

Biological databases

Ali Etemadi

Tehran University of Medical Sciences

Contents

- Database defined
- Entrez
- Literature review in PubMed
- Types of biological Databases
- GenBank & file format
- Genomic DNA Databases
- RNA databases
- Protein Databases (Uniprot)
- The Reference Sequence (RefSeq) Project
- Gene
- Other NCBI Databases

Database defined

- A database is an **organized collection of structured information**, or data, typically **stored electronically** in a computer system. A database is usually controlled by a database management system (DBMS). Together, the **data** and the **DBMS**, along with the **applications** that are associated with them, are referred to as a **database system**, often shortened to just **database**.
- Data within the most common types of databases in operation today is typically modeled in **rows and columns** in a series of tables to make processing and data querying efficient. The data can then be easily accessed, managed, modified, updated, controlled, and organized.

A simple database

A1	fx	A	B	C	D	E	F	G	H
1		اسمی دانشجویان	سال ورود	ایمیل	جلسه ۲۵ بهمن تاریخ ارایه (اعلام میشود)	۲۹ فروردین	۲ اسفند	۵ اردیبهشت	
2		بشار مسیبی	۱۴۰۱		۲۰ دی	*			
3		امیرحسین اکبرپور	۱۴۰۱		۵ اردیبهشت				
4		فاطمه خاری	۱۴۰۱		۱۸ بهمن	*			
5		معصومه امیرلو	۱۴۰۱						
6		مصطفی گنجوری	۱۴۰۰						
7		سکینه منزوی	۱۴۰۰	S.monzavi@yahoo.com		*			
8		احسان فیروزی مجد	۱۴۰۰						
9		زهراء باقریان	۱۴۰۰						
10		نادیا پورمشیر	۱۳۹۹	Pourmoshir.nadia@gmail.com					
11		سید حسین مرادی	۱۳۹۹	hosseinmoradiran@gmail.com		*			
12		آزاده اشکیان	۱۳۹۹	A.ashkiyan@gmail.com		*			
13		عاطفه بخششان	۱۳۹۸	atefehbakhshian1@gmail.com	۴ بهمن *	*			
14		الهام عرفانی ایزدیار	۱۳۹۸	Erfani.elham@gmail.com		*			
15		سجاد غورقانلو	۱۳۹۸						
16		نعمه روشن ضمیر	۱۳۹۸						
17		لدا یعقوبی	۱۳۹۷						
18		اسد کاظمی	۱۳۹۷						
19		سیده مهسا صالحی	۱۳۹۷						
~									

Object-Oriented Model

Object 1: Maintenance Report

Date	
Activity Code	
Route No.	
Daily Production	
Equipment Hours	
Labor Hours	

Object 1 Instance

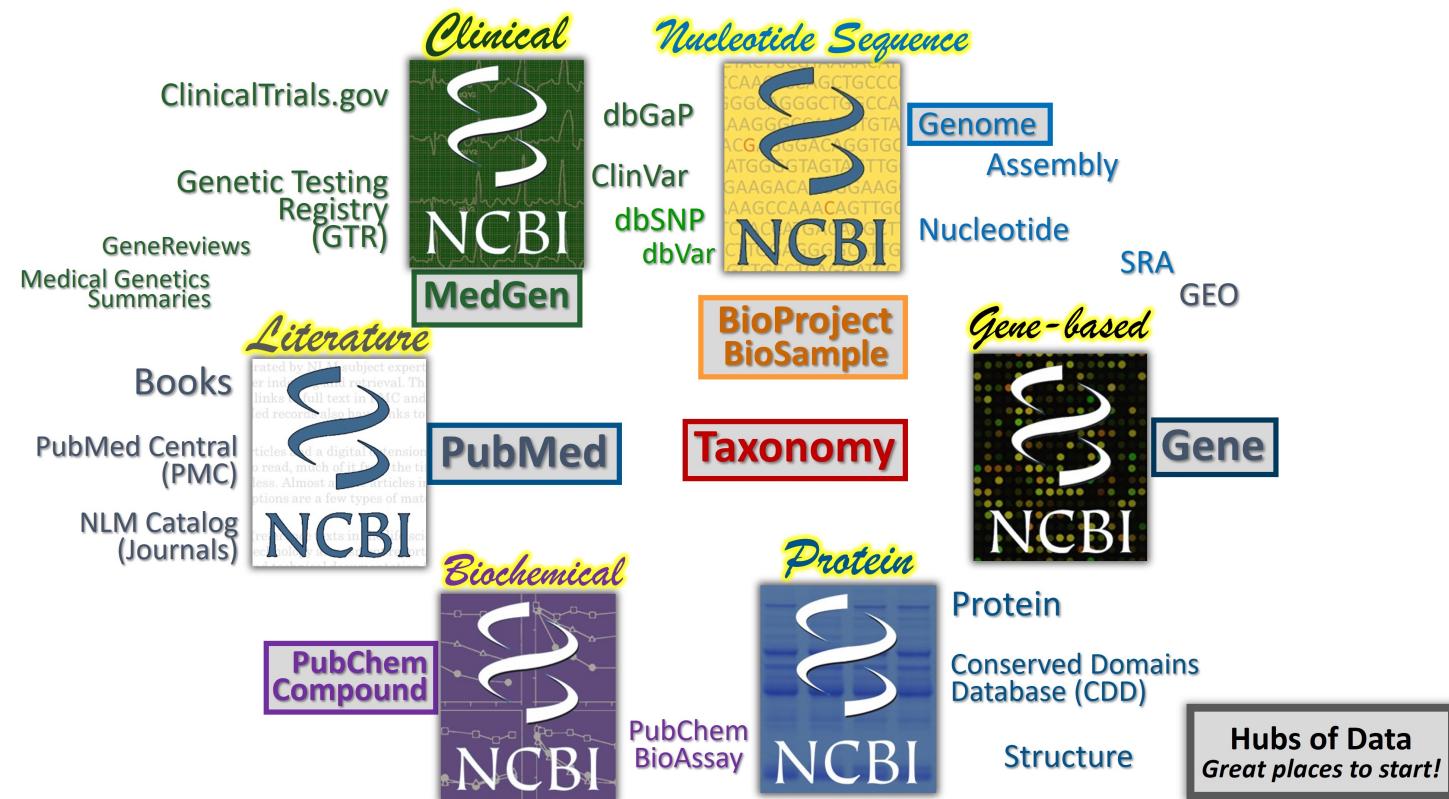
01-12-01	
24	
I-95	
2.5	
6.0	
6.0	

Object 2: Maintenance Activity

Activity Code	
Activity Name	
Production Unit	
Average Daily Production Rate	

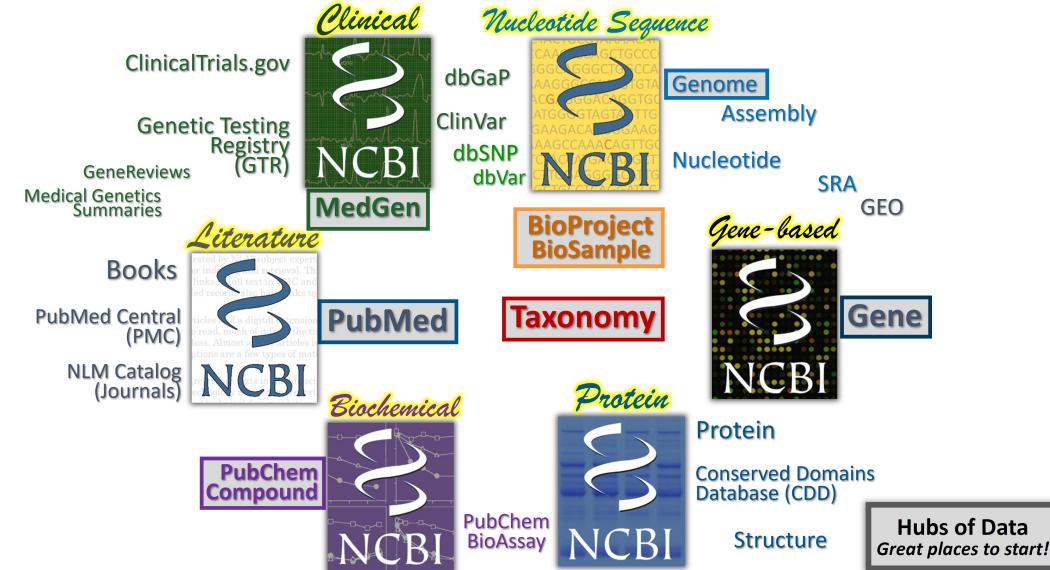
main centralized bioinformatics hubs

- National Center for Biotechnology Information (NCBI)
- European Bioinformatics Institute (EBI)



NCBI

- search service from the National Library of Medicine (NLM)
- more than 36 million citations for biomedical literature from MEDLINE, life science journals, and online books.
- Entrez *

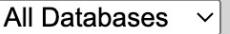


National Library of Medicine

National Center for Biotechnology Information

Log in

All Databases



Search

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information is a source of biomedical and genomic information.

[About the NCBI](#) | [Mission Statement](#)

Submit

Deposit data or manuscripts into NCBI databases



Develop

Use NCBI APIs and code libraries to build applications



Submit new data/manuscripts



Analyze

Identify an NCBI tool for your data analysis task



Download bulky data Using ftp

All available NCBI tools



News & Blog

Foreign Contamination Screen Tool: Now Available in Galaxy!

07 Mar 2024

Check out our latest enhancements! Do you submit genome assembly data to

Celebrating the Building Transparent ML/AI Solutions for Biological Research Codeathon

05 Mar 2024

Machine Learning and Artificial

Literature review in NCBI

The screenshot shows a web browser window with the URL <https://www.ncbi.nlm.nih.gov/search/all/?term=beta%20globin>. The search term 'beta globin' is entered in the search bar. The results are displayed in three main sections: Literature, Genomes, and Clinical. The Literature section is highlighted with a red rounded rectangle. It includes links to Bookshelf (471), MeSH (4), NLM Catalog (13), PubMed (11,717), and PubMed Central (42,097). The Genomes section includes Assembly (0), BioCollections (0), BioProject (169), BioSample (922), Genome (0), Nucleotide (8,253), SRA (1,872), and Taxonomy (0). The Clinical section includes links to various clinical databases like ClinicalTrials.gov, dbSNP, dbVar, GTR, MedGen, and OMIM.

Results found in 26 databases (1 error)

Literature	
Bookshelf	471
MeSH	4
NLM Catalog	13
PubMed	11,717
PubMed Central	42,097
Genomes	
Assembly	0
BioCollections	0
BioProject	169
BioSample	922
Genome	0
Nucleotide	8,253
SRA	1,872
Clinical	
ClinicalTrials.gov	45
ClinicalTrials.gov	0
ClinicalTrials.gov	0
dbSNP	27
dbVar	0
dbVar	0
dbVar	0
GTR	0
MedGen	0
OMIM	0

Bookshelf provides free online access to books and documents in life science and healthcare

controlled vocabulary thesaurus used for indexing articles for PubMed.

MeSH

MeSH protein design

Create alert Limits Advanced

Summary ▾

Send to: ▾

Search results

Items: 4

- [Structural Homology, Protein](#)
 - 1. The degree of 3-dimensional shape similarity between **proteins**. It can be an indication of distant AMINO ACID SEQUENCE HOMOLOGY and used for rational DRUG **DESIGN**.
Year introduced: 2003
- [Protein Engineering](#)
 - 2. Procedures by which **protein** structure and function are changed or created in vitro by altering existing or synthesizing new structural genes that direct the synthesis of **proteins** with sought-after properties. Such procedures may include the **design** of MOLECULAR MODELS of **proteins** using COMPUTER GRAPHICS or other molecular modeling techniques; site-specific mutagenesis (MUTAGENESIS, SITE-SPECIFIC) of existing genes; and DIRECTED MOLECULAR EVOLUTION techniques to create new genes.
Year introduced: 1989
- [FSD-1 protein, synthetic \[Supplementary Concept\]](#)
 - 3. FSD-1 - Full Sequence **Design** 1; has homology with Zif268; partial amino acid sequence given in first source
Date introduced: October 22, 1997
- [saccharide-binding proteins \[Supplementary Concept\]](#)
 - 4. facilitate the molecular **design** of drug carriers for greater target cell selectivity
Date introduced: April 18, 1988

"Supplementary Concept" feature refers to additional terms that are not included in the main MeSH hierarchy but are relevant to specific topics.

A screenshot of a web browser showing the NCBI search results for the term "beta globin". The search bar at the top contains "beta globin". Below it, a message says "Results found in 26 databases (1 error)". A large yellow circle highlights the "Literature" section, which includes links to Bookshelf (471), MeSH (4), NLM Catalog (13), PubMed (11,717), and PubMed Central (42,097). Another yellow circle highlights the "Genomes" section, which includes Assembly (0), BioCollections (0), BioProject (169), BioSample (922), Genome (0), Nucleotide (8,253), SRA (1,872), and Taxonomy (0). To the right of these sections, there is a list of other databases: GenBank, GDB, OMIM, ClinVar, ClinGen, OMIM, and Family Models. The bottom right corner of the page has a "linkOut" button.

Search NCBI

beta globin

Results found in 26 databases (1 error)

Literature

Bookshelf	471
MeSH	4
NLM Catalog	13
PubMed	11,717
PubMed Central	42,097

Genomes

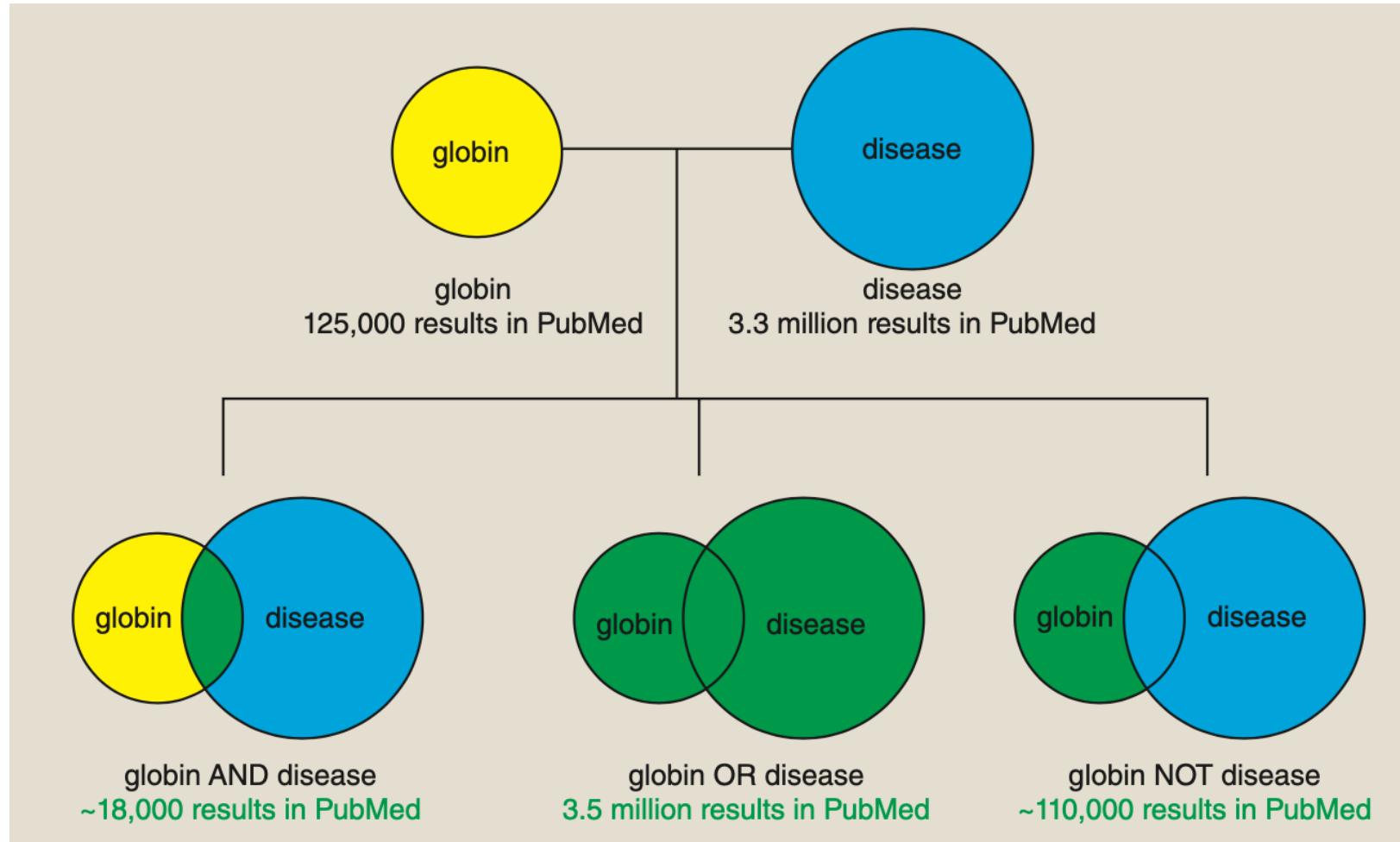
Assembly	0
BioCollections	0
BioProject	169
BioSample	922
Genome	0
Nucleotide	8,253
SRA	1,872
Taxonomy	0

journals, books,
audiovisuals,
computer software,
electronic resources
and other materials.

Pubmed
Pubmed central
linkOut

- AND, OR, and NOT must be capitalized
- For specific phrase, put it in “quotation marks”
- Asterisk* meaning all> “glob*”

Boolean operators



Biological databases: why we need them?

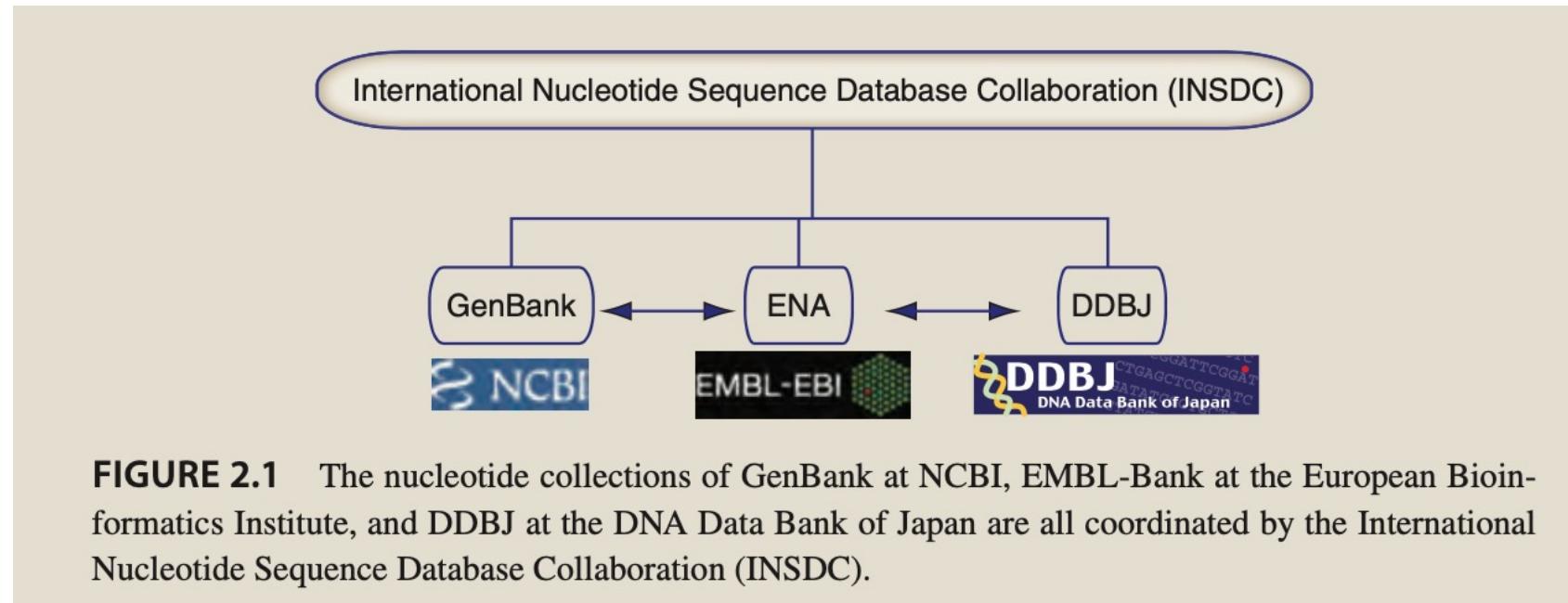
- In 1995 complete genome of a living organism sequenced (haemophilus influenza)
 - Few years later >> 1000s organisms
 - Now more than 260,000 organisms
-
- The goal is to introduce databases that store these huge information and strategy to extract information from them.

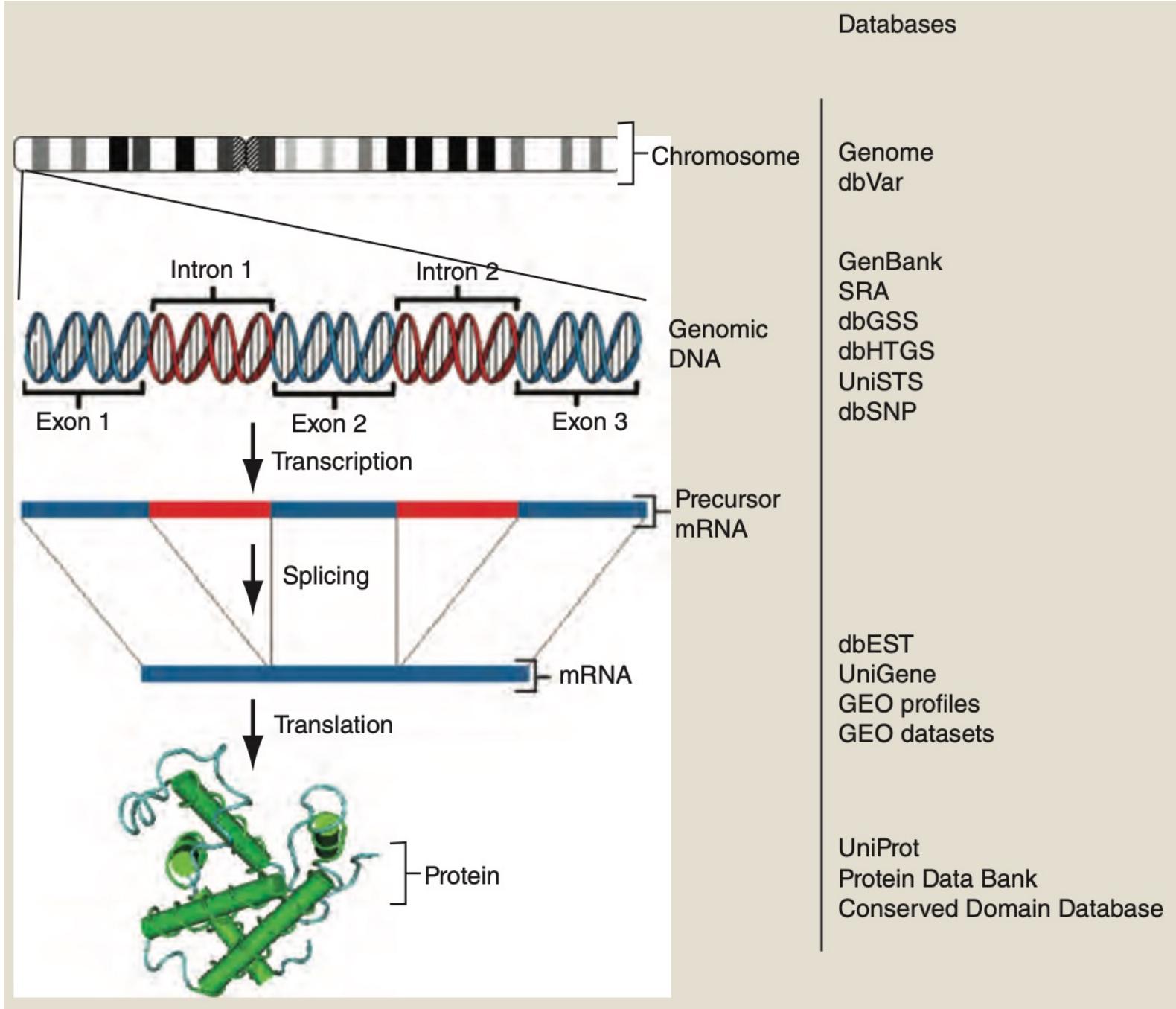
Types of Databases

- 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, RefSNP, NCBI Protein, Structure, Conserved Domain

Three main publicly access databases to access NAs and proteins

- Sequence information is equivalent in these databases
- Daily sync (release occurs every two months)
- Submit directly and quality check at time of submission
- Tbl2asn in NCBI automate creation of sequence records





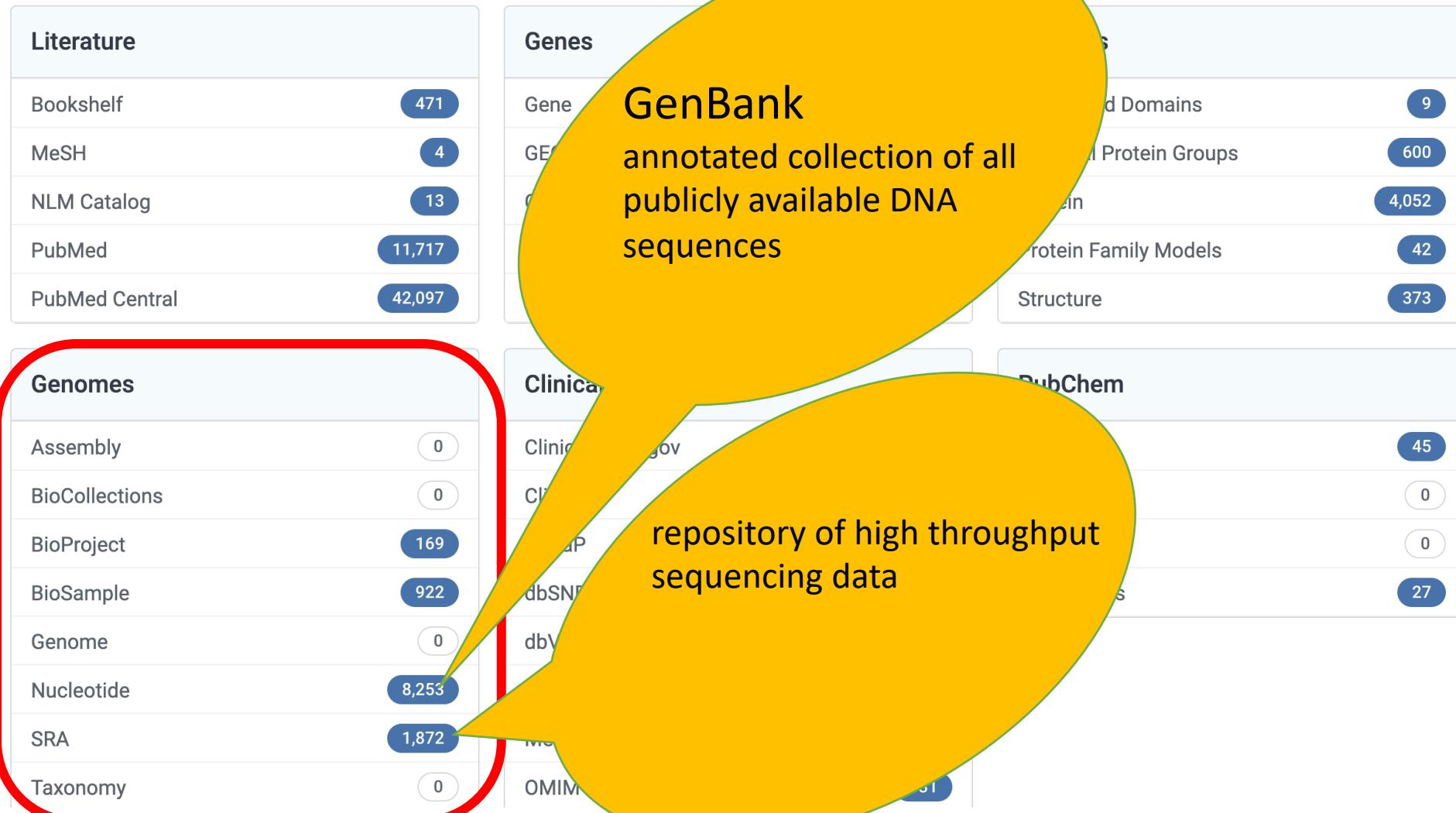
Types of Data in GenBank

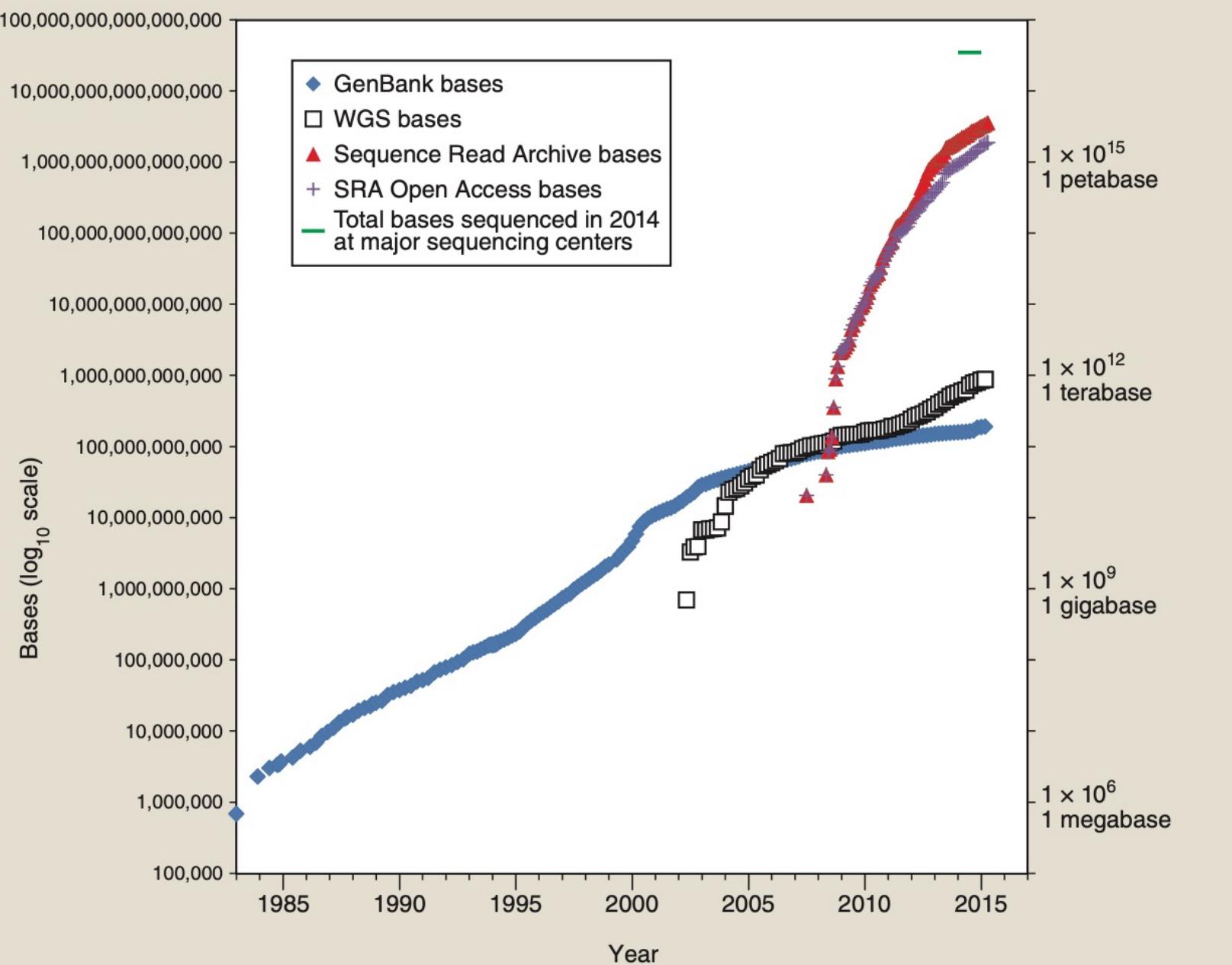
- DNA, RNA-based, and protein sequences are stored in discrete databases
- Because RNA is relatively unstable, it is typically converted to complementary DNA (cDNA), and a variety of databases contain cDNA sequences corresponding to RNA transcripts.
- The human genome organization (HUGO) Gene Nomenclature Committee (HGNC) has the critical task of assigning official names to genes and proteins.

دانلود rosetta_stuff Editorial Manager® Paraphrasing Tool... FDA & EMA's New... Review Reports: D... Intertek Pharmace... ScienceDaily: You... VPN

Search NCBI beta globin Search

Results found in 26 databases (1 error)





Base pairs	Unit	Abbreviation	Example
1	1 base pair	1 bp	
1000	1 kilobase pair	1 kb	
10^9	1 megabase pair	1 Mb	
10^{12}	1 gigabase pair	1 Gb	
10^{15}	1 terabase pair	1 Tb	
	1 petabase pair	1 Pb	

Size of a typical coding region of a gene
 Size of a typical bacterial genome
 The human genome is 3 billion base pairs

SRA sequences are short (typically 50–400 base pairs, refer to read in NGS technologies)

GenBank

- <https://www.ncbi.nlm.nih.gov/genbank/>
- an annotated collection of all publicly available DNA and protein sequences **exclude NGS data**
- Bibliographic and biological annotations
- Doubled in every ~18 months (from 1982-2015)
- WGS sequences are generated by high throughput sequencing.
- 1 Gb (billion) in a day
- whole human genome is about 3 Gb

Contents of DNA, RNA, and Protein databases

- Over 310,000 different species are represented in GenBank, with over 1000 new species added per month
- GenBank consist mostly of eukaryotes

1. PRI: primate sequences
2. ROD: rodent sequences
3. MAM: other mammalian sequences
4. VRT: other vertebrate sequences
5. INV: invertebrate sequences
6. PLN: plant, fungal, and algal sequences
7. BCT: bacterial sequences
8. VRL: viral sequences
9. PHG: bacteriophage sequences
10. SYN: synthetic sequences
11. UNA: unannotated sequences
12. EST: expressed sequence tags
13. PAT: patent sequences
14. STS: sequence-tagged sites
15. GSS: genome survey sequences
16. HTG: high-throughput genomic sequences
17. HTC: high-throughput cDNA sequences
18. ENV: environmental sampling sequences
19. CON: constricted sequences
20. TSA: transcriptome shotgun assembly sequences.

A Traditional GenBank Record

The Flatfile Format

Header

Feature Table

Sequence

The Header

LOCUS	AY182241	1931 bp	mRNA	linear	PLN	04-MAY-2004
DEFINITION	Malus x domestica (E,E)-alpha-farnesene synthase (AFS1) mRNA, complete cds.					
ACCESSION	AY182241					
VERSION	AY182241.2	GI:32265057				
KEYWORDS	.					
SOURCE	Malus x domestica (cultivated apple)					
ORGANISM	Malus x domestica Eukaryota; Viriplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicots; core eudicots; rosids; eurosids I; Rosales; Rosaceae; Maloideae; Malus.					
REFERENCE	1 (bases 1 to 1931)					
AUTHORS	Pechous,S.W. and Whitaker,B.D.					
TITLE	Cloning and functional expression of an (E,E)-alpha-farnesene synthase cDNA from peel tissue of apple fruit					
JOURNAL	Planta 219, 84-94 (2004)					
REFERENCE	2 (bases 1 to 1931)					
AUTHORS	Pechous,S.W. and Whitaker,B.D.					
TITLE	Direct Submission					
JOURNAL	Submitted (18-NOV-2002) PSI-Produce Quality and Safety Lab, USDA-ARS, 10300 Baltimore Ave. Bldg. 002, Rm. 205, Beltsville, MD 20705, USA					
REFERENCE	3 (bases 1 to 1931)					
AUTHORS	Pechous,S.W. and Whitaker,B.D.					
TITLE	Direct Submission					
JOURNAL	Submitted (25-JUN-2003) PSI-Produce Quality and Safety Lab, USDA-ARS, 10300 Baltimore Ave. Bldg. 002, Rm. 205, Beltsville, MD 20705, USA					
REMARK	Sequence update by submitter					
COMMENT	On Jun 26, 2003 this sequence version replaced gi:27804758.					

Header: Locus Line

LOCUS	AY182241	1931 bp	mRNA	linear	PLN	04-MAY-2004
ACCESSION	AY182241	Complete cds.				
VERSION	AY182241.2					
KEYWORDS						
SOURCE						
Locus name		Length	Molecule type	Division	Modification Date	
REFERENCE	1 (bases 1 to 1931)					
AUTHORS	Pechous, S.W. and Whitaker, B.D.					
TITLE	Cloning and functional expression of an (E,E)-alpha-farnesene synthase cDNA from peel tissue of apple fruit					
JOURNAL	Planta 219, 84-94 (2004)					
REFERENCE	2 (bases 1 to 1931)					
AUTHORS	Pechous, S.W. and Whitaker, B.D.					
TITLE	Direct Submission					
JOURNAL	Submitted (18-NOV-2002) PSI-Produce Quality and Safety Lab, USDA-ARS, 10300 Baltimore Ave. Bldg. 002, Rm. 205, Beltsville, MD 20705, USA					
REFERENCE	3 (bases 1 to 1931)					
AUTHORS	Pechous, S.W. and Whitaker, B.D.					
TITLE	Direct Submission					
JOURNAL	Submitted (25-JUN-2003) PSI-Produce Quality and Safety Lab, USDA-ARS, 10300 Baltimore Ave. Bldg. 002, Rm. 205, Beltsville, MD 20705, USA					
REMARK	Sequence update by submitter					
COMMENT	On Jun 26, 2003 this sequence version replaced gi:27804758.					

Header: Database Identifiers

LOCUS AY182241 1931 bp mRNA linear PLN 04-MAY-2004
DEFINITION Malus x domestica (E,E)-alpha-farnesene synthase cDNA, complete cds.
ACCESSION AY182241

ACCESSION AY182241

Accession

- Stable
- Reportable
- Universal

VERSION AY182241.2 GI:32265057

Version

Tracks changes in sequence

GI number

NCBI internal use

Spermatophytes; Magnoliophyta; eudicotyledons; core eudicots; Malpighiales; Rosales; Rosaceae; Malus; Malus. 1931)
Whitaker, B.D. synthase cDNA from peel tissue of apple fruit
Planta 219, 84-94 (2004)
2 (bases 1 to 1931)
Pechous, S.W. and Whitaker, B.D.
Direct Submission
Submitted (18-NOV-2002) PSI-Produce Quality and Safety Lab,
USDA-ARS, 10300 Baltimore Ave. Bldg. 002, Rm. 205, Beltsville, MD
20705, USA
3 (bases 1 to 1931)
Pechous, S.W. and Whitaker, B.D.
Direct Submission
Submitted (25-JUN-2003) PSI-Produce Quality and Safety Lab,
USDA-ARS, 10300 Baltimore Ave. Bldg. 002, Rm. 205, Beltsville, MD
20705, USA
Sequence update by submitter
On Jun 26, 2003 this sequence version replaced gi:27804758.

Header: Organism

LOCUS AY182241 1931 bp mRNA linear PLN 04-MAY-2004
DEFINITION *Malus x domestica* (E,E)-alpha-farnesene synthase (AFS1) mRNA, complete cds.
ACCESSION AY182241
VERSION AY182241.2 GI:32265057
KEYWORDS .

SOURCE *Malus x domestica* (cultivated apple)

ORGANISM *Malus x domestica*

Eukaryota; Viridiplantae; Streptophyta; Embryophyta;
Tracheophyta; Spermatophyta; Magnoliophyta; eudicots;
core eudicots; rosids; eurosids I; Rosales; Rosaceae;
Maloideae; Malus.

synthase cDNA from peel tissue or apple fruit

JOURNAL *Planta* 219, 84-94 (2004)

REFERENCE 2 (bases 1 to 1931)

AUTHORS Pechous, S.W. and Whitaker, B.D.

TITLE Direct Submission

JOURNAL Submitted (18-NOV-2002) PSI-Produce Quality and Safety Lab,
USDA-ARS, 10300 Baltimore Ave. Bldg. 002, Rm. 205, Beltsville, MD
20705, USA

REFERENCE 3 (bases 1 to 1931)

AUTHORS Pechous, S.W. and Whitaker, B.D.

TITLE Direct Submission

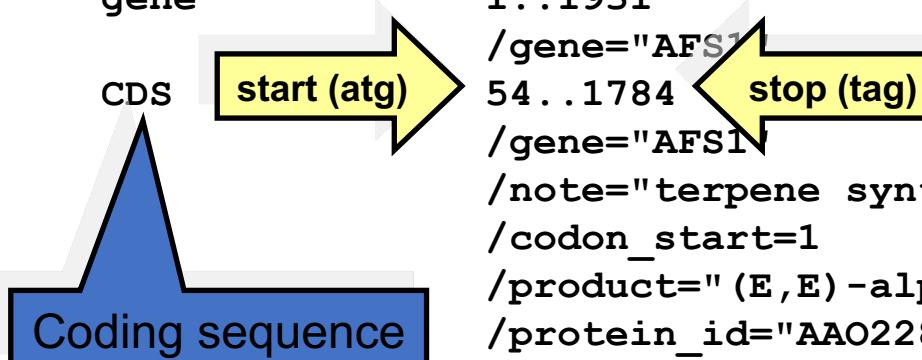
JOURNAL Submitted (25-JUN-2003) PSI-Produce Quality and Safety Lab,
USDA-ARS, 10300 Baltimore Ave. Bldg. 002, Rm. 205, Beltsville, MD
20705, USA

REMARK Sequence update by submitter

COMMENT On Jun 26, 2003 this sequence version replaced gi:27804758.

NCBI-controlled taxonomy

The Feature Table

FEATURES	Location/Qualifiers
source	1..1931 /organism="Malus x domestica" /mol_type="mRNA" /cultivar="'Law Rome'" /db_xref="taxon:3750" /tissue_type="peel"
gene	1..1931 /gene="AFS1" 54..1784 /gene="AFS1" /note="terpene synthase" /codon_start=1 /product="(E,E)-alpha-farnesene synthase" /protein_id="AAO22848.2" /db_xref="GI:32265058" /translation="MEFRVHLQADNEQKIFQNQMKPEPEASYLINQRRSANYKPNIWK NDFLDQSLISKYDGDEYRKLSSEKLIEEVKIVIYISAETMDLVAKLELIDSVRKLGLANLF EKEIKEALDSIAAIESDNLGTRDDLYGTALHFKILRQHGYKVSQDIFGRFMDEKGTL NHFAHLKGMLELFEASNLLGEFEGEDILDEAKASLTALRDSGHICYPDSNLSRDVVHS LELPSHRRVQWFDVKWQINAYEKDICRVNATLLELAKLNFnVVQAQLQKNLREASRW ANLGIADNLKFARDRLVECFACAVGVAFEPEHSSFRICLTKVINLVIIDDVYDIYGS EEELKHFTNAVDRWDSRETEQLPECMKMCFQVLYNTTCEIAREIEENGWNQVLQPLT KVWADFCKALLVEAEWYNKSHIPTLEEYLNGCISSSVSLLVHSFFSITHEGT DFLHKNEDLLYNISLIVRLNNDLGTSAAEQERGDSPSSIVCYMREVNASEETARKNI GMIDNAWKKVNGKCFTTNQVPFLSSFMNNATNMARVAHSILYKDGDGF DQEKGPRTHI
CDS	
Implied protein	

The Reference Sequence (RefSeq) Project

- The goal of RefSeq is to provide the **best representative** sequence for each normal (i.e., **nonmutated**) transcript produced by a gene and for each normal protein product
- GenBank is nearly **redundant** (hundreds of GenBank accession numbers). However, there will be only **one** RefSeq entry corresponding to a given gene or gene product, or several RefSeq entries if there are **splice variants**.
- curated by the staff at **NCBI**

The screenshot shows the NCBI Nucleotide search results page. The search query is "Nucleotide (myoglobin) AND "Homo sapiens""[porgn:txid9606]. The results list 10 items, all of which are mRNA entries for myoglobin in Homo sapiens. Item 1 is a 1,286 bp linear mRNA with accession NM_001382809.1 and GI 1844139640. Item 2 is a 955 bp linear mRNA with accession NM_001382812.1 and GI 1844139609. Item 3 is a 1,243 bp linear mRNA with accession NM_001382811.1 and GI 1844139539. Each result includes links to Protein, PubMed, Taxonomy, GenBank, FASTA, and Graphics.

https://www.ncbi.nlm.nih.gov/nucleotide

An official website of the United States government [Here's how you know](#)

National Library of Medicine
National Center for Biotechnology Information

Nucleotide Help

Species: Homo sapiens (10) Summary 20 per page Sort by Default order Send to: Filter your results: All (10) Manage Filters

Molecule types: genomic DNA/RNA (0) clear Items: 10 Filters activated: mRNA, RefSeq. [Clear all](#)

mRNA (10) [Customize ...](#)

Source databases: RefSeq (10) clear

RefSeq (10) [Customize ...](#)

Sequence Type: Nucleotide (10)

Sequence length: Custom range...

Release date: Custom range...

Revision date: Custom range...

[Clear all](#)

[Show additional filters](#)

Analyze these sequences Run BLAST

Find related data Database: Select

Find items

Search details

myoglobin[All Fields] AND "Homo sapiens"[porgn] AND (biomol_mrna[PROP] AND refseq[filter])

Search See more...

(myoglobin) AND "Homo sapiens"[porgn:txid9606]

- What are differences between :
- NM_005368.3
- NM_203377.1

- What is transcript variant ?
- Many genes have more than one transcript due to alternative splicing. The source of the transcript distinctly varies, and may be regulated and expressed under different physiological conditions.

RefSeq Accession Numbers

- non-redundancy
- updates to reflect current sequence data and biology
- stewardship by NCBI staff and collaborators

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

Chromosomes

NC_123455

Microbial replicons, organelle

Assemblies

NT_123456

Contig

NW_123456

WGS Supercontig



Steps to Assemble a Genome

Some Terminology

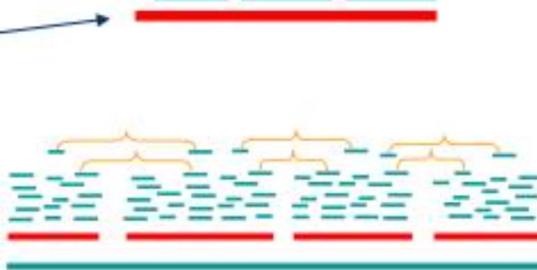
read a 500-900 long word that comes out of sequencer



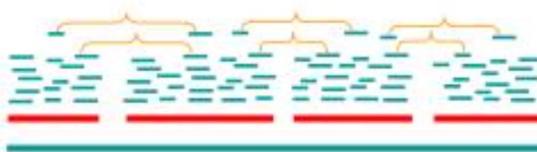
mate pair a pair of reads from two ends of the same insert fragment



contig a contiguous sequence formed by several overlapping reads with no gaps



supercontig an ordered and oriented set (scaffold) of contigs, usually by mate pairs



consensus sequence sequence derived from the multiple alignment of reads in a contig

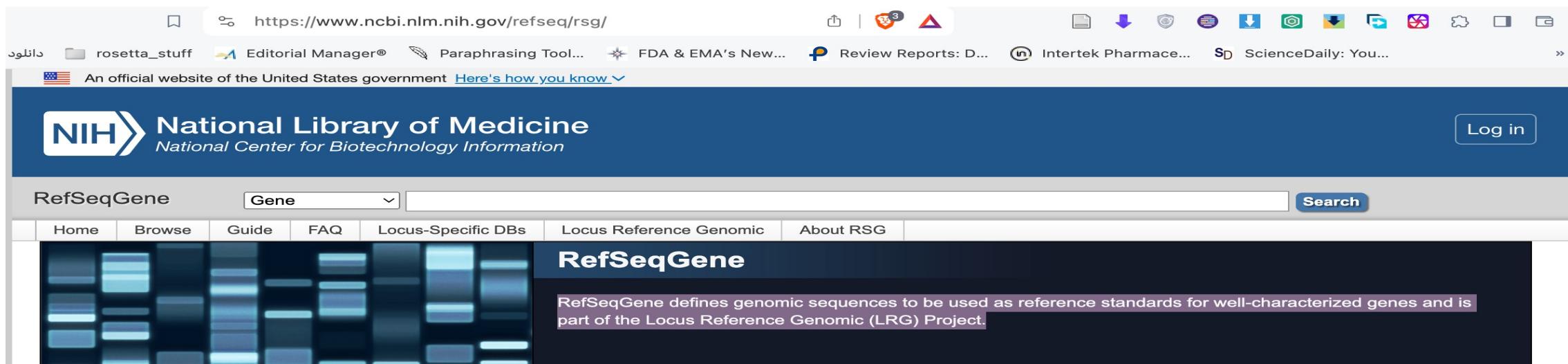
..ACGATTACAATAGGTT..

Other ACCESSION NUMBERS

- GenBank/EMBL/DDBJ nucleotide: One letter followed by five digits (e.g., X02775);
- Protein structure records: one digit followed by three letters (e.g., 1TUP)
- GenInfo (GI): gene identifier GI:28302128

RefSeq limitation

- changing version numbers
- a patient may have a **variant** at a specific nucleotide position in the beta globin gene corresponding to NM_000518.3 but (as often happens) the version number is not given. Once the record is subsequently updated to **NM_000518.4**, anyone studying this variant might be unsure of the correct position of the variant since it depends on which sequence version was used.
- **RefSeqGene:** RefSeqGene defines genomic sequences to be used as reference standards for well-characterized genes



دانلود rosetta_stuff Editorial Manager® Paraphrasing Tool... FDA & EMA's New... Review Reports: D... Intertek Pharmace... ScienceDaily: You... VPN

Search NCBI beta globin Search

Results found in 26 databases (1 error)

Literature	
Bookshelf	471
MeSH	4
NLM Catalog	13
PubMed	11,717
PubMed Central	42,097

Genomes	
Assembly	0
BioCollections	0
BioProject	169
BioSample	922
Genome	0
Nucleotide	8,253
SRA	1,872
Taxonomy	0

Effective June 2024, NCBI's Genome, Gene and assembly resources will no longer be available. Check <https://www.ncbi.nlm.nih.gov/datasets/>

Clinical	
Clinical Trials.gov	68
dbGaP	1,279
dbSNP	0
dbVar	216
GTR	29
MedGen	33
OMIM	131

PubChem	
BioAssays	45
Compounds	0
Pathways	0
Substances	27

Genomes	
Assembly	0
BioCollections	0
BioProject	169
BioSample	922
Genome	0
Nucleotide	8,253
SRA	1,872
Taxonomy	0

Clinical	
Clinical Trials.gov	68
dbGaP	1,279
dbSNP	0
dbVar	216
GTR	29
MedGen	33
OMIM	131

PubChem	
BioAssays	45
Compounds	0
Pathways	0
Substances	27

[دانلود کتاب های دان...](#)[rosetta_stuff](#)[Editorial Manager®](#)[Paraphrasing Tool...](#)[FDA & EMA's New...](#)[Review Reports: D...](#)[Intertek Pharmace...](#)[All Bookmarks](#)

An official website of the United States government [Here's how you know](#) ▾

 Search NCBI ...[Log in](#)[NCBI Datasets](#)[Taxonomy](#)[Genome](#)[Gene](#)[Command-line tools](#)[Documentation](#)

NCBI Datasets

A one-stop shop for finding, browsing, and downloading genomic data

 Enter a species

Examples: [Amphiprion ocellaris](#) [Salmonella](#) [whooping crane](#)



Gene Expression Omnibus (GEO)

Dataset: not curated
Profile: curated

Bookshelf	4
MeSH	4
NLM Catalog	13
PubMed	11,717
PubMed Central	42,097

Genomes	
Assembly	0
BioCollections	0
BioProject	169
BioSample	922
Genome	0
Nucleotide	8,253
SRA	1,872
Taxonomy	0

Genes

Gene	247
GEO DataSets	1,276
GEO Profiles	5,246
HomoloGene	error
PopSet	93

Proteins

Conserved Domains	9
Identical Protein Groups	600
Protein	4,052
Protein Family Models	42
Structure	373

Clinical

ClinicalTrials.gov	68
ClinVar	1,279
dbGaP	0
dbSNP	0
dbVar	216
GTR	29
MedGen	33
OMIM	131

PubChem

BioAssays	45
Compounds	0
Pathways	0
Substances	27

What is Gene Expression Omnibus (GEO)?

- Public **repository** for high-throughput functional genomic data
- Archives and **freely** distributes microarray, next-generation sequencing, and other data like mass spect

Why Submit Your Data to GEO?

- Fulfill funders' and journals' requirements for publication
- **Long-term** archiving at a centralized repository
- Integration with other NCBI resources for increased **usability and visibility**

What kinds of data will GEO accept?

- Gene expression profiling by microarray or next-generation sequencing
- Non-coding RNA profiling by microarray or next-generation sequencing
- **Chromatin immunoprecipitation** (ChIP) profiling by microarray or next-generation sequencing
- **Genome methylation** profiling by microarray or next-generation sequencing
- High-throughput **RT-PCR Genome variation** profiling by array (arrayCGH)
- **SNP arrays**
- **Serial Analysis of Gene Expression** (SAGE)
- Protein arrays

Results found in 26 databases (1 error)

Literature	
Bookshelf	471
MeSH	4
NLM Catalog	13
PubMed	11,717
PubMed Central	42,097

Genes	
Gene	247
GEO DataSets	1,276
GEO Profiles	5,246
HomoloGene	error
PopSet	93

- Proteins
- Conserved Domains
- Identical Protein Groups
- Protein
- Protein Families
- Structures

genomic variation and evolutionary history

Category	Count
Genomes	0
Assembly	0
BioCollections	0
BioProject	169
BioSample	922
Genome	0
Nucleotide	8,253
SRA	1,872
Taxonomy	0

Clinical	
ClinicalTrials.gov	68
ClinVar	1,279
dbGaP	0
dbSNP	0
dbVar	216
GTR	29
MedGen	33
OMIM	131

The image shows a large yellow circle with a red border and a white center containing the number 0. The word "Substances" is written in black text at the bottom left. A small blue circle with the number 27 is in the bottom right corner.

genomic variation and its relationship to human health.

ClinVar

ClinVar Genomic variation as it relates to human health

Search by gene symbols, location, HGVS expressions, c-dot, p-dot, conditions, ...

Search ClinVar

Advanced search

About Access Submit Stats FTP Help



NM_000290.4(PGAM2):c.266A>C (p.Glu89A...)

Cite

Follow

Print

Download

On this page

Classification

Variant Data

Genes

Germline

Condition

Submission

Function

Citations

Text mined

Germline

Classification
★☆☆☆ (3)



Conflicting classifications of pathogenicity
Likely pathogenic(1); Uncertain significance(1)
criteria provided, conflicting classifications



Somatic

No data submitted for somatic clinical impact

Somatic

No data submitted for oncogenicity



دانلود rosetta_stuff Editorial Manager® Paraphrasing Tool... FDA & EMA's New... Review Reports: D... Intertek Pharmace... ScienceDaily: You... VPN

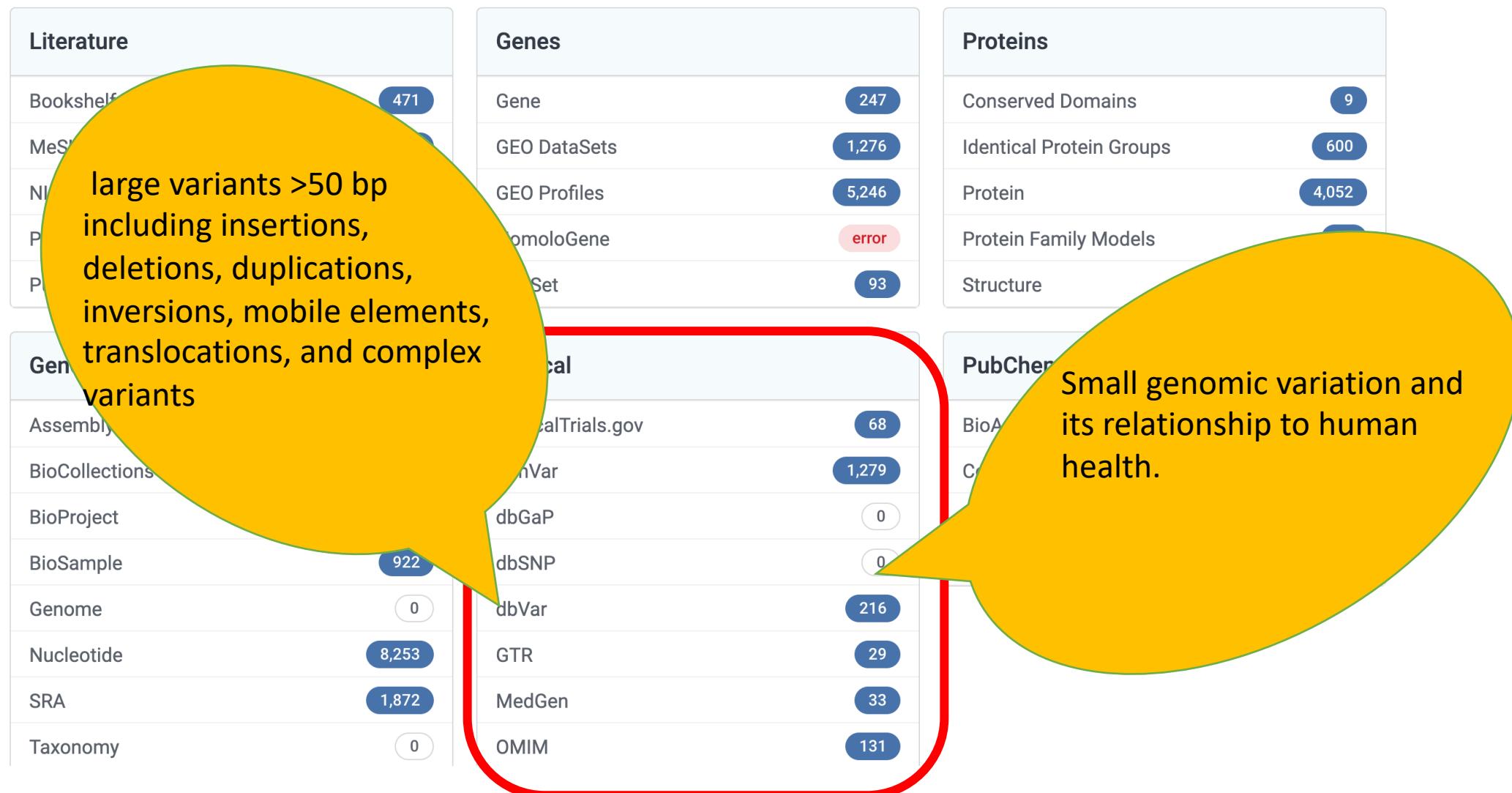
Search NCBI beta globin Search

Results found in 26 databases (1 error)

Literature	Genes	Proteins
Bookshelf 471	Gene 247	Conserved Domains 9
MeSH 4	GEO DataSets 1,276	Identical Protein Groups 600
NLM Catalog 13	GEO Profiles 5,246	Protein
PubMed 11,717	HomoloGene error	Protein Family Model
PubMed Central 42,097	PopSet 93	Structure
Genomes	Clinical	Pub
Assembly 0	ClinicalTrials.gov 68	B
BioCollections 0	ClinVar 1,279	Sub
BioProject 169	dbGaP 6	27
BioSample 922	dbSNP 0	
Genome 0	dbVar 216	
Nucleotide 8,253	GTR 29	
SRA 1,872	MedGen 33	
Taxonomy 0	OMIM 131	

studies that have investigated
the interaction of genotype
and phenotype in Humans.

Results found in 26 databases (1 error)



دانلود rosetta_stuff

Editorial Manager®

Paraphrasing Tool...

FDA & EMA's New...

Review Reports: D...

Intertek Pharmace...

ScienceDaily: You...

Search NCBI

beta globin

Search

Results found in 26 databases (1 error)

Online Mendelian Inheritance in Man (OMIM) is a catalog of human genes and genetic disorders. links to PubMed articles and sequence information.

Genes

Gene	247
GEO DataSets	1,276
GO Profiles	5,246
HomoloGene	error
Protein	93
PubChem	68
BioAssays	1,279
Compounds	0
Pathways	0
Substances	216
SRA	29
Taxonomy	33
OMIM	131

Proteins

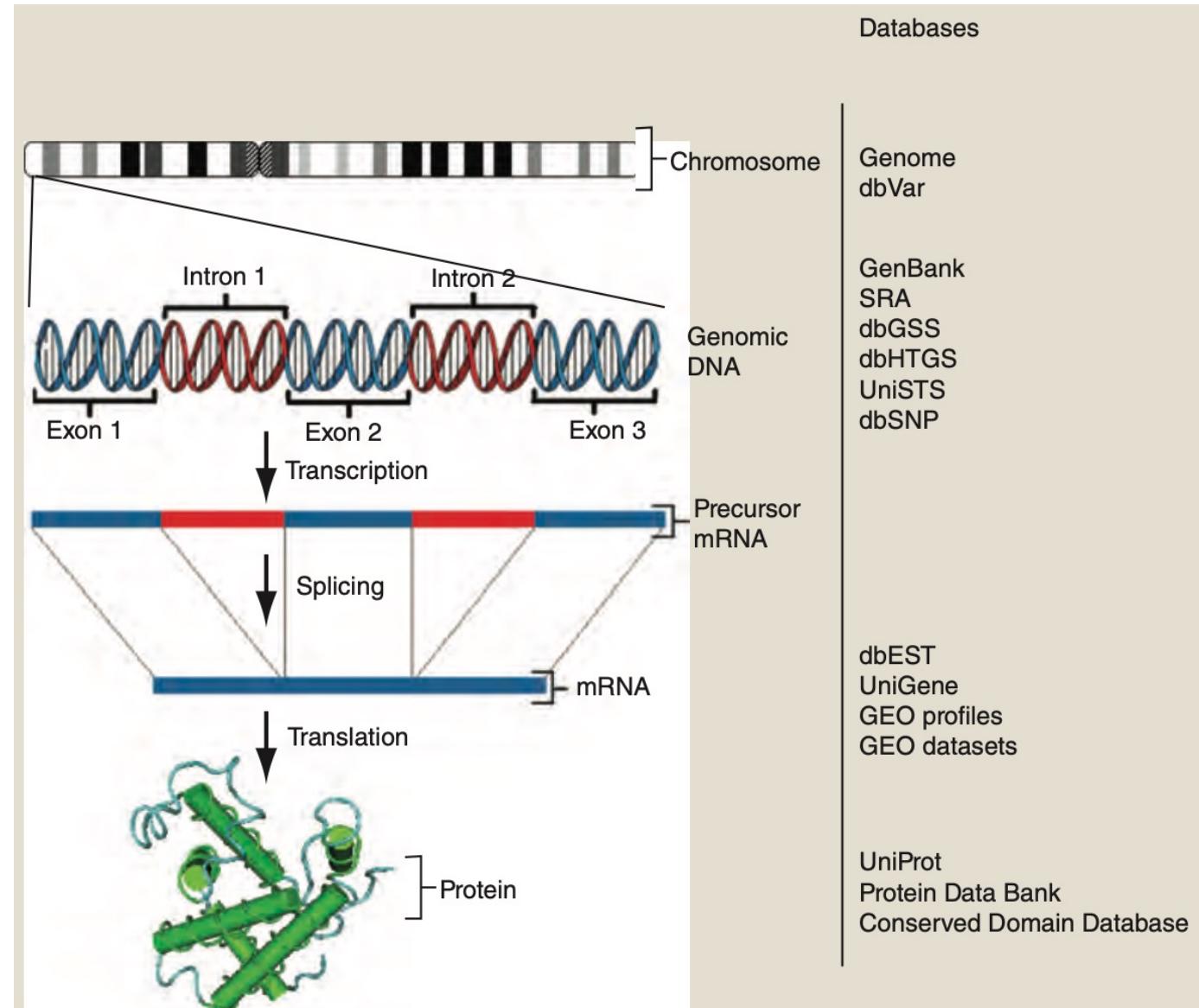
Conserved Domains	9
Identical Protein Groups	600
Protein	4,052
Protein Family Models	42
Structure	373

PubChem

BioAssays	45
Compounds	0
Pathways	0
Substances	27

Protein Databases

- UniProt
- Protein Data Bank (PDB)
- MMDB: Molecular Modeling Data Base
- Conserved domain database



UniProt

- most **comprehensive, centralized** protein sequence catalog
- consists of a combination of **three key databases**
- Swiss-Prot is considered the best-annotated protein database,
- translated EMBL (TrEMBL) Nucleotide Sequence Database Library provides automated (rather than manual) annotations of proteins
- PIR maintains the Protein Sequence Database,

UniProt

- UniProt is organized in three database layers
- The UniProt Knowledgebase (**UniProtKB**): central database
- The UniProt Reference Clusters (**UniRef**): nonredundant reference clusters based on UniProtKB, sharing at least 50%, 90%, or 100% identity
- The UniProt Archive, **UniParc**: stable, nonredundant archive of protein sequences from a wide variety of sources (including model organism databases, patent offices, RefSeq, and Ensembl)

Protein Data Bank (PDB)

- Established in 1971
 - Funded by NSF, DOE, NIH
 - Operated by Rutgers, SDSC, NIST
- Purpose: Make protein structure data available to the entire scientific community
- In the beginning: “less than a dozen” protein structures
- Currently has 27,112 protein structures
- Growing at 20% per year
- New structures 50 times larger than those in 1971 are commonplace

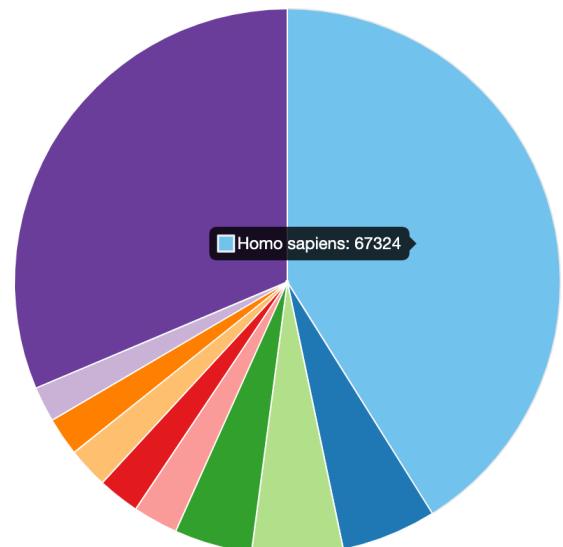
Number of Structures Released Annually Total Number of Entries Available

250000

By Scientific Name of Source Organism

The scientific name associated with this taxonomy code aggregated by the [NCBI Taxonomy Database](#). This name corresponds to the taxonomy identifier assigned by the PDB depositor, and includes natural, engineered, and synthetic source organisms.

200000



- Homo sapiens (67,324)
- Mus musculus (9,215)
- synthetic construct (8,846)
- Escherichia coli (7,553)
- Escherichia coli K-12 (4,370)
- Severe acute respiratory syndrome coronavirus 2 (4,064)
- Rattus norvegicus (3,953)
- Bos taurus (3,730)
- Saccharomyces cerevisiae (3,447)
- Others (51,368)

Number of Entries

100000

50000

0

1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024

Year

2024

Number of Structures Released Annually: 2756
Total Number of Entries Available: 216864

Why is the PDB Important?

- Rapid, extensive access to new structure data
- “Collective Leverage” for ...
 - Understanding molecular machinery
 - Rational drug design
 - Engineering new molecules
 - Structural genomics
 - etc...

PDB Website

<https://www.rcsb.org/>

Enter what you know...

https://www.rcsb.org/

RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19 MyPDB Contact us

RCSB PDB PROTEIN DATA BANK

216,864 Structures from the PDB
1,068,577 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entry ID(s), or sequence Advanced Search | Browse Annotations Include CSM Help

PDB-101 PDB EMDDataResource NAKB wwPDB Foundation PDB-Dev

Access Computed Structure Models (CSMs) of all available model organisms Learn more

Welcome

Deposit

Search

Visualize

Analyze

Download

Learn

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

- Experimentally-determined 3D structures from the **Protein Data Bank (PDB)** archive
- Computed Structure Models (CSM)** from AlphaFold DB and ModelArchive

These data can be explored in context of external annotations providing a structural view of biology.

Explore NEW Features

PDB-101 Training Resources

March Molecule of the Month

Hyaluronidases

Search for 8CUS pdb

Summary tab → Structure Summary

Overall model → Biological Assembly 1

PDB info → 8CUS

Structure quality info → Experimental Data Snapshot

Structure symmetry info → Global Symmetry: Asymmetric - C1

RCB PDB-101 wwPDB EMDDataResource NAKB wwPDB Foundation PDB-Dev

Structure Summary Structure Annotations Experiment Sequence Genome Versions

Display Files Download Files Data API

Biological Assembly 1 ?

Display Files Download Files Data API

Overall model

PDB info

Structure quality info

Structure symmetry info

8CUS

Accurate computational design of genetically encoded 3D protein crystals

PDB DOI: <https://doi.org/10.2210/pdb8CUS/pdb>

Classification: DE NOVO PROTEIN

Organism(s): synthetic construct

Expression System: Escherichia coli

Mutation(s): No ⓘ

Deposited: 2022-05-17 Released: 2023-11-01

Deposition Author(s): Bera, A.K., Li, Z., Baker, D.

Funding Organization(s): Howard Hughes Medical Institute (HHMI)

Experimental Data Snapshot

Method: X-RAY DIFFRACTION

Resolution: 3.98 Å

R-Value Free: 0.218

R-Value Work: 0.176

R-Value Observed: 0.179

wwPDB Validation ⓘ

3D Report Full Report

Metric Percentile Ranks Value

Rfree 0.216

Clashscore 4

Ramachandran outliers 0.9%

Sidechain outliers 2.6%

Worse Better

Percentile relative to all X-ray structures

Percentile relative to X-ray structures of similar resolution

This is version 1.1 of the entry. See complete history.

Macromolecule Content

- Total Structure Weight: 39.66 kDa ⓘ
- Atom Count: 2,551 ⓘ
- Modelled Residue Count: 336 ⓘ
- Deposited Residue Count: 364 ⓘ
- Unique protein chains: 2

Literature

Accurate computational design of three-dimensional protein crystals.

Li, Z., Wang, S., Nattermann, U., Bera, A.K., Borst, A.J., Yaman, M.Y., Bick, M.J., Yang, E.C., Sheffler, W., Lee, B., Seifert, S., Hura, G.L., Nguyen, H., Kang, A., Dalal, R., Lubner, J.M., Hsia, Y., Haddox, H., Courbet, A., Dowling, Q., Miranda, M., Favor, A., Etemadi, A., Edman, N.I., Yang, W., Weidle, C., Sankaran, B., Negahdari, B., Ross, M.B., Ginger, D.S., Baker, D.
 (2023) Nat Mater 22: 1556-1563

PubMed: [37845322](#) Search on PubMed
 DOI: <https://doi.org/10.1038/s41563-023-01683-1>

Primary Citation of Related Structures:
 8CUS, 8CUT, 8CUU, 8CUV, 8CUW, 8CUX, 8CWS, 8CWY, 8CWZ, 8FAR, 8SZZ

PubMed Abstract:
 Protein crystallization plays a central role in structural biology. Despite this, the process of crystallization remains poorly understood and highly empirical, with crystal contacts, lattice packing arrangements and space group preferences being largely unpredictable. Programming protein crystallization through precisely engineered side-...
[View More](#)

Organizational Affiliation:
 Department of Biochemistry, University of Washington, Seattle, WA, USA.

Experimental Data & Validation

Experimental Data

Method: X-RAY DIFFRACTION
 Resolution: 3.98 Å
 R-Value Free: 0.218
 R-Value Work: 0.176
 R-Value Observed: 0.179
 Space Group: [I4 3 2](#)

Unit Cell:

Length (Å)	Angle (°)
a = 235.728	α = 90
b = 235.728	β = 90
c = 235.728	γ = 90

Software Package:

Software Name	Purpose
PHENIX	refinement
PHENIX	refinement
XDS	data reduction
HKL-3000	data scaling
PHASER	phasing

[View more in-depth experimental data](#)

Macromolecules

Find similar proteins by: Sequence (by identity cutoff) | 3D Structure

Entity ID: 1

Molecule	Chains ⓘ	Sequence Length	Organism	Details	Image
I432-1(NaCl) Chain A	A	148	synthetic construct	Mutation(s): 0 ⓘ	

Entity Groups ⓘ

Sequence Clusters [30% Identity](#) [50% Identity](#) [70% Identity](#) [90% Identity](#) [95% Identity](#) [100% Identity](#)

Sequence Annotations

Reference Sequence 8CUS_1

8CUS_1 UNMODELED A HYDROPHITY DISORDER DISORDERED BINDING

MRGHIIHHHHGSSALAYVMLGLLSSLNRLSLAAEAYKKAIELDPNDAIALLLGVSLEKLKRLDEAAEAYKKAIELKPNDASAKELGKVLEKLGRLEAAKAYAEAIKLDPSOAEAKELGKVLEKLQLAERAYQLAELDPND

Manuscript info

Chain info

Experimental data

Structure Summary

Structure

Annotations

Experiment

Sequence

Genome

Versions

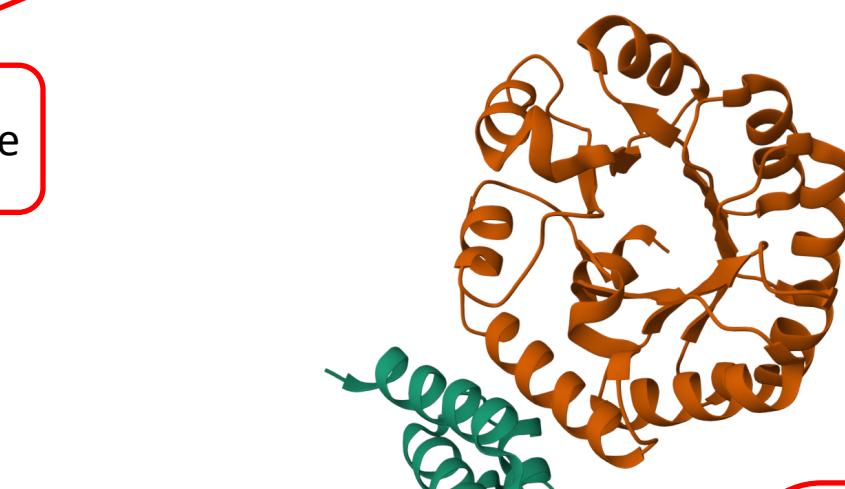
8CUS₁

Accurate computational design of genetically encoded 3D protein crystals

Structure Tab

Sequence of	8CUS Accur...	Chain	1: I432-1(NaC...)	A
12	22	32	42	52
RGHHHHHHGSSALAYVMLGLLSSLNRSLAAEAYKKAIELDPNDALAWLLGSVLEKLKRLDEAAEAYKKAIELKPNDASAWKELGVLEKLGRDLDEAKAYAEAIAKLDPSD				

AEAAKELGKVLEKLGQLELAERAYQLIAIELDPND

**Protein Sequence****Display Files** ▾ **Download Files** ▾**Structure**

8CUS | Accurate computational de...

Type Assembly

Asm Id 1: Author And Softwar...

Dynamic Bonds × Off

Nothing Focused

Measurements**Structure Motif Search****Components** 8CUS

Preset + Add

Polymer Cartoon

Unit Cell I 4 3 2

Density**Quality Assessment****Assembly Symmetry****Display Files** ▾ **Download Files** ▾

FASTA Sequence
mmCIF Format
mmCIF Format (Header)
PDB Format
PDB Format (Header)

Display Files ▾ **Download Files** ▾**FASTA Sequence**

PDBx/mmCIF Format

PDBx/mmCIF Format (gz)

BinaryCIF Format (gz)

PDB Format

PDB Format (gz)

PDBML/XML Format (gz)

Structure Factors (CIF)

Structure Factors (CIF - gz)

Validation Full PDF

Validation (XML - gz)

Validation (CIF - gz)

Biological Assembly 1 (CIF - gz) ⓘ

Biological Assembly 1 (PDB - gz)

fo-fc Map (DSN6)

2fo-fc Map (DSN6)

Map Coefficients (MTZ format)

[Structure Summary](#)[Structure](#)[Annotations](#)[Experiment](#)[Sequence](#)[Genome](#)[Versions](#)

8CUS

[Display Files ▾](#)[Download Files ▾](#)

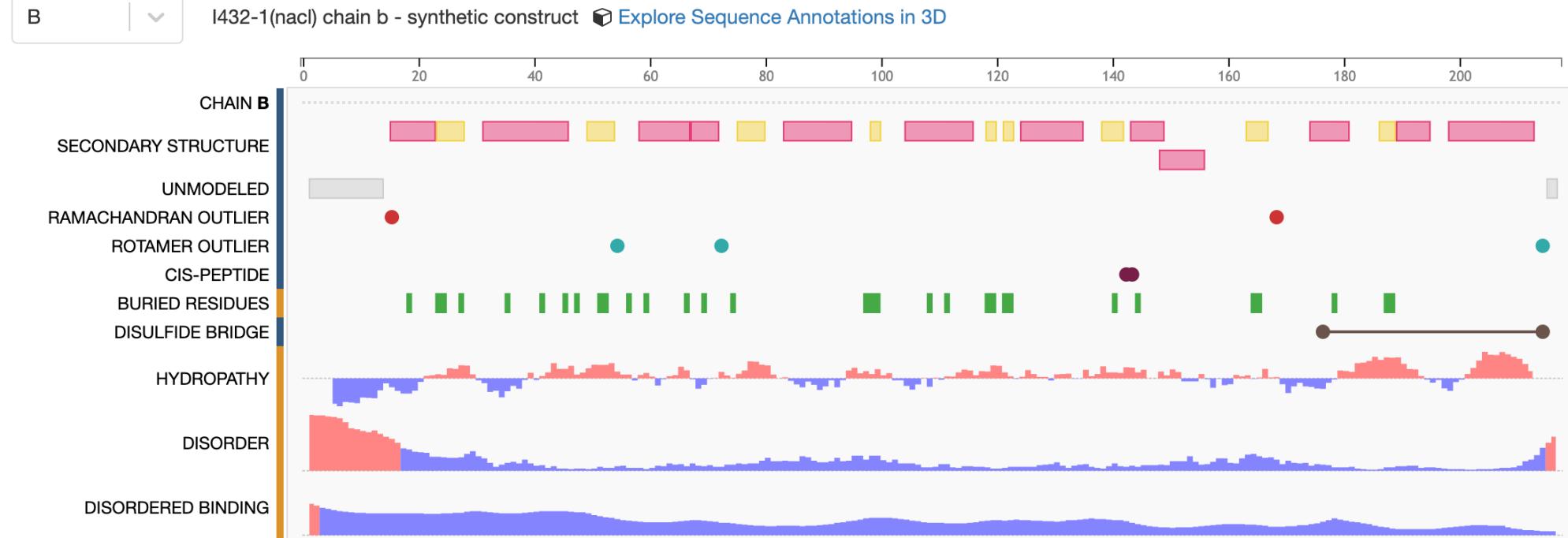
Accurate computational design of genetically encoded 3D protein crystals

[Help](#)**CHAIN**

B

I432-1(nacl) chain b - synthetic construct [Explore Sequence Annotations in 3D](#)

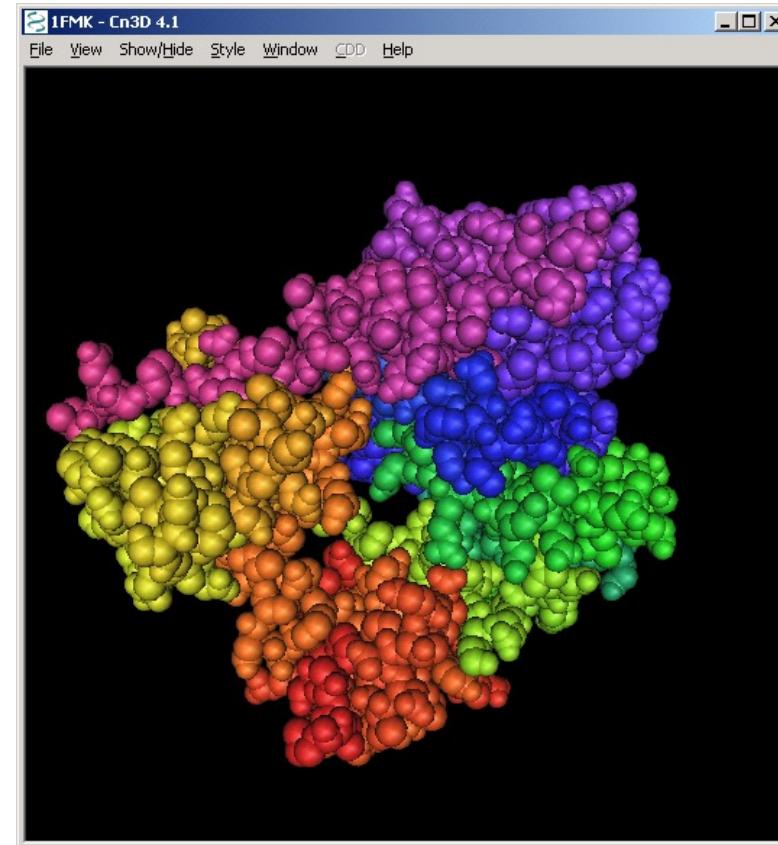
Protein sequence properties



MMDB: Molecular Modeling Data Base

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

Cn3D 4.1: C-Src



1FMK - Sequence/Alignment Viewer

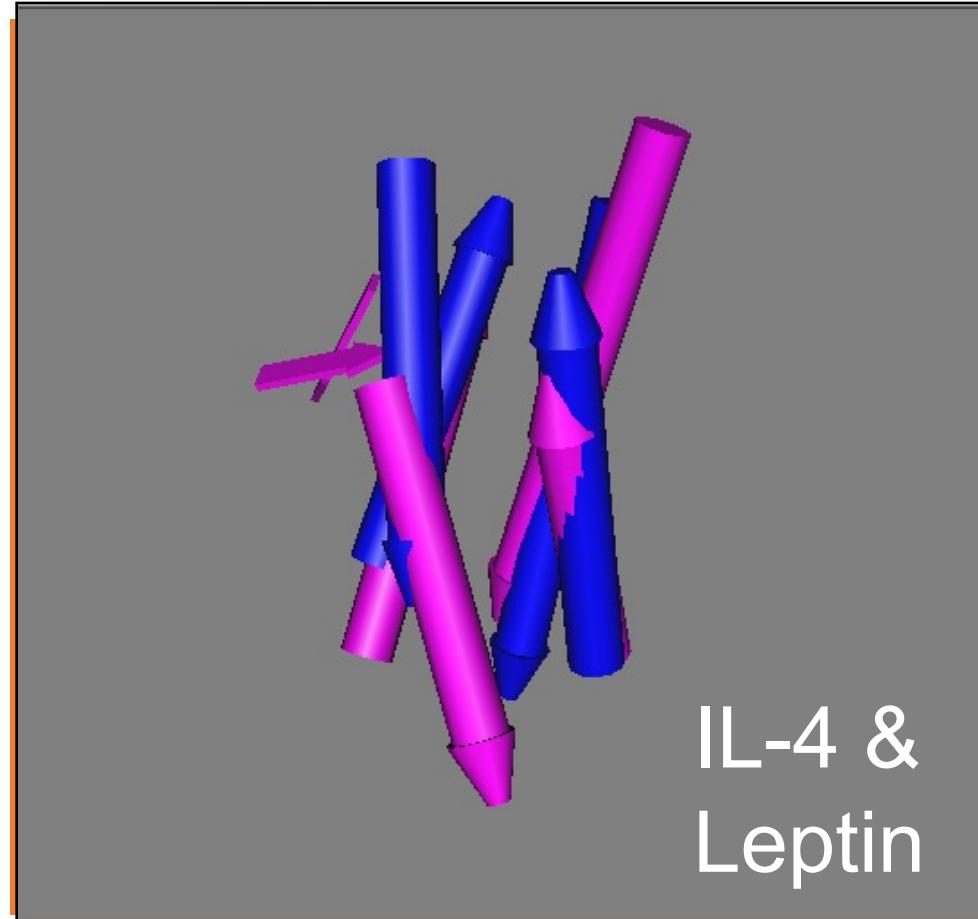
View Edit Mouse Mode Unaligned Justification Imports

1FMK	i v n n t e g d w w l a h s l s t g q t g y i p s n y v a p s d s i q a e e w y f g k i t r r e s e r l l n a e n p r g t f l v r e
------	---

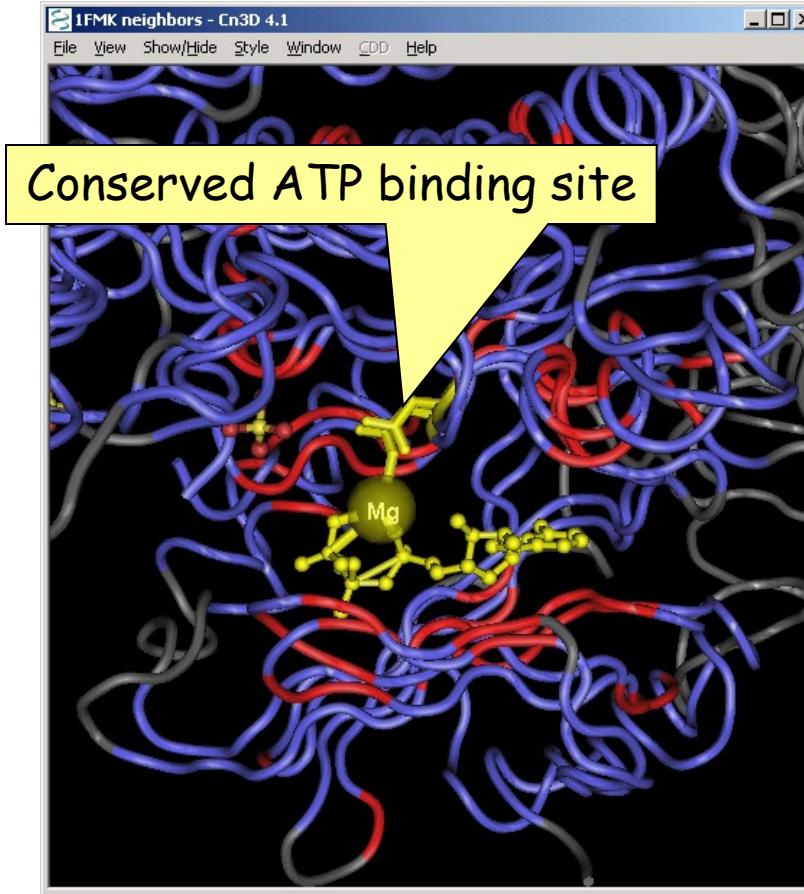
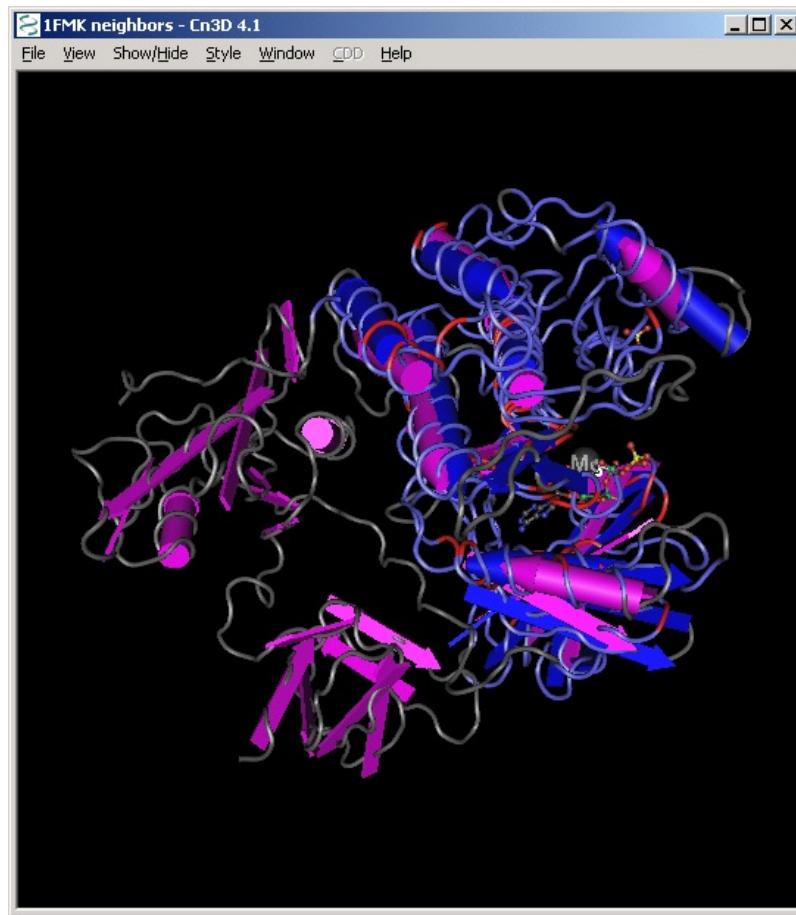
VAST: Structure Neighbors

Vector Alignment Search Tool

For each protein chain,
locate SSEs (secondary
structure elements),
and represent them as
individual vectors.
align the vectors



Cn3D 4.1: Structural Alignment



1FMK neighbors - Sequence/Ali

Src Kinase H. sapiens

View Edit Mouse Mode Unalign

1FMK RLPQLVDMAAQIASGMAYVERMNYVRDRLRAANILVGE~~~~~NLVCKVADFLGLARLi edney~~~
1CSN SVKTVAMAAKQMLARVQS I HEKSLVYRDIKPDNFLIGRpnsknANMIYVVDFGMVKfy rdpvt kqh

Casein kinase S. pombe

This panel shows a sequence alignment between two kinase proteins. The top row shows the sequence for Src Kinase from *H. sapiens*, and the bottom row shows the sequence for Casein kinase from *S. pombe*. The sequences are color-coded, and a yellow box highlights a conserved region near the C-terminal end of the sequence.

NCBI's Conserved Domain Database

- Multiple sequence alignments
- PSI-BLAST –based score matrices
- Sources SMART, PFAM, COGs, New NCBI curated domains
 - structure informed alignments
- Stats:
 - COGS 4,873
 - Pfam 5,193
 - Smart 653
 - NCBI CDD 316

NCBI CD: Tyrosine Kinase

 NCBI

Conserved Domain Database

PubMed	Nucleotide	Protein	Structure	CDD	Taxonomy	Help?				
CD: cd00192.1, TyrKc	PSSM-Id: 5392			Source: Smart						
Description: Tyrosine kinase, catalytic domain. Phosphotransferases; tyrosine-specific kinase subfamily. Enzymes with TyrKc domains belong to an extensive family of proteins which share a conserved catalytic core common to both serine/threonine and tyrosine protein kinases. Enzymatic activity of tyrosine protein kinases is controlled by phosphorylation of specific tyrosine residues in the activation segment of the catalytic domain or a C-terminal tyrosine (tail) residue with reversible conformational changes.										
Taxa: Eukaryota	References: 10 Pubmed Links			Related: cd00180 , COG3642 , smart00090 , COG2112 , pfam01163 , COG0515 , COG1718 , pfam00069 , smart00219 , smart00220						
Status: curated CD	Created: 1-Nov-2000			Updated: 17-Apr-2002						
Aligned: 78 rows	PSSM: 269 columns			Representative: Consensus						
Proteins: [Click here for CDART summary of Proteins containing cd00192]										
View 3D Structure with Cn3D using Virtual Bonds (To display structure, download Cn3D)										
View Alignment as Hypertext width 60 color at 2.0 bits feature catalytic loop										
Subset Rows up to 10 of the most diverse members										
.....*....*....*....*....*....*....*....*....*....										
Feature 1	consensus	1	KWEIPREDLTlg--kLGEGAFGEVYKGTLKgkd-----nksIDVAVKTLKEDase-----eqIKEFLRE	59						
1FGI_A		15	RWELPRDRVLVlg--gPLGEGAFGVVLAEGIglkd-----kpnrvTKVAVMLIKSDATE-----kdLSDLISE	76						
gi_532823		122	SWELNHADITMt--kVLGECAFGEVKLGLTLKmgs-----ttVDVAIKVAKLEKvtk-----eqIKEIIITE	179						
gi_7498408		98	PWEYMHSDVTAG--kVLGECAFGEVKFSGTLKlkd-----gssAEVAIKMTKVSAfls-----kikIKEMMNE	157						
gi_126497		875	ICLIPSSSELQTKldkKLGAGAFGTvFAGIYyPkra-----knvkIPVAAIKVQTxDQs-----qTDEMLEE	934						
gi_7498920		104	TVSSTYSTSQLSL--qLIGSDLRQLSLQGLLBSea-----vvLETIVGKGYFBNVyrgrmrpdpagrlipavktlkgerardiahIEKFLRE	187						
gi_7506508		72	PVEIDPSNQLKFs--sVLENGYFCKIRRGTLNinvtaee----knsnkLVDVLVKRPKNHIE-----KEMLENE	133						
gi_1103393		402	PLNVSDVNIREv--kQIGVGQFGAVVLAEMTlgsgsnvaslpkgsmnadgvALAVAKVILKPDVsd-----evLQSFDKE	473						
gi_478809		148	EWNVDMSQLFF---GLKFAHGAHSLRYHGvYk-----dEAVAVKILMVPEddngna-----lasr1EKQFIRE	207						
gi_7505469		143	SWQLRHNSVVYddtgKLGSGNFCIVYRGMLRkqg-----kiIPVAIKVSKNSDkesaa-----lmetRNLLAE	206						
.....*....*....*....*....*....*....*....*....*....										
Feature 1	consensus	110								
1FGI_A		60	AKIMRK-LKHPNIVRLLGVCTee-ePLMIVMEYMEGGDLTDYLKRp-----dsLSLDLsFALQIAKGMEYLEsKNFVHRDLAARNc	142						
gi_532823		77	MEMMKM <i>g</i> GHKNIINLLGACTqd-gPLYVIVEYASKGNLREYLQArppgleysy-npshnpeeqLSSKDLVsCAYQVARGMEYLAsKKCIHRDLAARNv	174						
gi_7498408		180	ARLMRG-FDHKNVVKCYGVAAid-ePLIVMELVPGGALDKYLQKNs-----svQWPBKLDLIAQVAGLAYIHSKNIICHRDIAARNI	260						
gi_126497		158	ARFIRN-FNHKNVVRLYGVAhde-qPLYILLELVKGGSQLDHMKKAt-----gtvVPiAEKVRCIGAARGIEYLNQNNCIHRDIAARNc	242						
gi_7498920		935	ATNMRf-LRHDNLLKIIGFCMhd-dGLKIVTIYRPLGNLQNFLKLHke-----NLGAREQVLYCYQIASGMQYLEKQRVVHRDLATRNv	1016						
gi_7506508		188	GVVMKh-LDHPHVSLLGISIspagNPWVVLPMEGGDlKTYTADPnr-----alCVLELLDFAHQVAQGMSYLAQHFVHRDLAARNc	270						
gi_1103393		134	LTI _{LR} lriGRHPNVIALIGYPLmgfRTVVVYESVELGNLQHVLKItknfelfvgvpeteFsgldefFCVsDLHsFAQIANGMEYLTHIPViHRLFLALRNi	233						
gi_478809		474	IKFMSq-LQHDSVQLLAICThs-kHPFIVMEYMEGDLNQFLQKYqmddd-----alyasnqippstLLYMAVQIASGMVYLSSlNYVHRDLATRNc	565						
gi_7505469		208	VTLLsr-LHHQNVIKFSAACRkp-pVYCIITEYLAEGSLRAYLHKLeh-----qtISLQKLIAFALDIARGMEYIHSQGVVHRDLKPENi	290						
		207	ARIMMN-YNNINVIRIYGVACdv-pPFMVCMEFCSGGSLEDALKKYgk-----dMEEFERQIILIDAARGMRYLHDKKCVHRDLASRNc	288						

Using Cn3D to model domains

