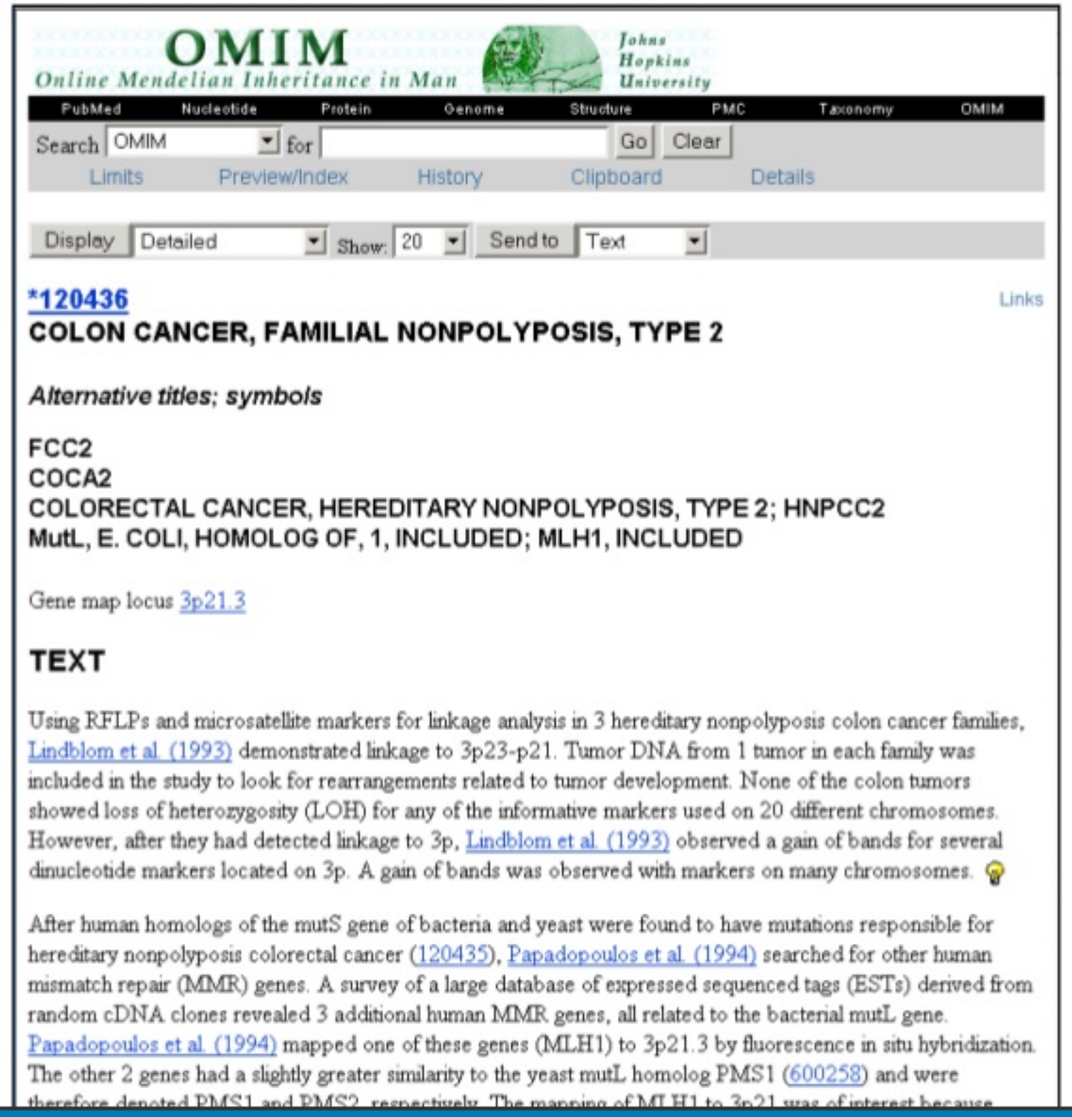
Books Link


1: Science. 1994 Mar 18;263(5153):1625-9.
Comment in:
Science. 19
Mutation of a
Papadopoulos N
CA, Haseltine V
Johns Hopkins C
Some cases of h
alterations in a m
expressed sequer
revealed t ad
mutL gene ne
centimorg of m
kindreds. tatic
in such kindreds,
results suggest th
HNPCC.
18 items in The NCBI Handbook
Natio ry of Medicine (US),
NCBI ov.
4 ite C. elegans II.
L.; Blumenthal, Thomas; Meyer, Barbara J.; Priess, James R., editors.
Cold Spring Harbor Laboratory Press; c1997.
The NCBI Handbook
Short Contents | Full Contents
Other books @ NCBI
Navigation
About this book
Part 3. Querying and Linking the Data
21. UniGene: A Unified View of the Transcriptome
Expressed Sequence Tags (ESTs)
Sequence Clusters
UniGene Cluster Browser
Protein Similarity Analysis
Digital Differential Display (DDD)
HomoloGene
References
Glossary
Search
GO
This book
All books
PubMed
The NCBI Handbook ➜ Part 3. Querying and Linking the Data
Created: October 9, 2002
Updated: August 13, 2003
21. UniGene: A Unified View of the Transcriptome
by Joan U. Pontius, Lukas Wagner, and Gregory D. Schuler
Summary
The task of assembling an inventory of all genes of Homo sapiens and other organisms began more than a decade ago with large-scale survey sequencing of transcribed sequences. The resulting Expressed Sequence Tags (ESTs) were a gold mine of novel gene sequences that provided an infrastructure for additional large-scale projects, such as gene maps, expression systems, and full-length cDNA projects. In addition, untold numbers of targeted gene-hunting projects have benefited from the availability of these sequences and the physical clone reagents. However, the high level of redundancy found among transcribed sequences, not to mention a variety of common experimental artifacts, made it difficult for many people to make effective use of the data. This problem was the motivation for the development of UniGene, a largely automated analytical system for producing an organized view of the transcriptome. In this chapter, we discuss

W. H. FREEMAN AND COMPANY
BIOCHEMISTRY
FIFTH EDITION
Jeremy M. Berg John L. Tymoczko Lubert Stryer

Figure 31.29. Structure of a Bromodomain. This four-helix-bundle domain binds peptides containing acetyllysine. An acetylated

OMIM: Human Disease Genes



OMIM
Online Mendelian Inheritance in Man
Johns Hopkins University

PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | OMIM

Search OMIM for [ ] Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Detailed Show: 20 Send to Text

*120436 Links

COLON CANCER, FAMILIAL NONPOLYPOSIS, TYPE 2

Alternative titles; symbols

FCC2
COCA2
COLORECTAL CANCER, HEREDITARY NONPOLYPOSIS, TYPE 2; HNPCC2
MutL, E. COLI, HOMOLOG OF, 1, INCLUDED; MLH1, INCLUDED

Gene map locus 3p21.3

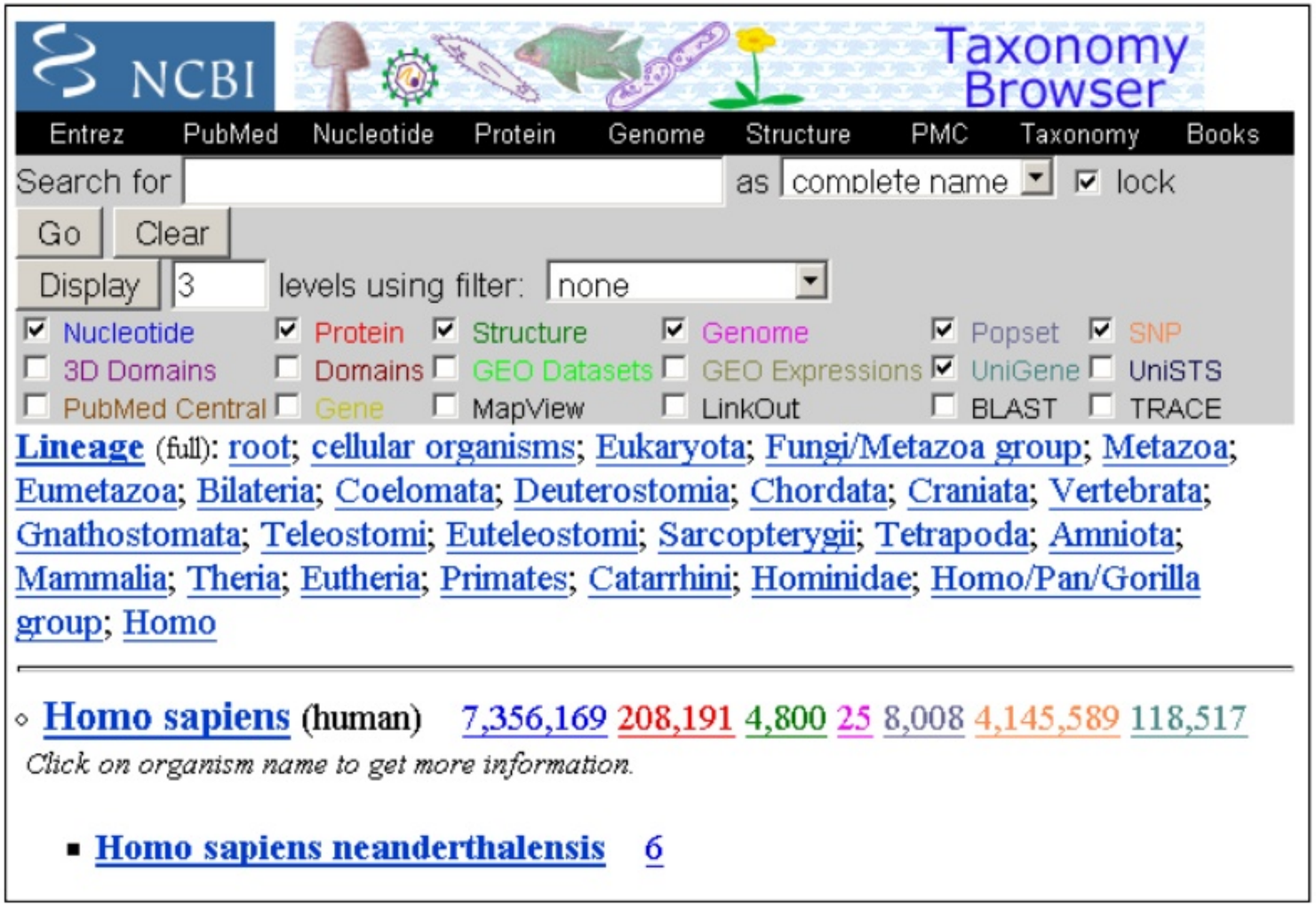
TEXT

Using RFLPs and microsatellite markers for linkage analysis in 3 hereditary nonpolyposis colon cancer families, Lindblom et al. (1993) demonstrated linkage to 3p23-p21. Tumor DNA from 1 tumor in each family was included in the study to look for rearrangements related to tumor development. None of the colon tumors showed loss of heterozygosity (LOH) for any of the informative markers used on 20 different chromosomes. However, after they had detected linkage to 3p, Lindblom et al. (1993) observed a gain of bands for several dinucleotide markers located on 3p. A gain of bands was observed with markers on many chromosomes.

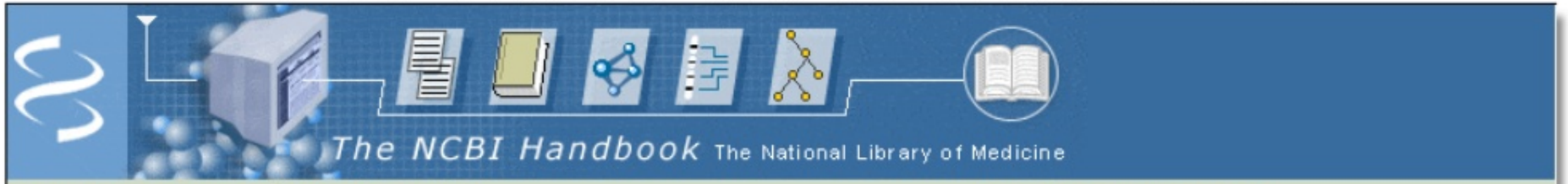
After human homologs of the mutS gene of bacteria and yeast were found to have mutations responsible for hereditary nonpolyposis colorectal cancer (120435), Papadopoulos et al. (1994) searched for other human mismatch repair (MMR) genes. A survey of a large database of expressed sequenced tags (ESTs) derived from random cDNA clones revealed 3 additional human MMR genes, all related to the bacterial mutL gene. Papadopoulos et al. (1994) mapped one of these genes (MLH1) to 3p21.3 by fluorescence in situ hybridization. The other 2 genes had a slightly greater similarity to the yeast mutL homolog PMS1 (600258) and were therefore denoted PMS1 and PMS2, respectively. The mapping of MLH1 to 3p21 was of interest because

Taxonomy Link

Taxonomy Link



For More Information…



Short Contents | Full Contents Other books @ NCBI

The NCBI Handbook

Part 1. The Databases

1. GenBank: The Nucleotide Sequence Database
Ilene Mizrachi.
Created: October 9, 2002, Updated: July 27, 2004

2. PubMed: The Bibliographic Database
Kathi Canese, Jennifer Jentsch, and Carol Myers.
Created: October 9, 2002, Updated: August 13, 2003

3. Macromolecular Structure Databases
Eric Sayers and Steve Bryant.
Created: October 9, 2002, Updated: August 13, 2003

4. The Taxonomy Project
Scott Federhen.
Created: October 9, 2002, Updated: August 13, 2003

5. The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation
Adrienne Kitts and Stephen Sherry.
Created: October 09, 2002, Updated: June 16, 2006

6. The Gene Expression Omnibus (GEO): A Gene Expression and Hybridization Repository
Ron Edgar and Alex Lash

Search
Go
This book All books
PubMed

Thank You!