



INTRODUCTION TO BIOINFORMATICS

Please take the initial BIOINF525 questionnaire:
< <http://tinyurl.com/bioinf525-questions> >

Barry Grant
University of Michigan
www.thegrantlab.org



Barry Grant, Ph.D.
bjgrant@umich.edu



Ryan Mills, Ph.D.
remills@umich.edu



Lauren Jepsen (GSI)
ljepson@umich.edu

COURSE LOGISTICS

Lectures: Tuesdays 2:30-4:00 PM
Rm. 2062 Palmer Commons

Labs: Thursdays 2:30-4:00 PM
Rm. 2036 Palmer Commons

Website: <http://tinyurl.com/bioinf525-w17>

Lecture, lab and background reading material
plus homework and course announcements

MODULE OVERVIEW

Objective: Provide an introduction to the practice of bioinformatics as well as a practical guide to using common bioinformatics databases and algorithms

1.1. ▶ *Introduction to Bioinformatics*

1.2. ▶ *Sequence Alignment and Database Searching*

1.3 ▶ *Structural Bioinformatics*

1.4 ▶ *Genome Informatics: High Throughput Sequencing Applications and Analytical Methods*

TODAYS MENU

Overview of bioinformatics

- The what, why and how of bioinformatics?
- Major bioinformatics research areas.
- Skepticism and common problems with bioinformatics.

Bioinformatics databases and associated tools

- Primary, secondary and composite databases.
 - Nucleotide sequence databases (GenBank & RefSeq).
 - Protein sequence database (UniProt).
 - Composite databases (PFAM & OMIM).

Database usage vignette

- Searching with ENTREZ and BLAST.
- Reference slides and handout on major databases.

HOMework

- Complete the **initial course questionnaire**:
<http://tinyurl.com/bioinf525-questions>
- Check out the “**Background Reading**” material online:
[PDF1 \(bioinformatics review\)](#),
[PDF 2 \(bioinformatics challenges\)](#).
- Complete the **lecture 1.1 homework questions**:
<http://tinyurl.com/bioinf525-quiz1>

Q. What is Bioinformatics?

“Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.”

... Bioinformatics is a hybrid of biology and computer science

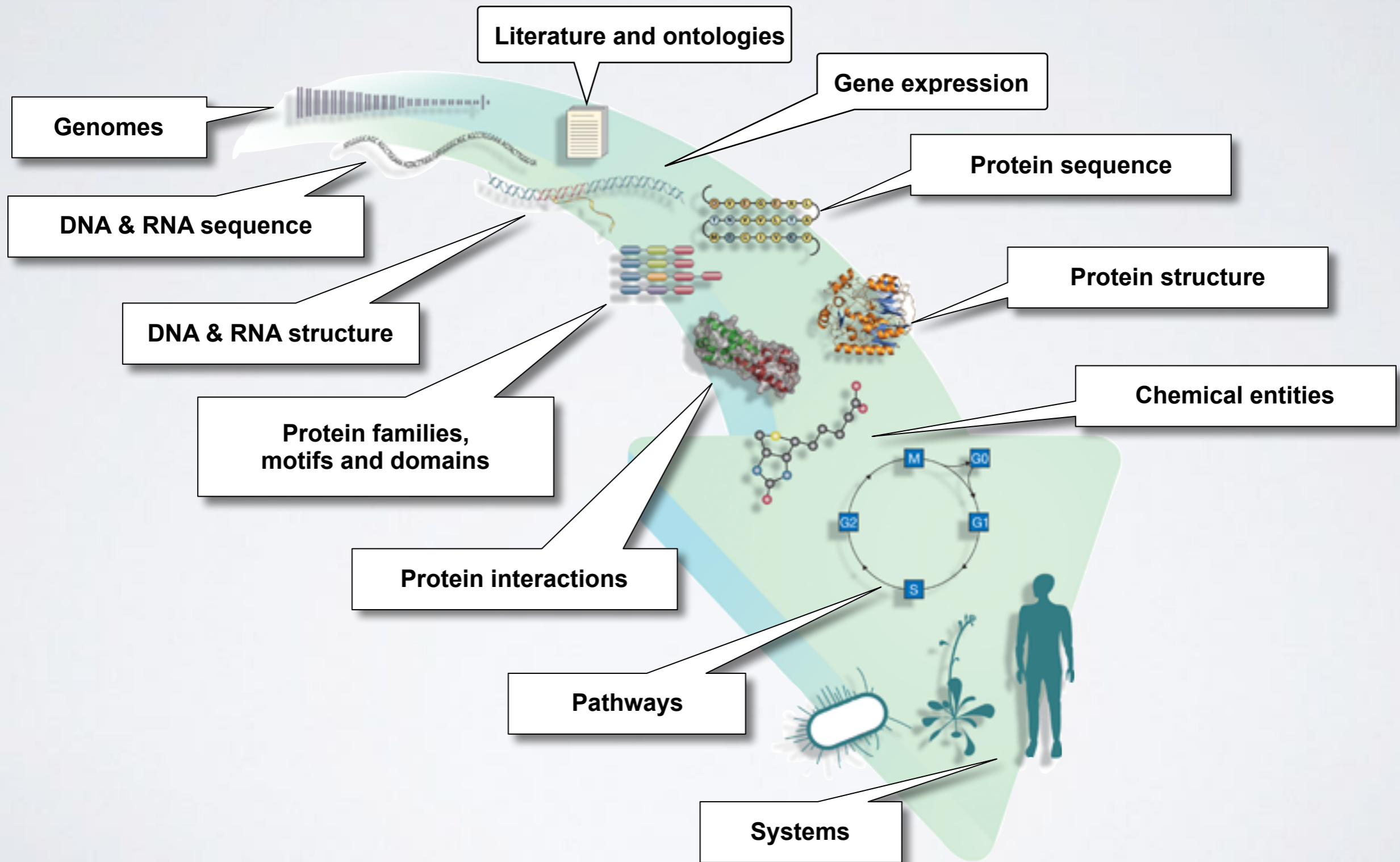
... **Bioinformatics is computer aided biology!**

Computer based management and analysis of biological and biomedical data with useful applications in many disciplines, particularly genomics, proteomics, metabolomics, etc...

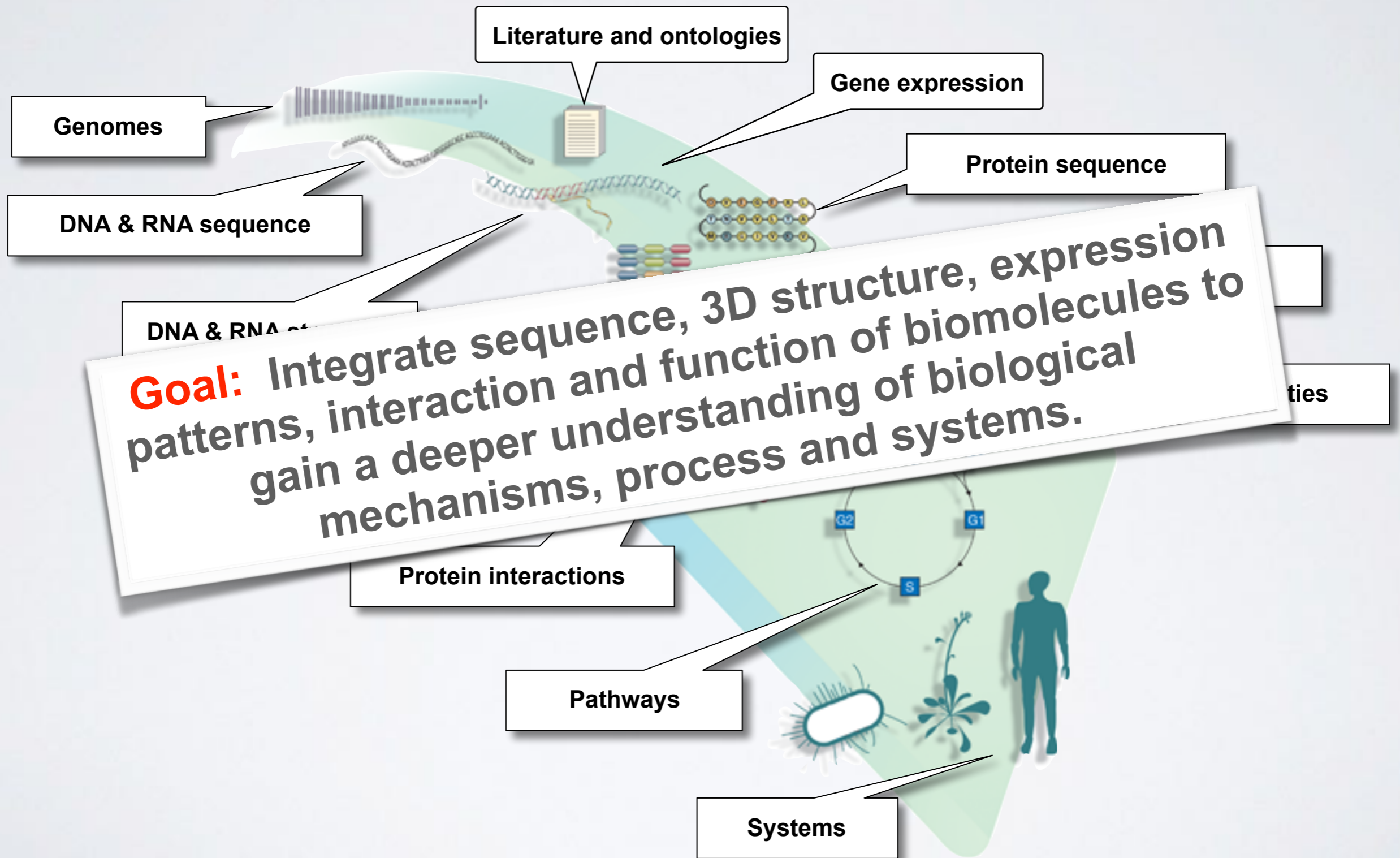
MORE DEFINITIONS

- ▶ “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “**informatics**” techniques (derived from disciplines such as applied maths, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a **large-scale**.
Luscombe NM, et al. Methods Inf Med. 2001;40:346.
- ▶ “Bioinformatics is research, development, or application of **computational approaches** for expanding the use of **biological, medical, behavioral or health data**, including those to **acquire, store, organize** and **analyze** such data.”
National Institutes of Health (NIH) (<http://tinyurl.com/l3gxr6b>)

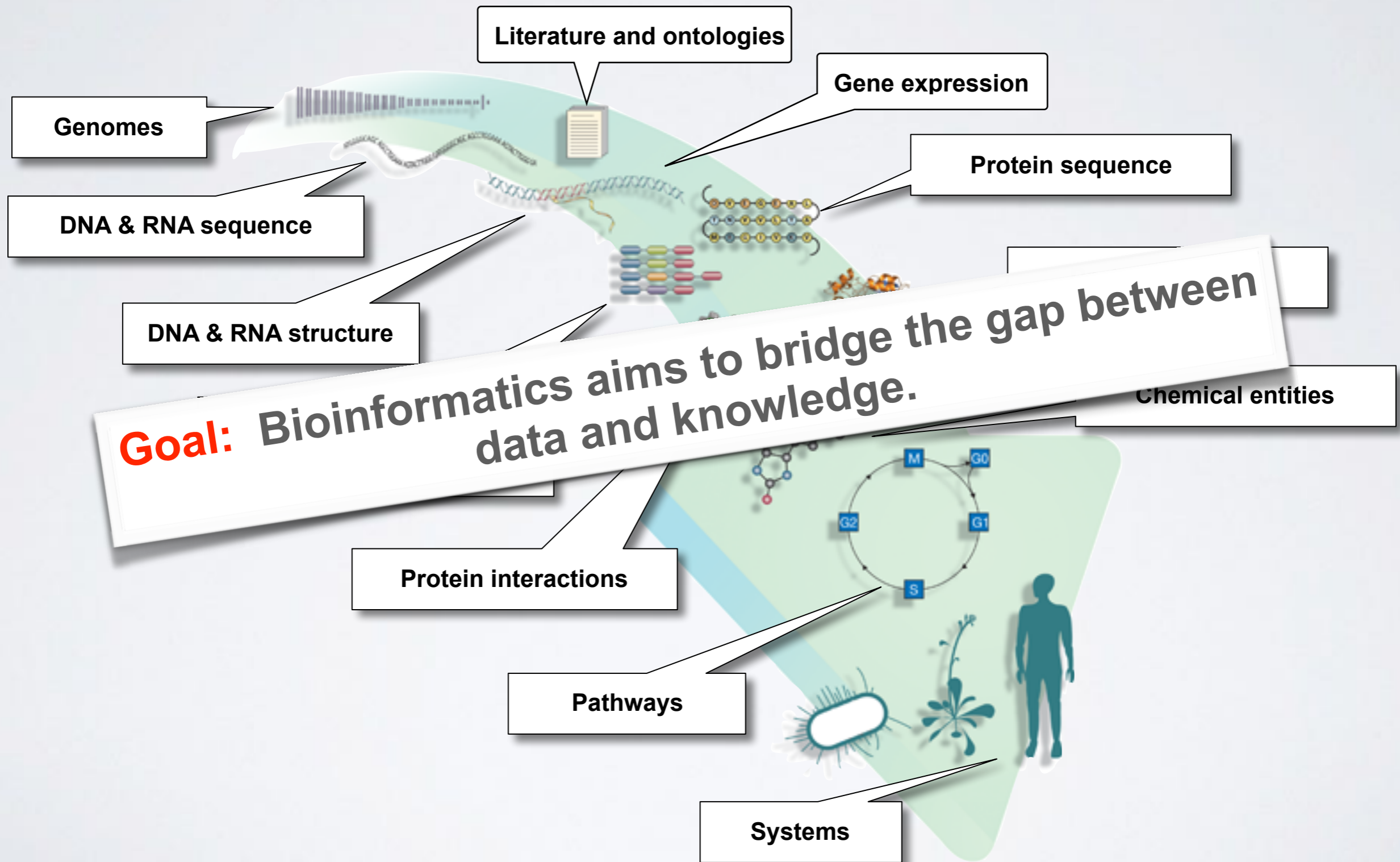
Major types of Bioinformatics Data



Major types of Bioinformatics Data



Major types of Bioinformatics Data



BIOINFORMATICS RESEARCH AREAS

Include but are not limited to:

- Organization, classification, dissemination and analysis of biological and biomedical data (particularly '-omics' data).
- Biological sequence analysis and phylogenetics.
- Genome organization and evolution.
- Regulation of gene expression and epigenetics.
- Biological pathways and networks in healthy & disease states.
- Protein structure prediction from sequence.
- Modeling and prediction of the biophysical properties of biomolecules for binding prediction and drug design.
- Design of biomolecular structure and function.

With applications to Biology, Medicine, Agriculture and Industry

Where did bioinformatics come from?

Bioinformatics arose as molecular biology began to be transformed by the emergence of molecular sequence and structural data

Recap: The key dogmas of molecular biology

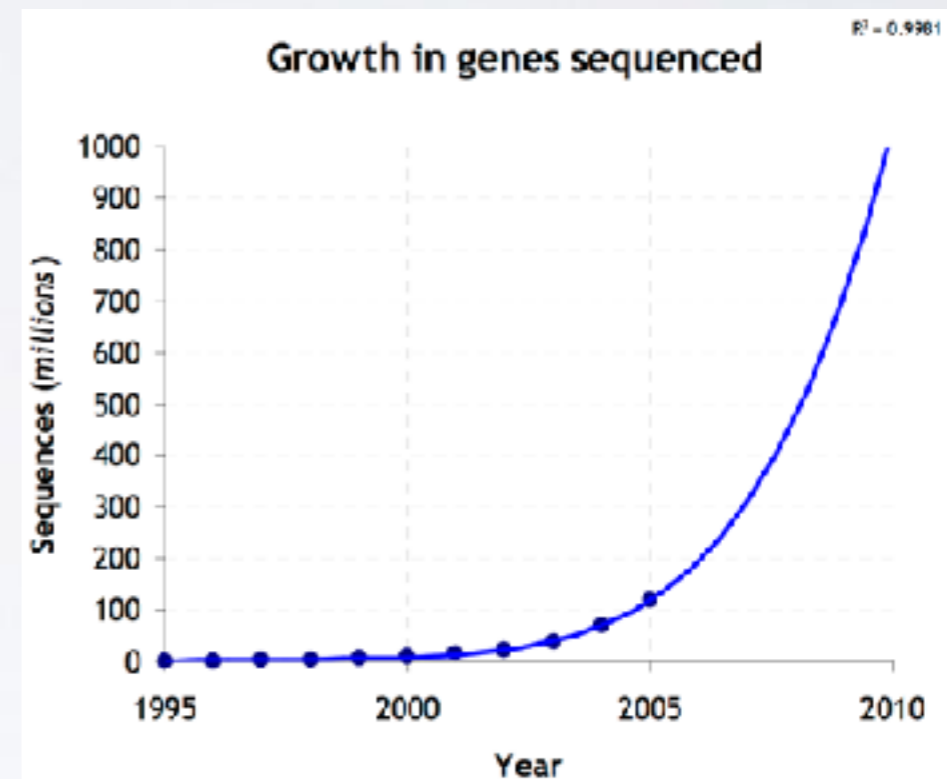
- *DNA sequence determines protein sequence.*
- *Protein sequence determines protein structure.*
- *Protein structure determines protein function.*
- *Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.*

Bioinformatics is now essential for the archiving, organization and analysis of data related to all these processes.

Why do we need Bioinformatics?

Bioinformatics is necessitated by the rapidly expanding quantities and complexity of biomolecular data

- Bioinformatics provides methods for the efficient:
 - ▶ **storage**
 - ▶ **annotation**
 - ▶ **search and retrieval**
 - ▶ **data integration**
 - ▶ **data mining and analysis**

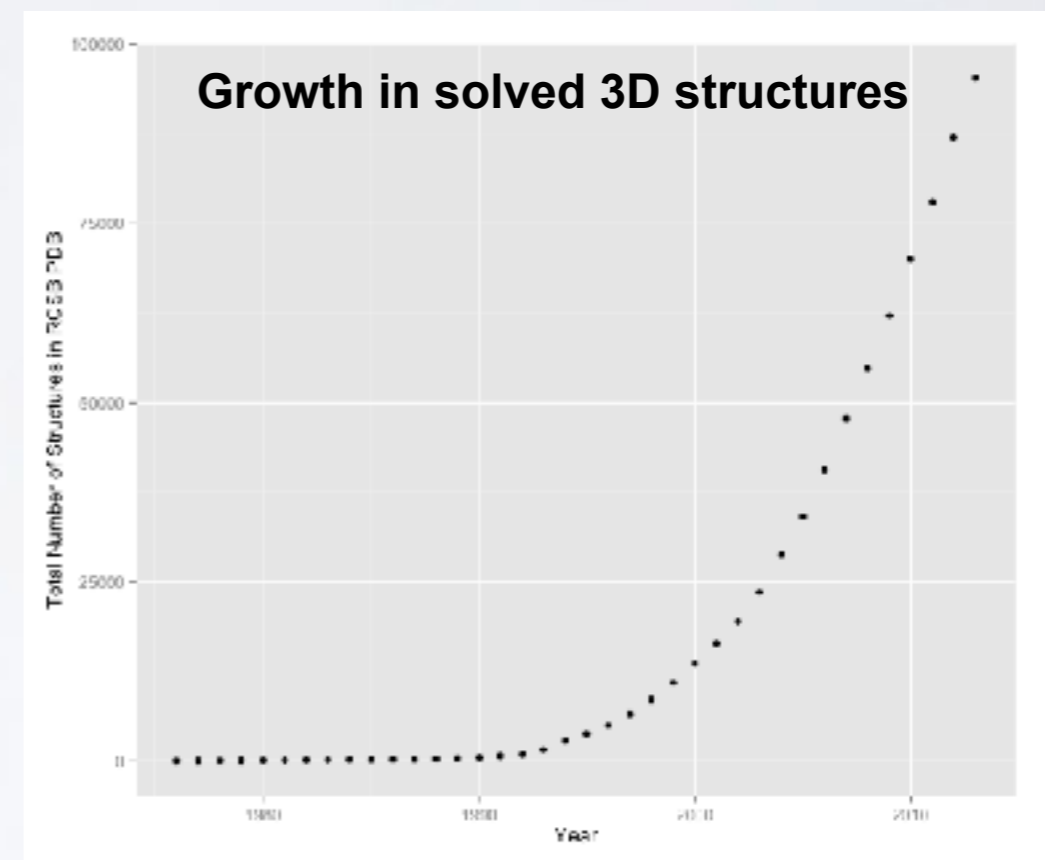


E.G. data from sequencing, structural genomics, microarrays, proteomics, new high throughput assays, *etc...*

Why do we need Bioinformatics?

Bioinformatics is necessitated by the rapidly expanding quantities and complexity of biomolecular data

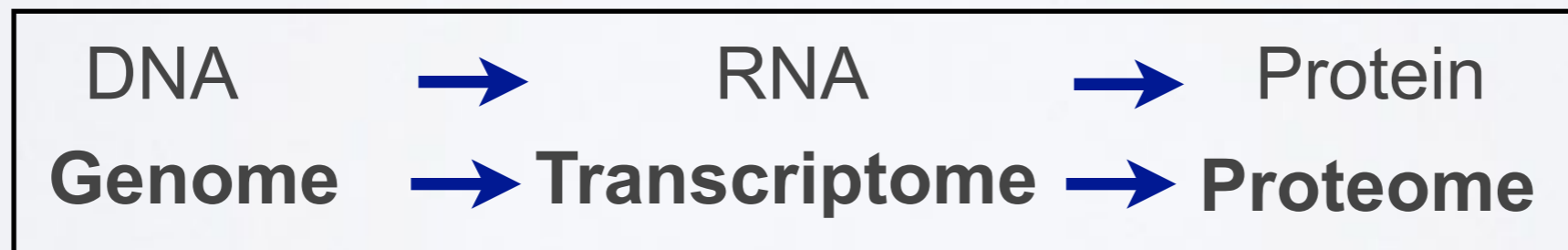
- Bioinformatics provides methods for the efficient:
 - ▶ **storage**
 - ▶ **annotation**
 - ▶ **search and retrieval**
 - ▶ **data integration**
 - ▶ **data mining and analysis**



E.G. data from sequencing, structural genomics, microarrays, proteomics, new high throughput assays, *etc...*

How do we do Bioinformatics?

- A “*bioinformatics approach*” involves the application of **computer algorithms**, **computer models** and **computer databases** with the broad goal of understanding the action of both individual genes, transcripts, proteins and large collections of these entities.



How do we actually do Bioinformatics?

Pre-packaged tools and databases

- ▶ Many online
- ▶ New tools and time consuming methods frequently require downloading
- ▶ Most are free to use

Tool development

- ▶ Mostly on a UNIX environment
- ▶ Knowledge of programming languages frequently required (Python, **R**, Perl, C Java, Fortran)
- ▶ May require specialized or high performance computing resources....

Skepticism & Bioinformatics

We have to approach computational results the same way we do wet-lab results:

- Do they make sense?
- Is it what we expected?
- Do we have adequate controls, and how did they come out?
- Modeling is modeling, but biology is different...
What does this model actually contribute?
- Avoid the miss-use of 'black boxes'

Common problems with Bioinformatics

Confusing multitude of tools available

- ▶ Each with many options and settable parameters

Most tools and databases are written by and for nerds

- ▶ Same is true of documentation - if any exists!

Most are developed independently

Notable exceptions are found at the:

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

General Parameters

Max target sequences
Select the maximum number of aligned sequences to display

Short queries Automatically adjust parameters for short input sequences

Expect threshold

Word size

Max matches in a query range

Scoring Parameters

Matrix

Gap Costs

Compositional adjustments

Filters and Masking

Filter Low complexity regions

Mask Mask for lookup table only
 Mask lower case letters

PSI/PHI/DELTA BLAST

Upload PSSM no file selected
Optional

PSI-BLAST Threshold

Pseudocount

Even Blast has many settable parameters

Related tools with different terminology

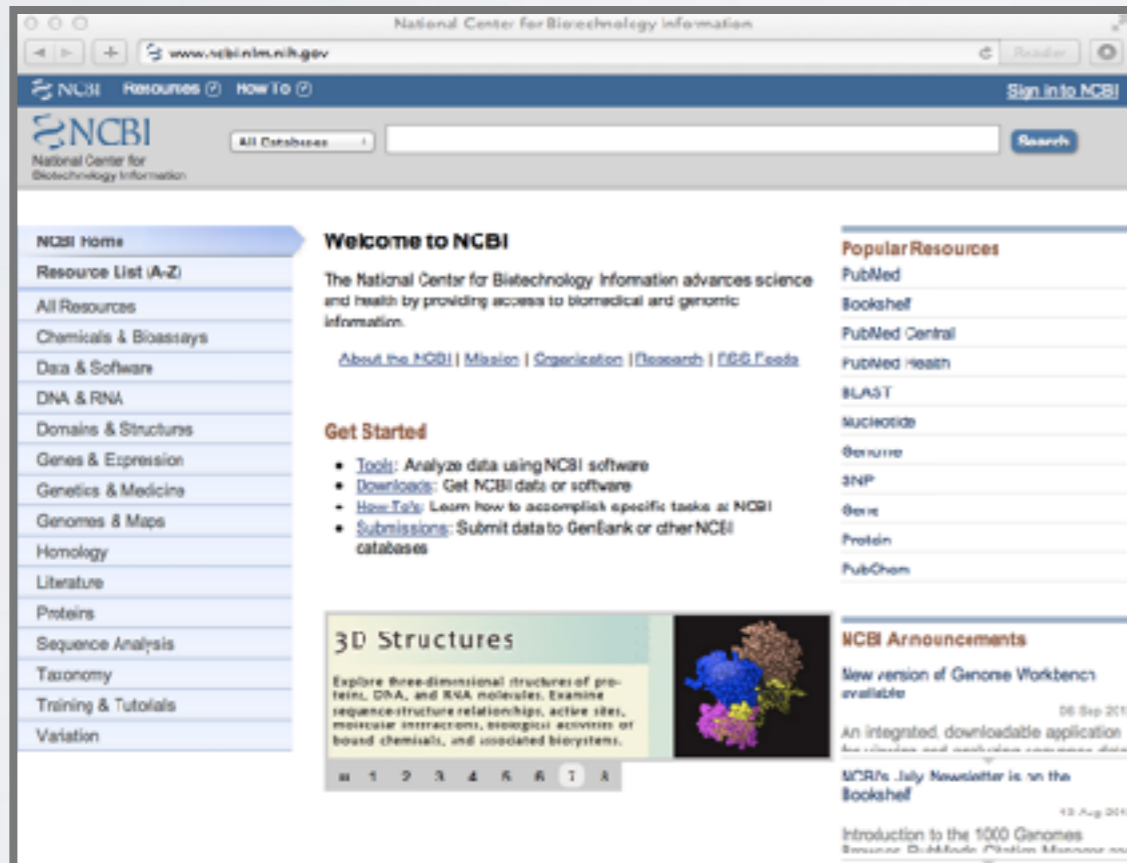
STEP 3 - Set your parameters

PROGRAM

MATRIX	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE
<input type="text" value="BLOSUM50"/>	<input type="text" value="-10"/>	<input type="text" value="-2"/>	<input type="text" value="2"/>	<input type="text" value="10"/>	<input type="text" value="0 (default)"/>
DNA STRAND	HISTOGRAM	FILTER	STATISTICAL ESTIMATES		
<input type="text" value="N/A"/>	<input type="text" value="no"/>	<input type="text" value="none"/>	<input type="text" value="Regress"/>		
SCORES	ALIGNMENTS	SEQUENCE RANGE	DATABASE RANGE	MULTI HSPs	
<input type="text" value="50"/>	<input type="text" value="50"/>	<input type="text" value="START-END"/>	<input type="text" value="START-END"/>	<input type="text" value="no"/>	
SCORE FORMAT					
<input type="text" value="Default"/>					

Key Online Bioinformatics Resources: NCBI & EBI

The NCBI and EBI are invaluable, publicly available resources for biomedical research



The screenshot shows the NCBI website homepage. The browser address bar displays 'www.ncbi.nlm.nih.gov'. The page features a navigation menu on the left with categories like 'All Resources', 'Chemicals & Bioassays', 'Data & Software', 'DNA & RNA', 'Domains & Structures', 'Genes & Expression', 'Genetics & Medicine', 'Genomes & Maps', 'Homology', 'Literature', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Variation'. The main content area includes a 'Welcome to NCBI' message, a 'Get Started' section with links to 'Tools', 'Downloads', 'How To's', and 'Submissions', and a 'Popular Resources' list containing 'PubMed', 'Bookshelf', 'PubMed Central', 'PubMed Health', 'BLAST', 'Nucleotide', 'Genome', 'SNP', 'Gene', 'Protein', and 'PubChem'. There is also a '3D Structures' section and 'NCBI Announcements'.

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



The screenshot shows the EBI website homepage. The browser address bar displays 'EMBL-European Bioinformatics Institute'. The page features a header with 'The European Bioinformatics Institute' and 'Part of the European Molecular Biology Laboratory'. The main content area includes a search bar with the text 'Find a gene, protein or chemical:', a 'Popular' section with links to 'Services', 'Research', 'Training', 'Industry', 'European Coordination', and 'EMBL ALUMNI', and a 'News from EMBL-EBI' section. There is also a 'Visit EMBL.org' section and a 'Partners Animal Genome conferences' section.

<https://www.ebi.ac.uk>

National Center for Biotechnology Information (NCBI)

- Created in 1988 as a part of the National Library of Medicine (NLM) at the National Institutes of Health
- NCBI's mission includes:
 - ▶ Establish **public databases**
 - ▶ Develop **software tools**
 - ▶ **Education** on and dissemination of biomedical information
- We will cover a number of core NCBI databases and software tools in the lecture



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

National Center for Biotechnology Information

www.ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

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 advances science and health by providing access to biomedical and genomic information.

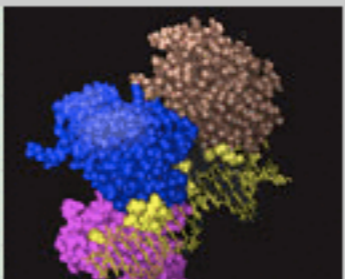
[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

Get Started

- [Tools](#): Analyze data using NCBI software
- [Downloads](#): Get NCBI data or software
- [How-To's](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

3D Structures

Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals, and associated biosystems.



Popular Resources

PubMed

Bookshelf

PubMed Central

PubMed Health

BLAST

Nucleotide

Genome

SNP

Gene

Protein

PubChem

NCBI Announcements

New version of Genome Workbench available

06 Sep

An integrated, downloadable applicati

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

The image shows a screenshot of the National Center for Biotechnology Information (NCBI) website. The browser address bar displays 'www.ncbi.nlm.nih.gov'. The page features a navigation menu on the left with categories like 'NCBI Home', 'Resource List (A-Z)', and various biological topics. The main content area includes a 'Welcome to NCBI' message and a 'Get Started' section with links to 'Tools', 'Downloads', 'How-To's', and 'Submissions'. A 'Popular Resources' box is overlaid on the right side of the page, listing several key services: PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. Red arrows point to PubMed, BLAST, and SNP, while a red bracket groups Nucleotide, Genome, SNP, Gene, and Protein.

National Center for Biotechnology Information

www.ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

All Databases Search

Popular Resources

- PubMed ←
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST ←
- Nucleotide
- Genome
- SNP ←
- Gene
- Protein
- PubChem

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 provides access to a wide range of biological information.

[About the NCBI](#) | [Mission](#) | [Our Services](#)

Get Started

- [Tools](#): Analyze data using NCBI tools
- [Downloads](#): Get NCBI data
- [How-To's](#): Learn how to access NCBI resources
- [Submissions](#): Submit data to NCBI databases

3D Structures

Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals, and associated biosystems.

Resources

Central Health

Announcements

New version of Genome Workbench available

06 Sep

An integrated, downloadable application

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

The screenshot shows the NCBI website homepage. At the top, there is a navigation bar with "NCBI", "Resources", and "How To" menus, and a "Sign in to NCBI" link. Below this is a search bar with a dropdown menu set to "All Databases" and a "Search" button. The main content area features a "Welcome to NCBI" message with the tagline "The National Center for Biotechnology Information advances science". To the left, there are links for "NCBI Home" and "Resource List (A-Z)". To the right, there is a "Popular Resources" section with a link to "PubMed".

Notable NCBI databases include:
GenBank, **RefSeq**, **PubMed**, dbSNP
and the search tools **ENTREZ** and **BLAST**

This screenshot shows a section of the NCBI website with a sidebar on the left containing links to "Homology", "Literature", "Proteins", "Sequence Analysis", "Taxonomy", "Training & Tutorials", and "Variation". The main content area is titled "databases" and features a "3D Structures" section with a description: "Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals, and associated biosystems." To the right of this text is a 3D molecular model. Further right, there are links for "Protein" and "PubChem". At the bottom right, there is an "NCBI Announcements" section with a headline "New version of Genome Workbench available" dated "06 Sep" and a sub-headline "An integrated, downloadable applicati".

Key Online Bioinformatics Resources: NCBI & EBI

The NCBI and EBI are invaluable, publicly available resources for biomedical research



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



<https://www.ebi.ac.uk>

European Bioinformatics Institute (EBI)

- Created in 1997 as a part of the European Molecular Biology Laboratory (EMBL)
- EBI's mission includes:
 - ▶ providing freely available **data and bioinformatics services**
 - ▶ and providing advanced **bioinformatics training**
- We will briefly cover several EBI databases and tools that have advantages over those offered at NCBI



The EBI maintains a number of high quality curated **secondary databases** and associated tools

The screenshot shows the EMBL-EBI website homepage. At the top, the browser address bar shows 'www.ebi.ac.uk'. The main header features the EMBL-EBI logo and navigation links for 'Services', 'Research', 'Training', and 'About us'. Below the header, the text reads 'The European Bioinformatics Institute' and 'Part of the European Molecular Biology Laboratory'. A paragraph describes the institute's mission: 'EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.' A search bar is provided with the prompt 'Find a gene, protein or chemical:' and a 'Search' button. Below the search bar, there are six colored tiles: 'Services' (teal), 'Research' (green), 'Training' (yellow), 'Industry' (blue), 'European Coordination' (orange), and 'EMBL ALUMNI' (white). The 'Services' tile is highlighted with a red border. To the right, a 'Popular' section lists links for 'Services', 'Research', 'Training', 'News', 'Jobs', 'Visit us', 'EMBL', and 'Contacts'. Below this is a 'Visit EMBL.org' section with the EMBL 40th anniversary logo (1974-2014). The 'Upcoming events' section features a banner for the 'Plant and Animal Genome conference (PAG XXIV)' on Sunday 10 - Tuesday 12 January 2016. At the bottom, there are three small image thumbnails.

The EBI maintains a number of high quality curated **secondary databases** and associated tools

The screenshot shows the EBI Services website. The browser address bar displays 'www.ebi.ac.uk/services'. The page features a teal header with the 'Services' title and navigation links for 'Services', 'Research', 'Training', and 'About us'. Below the header, there are tabs for 'Overview', 'A to Z', 'Data submission', and 'Support'. The main content area is titled 'Bioinformatics services' and includes a paragraph stating: 'We maintain the world's most comprehensive range of **freely available** and up-to-date molecular databases. Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our web services to access our resources programmatically. You can read more about our services in the journal *Nucleic Acids Research*.' To the right, a 'Popular' section lists services such as Ensembl, UniProt, PDBc, ArrayExpress, ChEMBL, BLAST, Europe PMC, Reactome, Train online, and Support. Below this is a 'Service news' section with a banner image of a butterfly and a protein structure. At the bottom right, there is a 'Training' section with a banner image of a person at a computer. The bottom of the page shows the start of a 'Programmatic access' section.

Services < EMBL-EBI

www.ebi.ac.uk/services

EMBL-EBI

Services Research Training About us

Services

Overview A to Z Data submission Support

Bioinformatics services

We maintain the world's most comprehensive range of **freely available** and up-to-date molecular databases. Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our web services to access our resources programmatically. You can read more about our services in the journal *Nucleic Acids Research*.

DNA & RNA
genes, genomes & variation

Gene expression
RNA, protein & metabolite expression

Proteins
sequences, families & motifs

Structures
Molecular & cellular structures

Systems
reactions, interactions & pathways

Chemical biology
chemogenomics & metabolomics

Ontologies
taxonomies & controlled vocabularies

Literature
Scientific publications & patents

Cross domain
cross-domain tools & resources

Popular

- Ensembl
- UniProt
- PDBc
- ArrayExpress
- ChEMBL
- BLAST
- Europe PMC
- Reactome
- Train online
- Support

Service news

Training

Programmatic access

The EBI maintains a number of high quality curated **secondary databases** and associated tools

The screenshot shows the EBI Services website. The main heading is "Services" with sub-navigation for "Overview", "A to Z", "Data submission", and "Support". The "Bioinformatics services" section describes the availability of molecular databases and tools. A grid of service categories is displayed, with "Proteins" highlighted. A "Popular" list on the right includes Ensembl, UniProt, PDBe, ArrayExpress, and ChEMBL. A "Training" banner is visible at the bottom right.

Services < EMBL-EBI

www.ebi.ac.uk/services

Services Research Training About us

Services

Overview A to Z Data submission Support

Bioinformatics services

We maintain the world's most comprehensive range of **freely available** and up-to-date molecular databases. Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our web services to access our resources programmatically. You can read more about our services in the journal Nucleic Acids Research.

DNA & RNA genes, genomes & variation	Gene expression RNA, protein & metabolite expression	Proteins sequences, families & motifs
Structures Molecular & cellular structures	Systems reactions, interactions & pathways	Chemical biology chemogenomics & metabolomics
Ontologies taxonomies & controlled vocabularies	Literature Scientific publications & patents	Cross domain cross-domain tools & resources

Programmatic access

Popular

- Ensembl
- UniProt
- PDBe
- ArrayExpress
- ChEMBL








Training

<https://www.ebi.ac.uk>

The EBI makes available a wider variety of **online tools** than NCBI

Proteins

Popular services

	UniProt: The Universal Protein Resource The gold-standard, comprehensive resource for protein sequence and functional annotation data.
	InterPro A database for the classification of proteins into families, domains and conserved sites.
	PRIDE: The Proteomics Identifications Database An archive of protein expression data determined by mass spectrometry.
	Pfam A database of hidden Markov models and alignments to describe conserved protein families and domains.
	Clustal Omega Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.
	HMMER - protein homology search Fast sensitive protein homology searches using profile hidden Markov models (HMMs). Variety of different search methods for querying against both sequence and HMM target databases.
	InterProScan 5 InterProScan 5 searches sequences against InterPro's predictive protein signatures. Please note that InterProScan 4.8 has been retired.

Quick links

- [Popular services in this category](#)
- [All services in this category](#)
- [Project websites in this category](#)

The EBI also provides a growing selection of **online tutorials** on EBI databases and tools

EMBL-EBI

Services Research Training About us

The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.

Find a gene, protein or chemical:

Search

Examples: [blast](#), [keratin](#), [bff1](#)...

Services

Research

Training

Industry

European Coordination

EMBL ALUMNI

Popular

- Services
- Research
- Training
- News
- Jobs
- Visit us
- EMBL
- Contacts

Visit **EMBL.org**

EMBL 40th anniversary logo

Upcoming events

Plant and Animal Genome conference (PAG XXIV)

Sunday 10 - Tuesday 12 January 2016

The EBI also provides a growing selection of **online tutorials** on EBI databases and tools

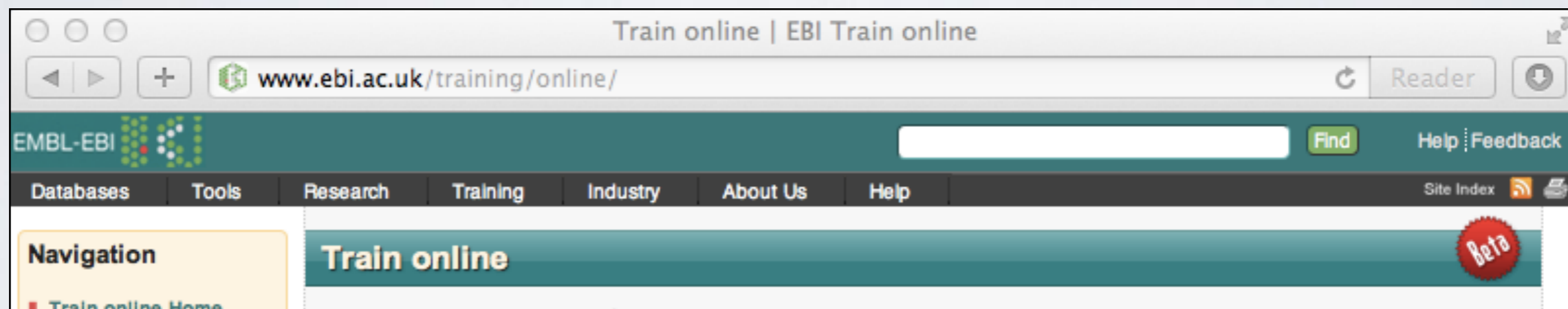
The screenshot shows a web browser window with the URL www.ebi.ac.uk/training/online/course/using-sequence-similarity-searching-tools-emb1-ebi. The page features the EMBL-EBI logo and navigation menus for Services, Research, Training, and About us. A prominent yellow banner reads "Train online". Below this, a breadcrumb trail indicates the current page: training > online > course-list > using-sequence-similarity-searching-tools-emb1-ebi.

The main content area is titled "Using sequence similarity searching tools at EMBL-EBI: webinar". On the left, a "Course content" sidebar lists the current webinar and "Contributors". A "Print Course" link is also visible. The central focus is a video player showing a webinar slide with the following text: "Using sequence similarity search tools at EMBL-EBI. Finding homologous sequences with BLAST, FASTA, PSI-Search etc." and a portrait of Andrew Cowley, with contact information: andrew.cowley@ebi.ac.uk and support@ebi.ac.uk. The video player shows a progress bar at 0:05 / 37:42.

On the right side, there are two utility boxes. The "Popular" box contains links for "Train online", "Find us", and "Funding". The "Find us at..." box lists various events and services: "Open days and career days", "Conference exhibitions", "EMBL courses and events", "Genome campus events", and "Science for schools".

Below the video player, a text block states: "This webinar focuses on how to use tools like **BLAST** and PSI-Search to find homologous sequences in EMBL-EBI databases, including tips on which tool and database to use, input formats, how to change parameters and how to interpret the results pages."

The EBI also provides a growing selection of **online tutorials** on EBI databases and tools



Notable EBI databases include:
ENA, UniProt, Ensembl
and the tools FASTA, BLAST, InterProScan,
MUSCLE, DALI, HMMER

Find a course

Browse by subject



[Genes and Genomes](#)



[Gene Expression](#)



[Interactions, Pathways and Networks](#)

**BIOINFORMATICS
DATABASES AND
ASSOCIATED TOOLS**

What is a database?

Computerized store of data that is organized to provide efficient retrieval.

- Uses standardized data (record) formats to enable computer handling

Key database features allow for:

- Adding, changing, removing and merging of records
- User-defined queries and extraction of specified records

Desirable features include:

- Contains the data you are interested in
- Allows fast data access
- Provides annotation and curation of entries
- Provides links to additional information (possibly in other databases)
- Allows you to make discoveries

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, Biolmage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty_cDB, DIP, DOGS, DOMO, DPD, DPlInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlilesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5 Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-U's, MPDB, MRR, MutBase, MycDB, NDB, NRSub, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc !!!!

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCCP, Beanref, Biolmage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVM, TKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, AP, ChickGBASE, Colibri, COPE, CottonDB, bEST, dbSTS, DDBJ, DGP, DictyDb, CDC, ECGC, EC02DBASE, OTHER, FlyBase, Link, G, HAEMB, H, HZRGbase, IMG, Kabat, KDNA, K, DB, Medline, Mendel, MEROPS, MGDB, MGI, MHC, MMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, Myc, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS- MODEL Repository, SWISS-PROT, TeIDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc !!!!

There are lots of Bioinformatics Databases

For a annotated listing of major bioinformatics databases please see the online handout

[Handout_Major_Databases.pdf](#) >

Side-note: Databases come in all shapes and sizes



Databases can be of variable quality and often there are multiple databases with overlapping content.

Finding Bioinformatics Databases

The screenshot shows a web browser window displaying the Oxford Journals website. The browser's address bar shows the URL www.oxfordjournals.org/nar/database/cat/3. The page title is "Oxford Journals | Life Sciences | Nucleic Acids Research | Database Summary Paper Categories". The main heading is "Nucleic Acids Research". Below the heading, there is a navigation menu with links for "ABOUT THIS JOURNAL", "CONTACT THIS JOURNAL", "SUBSCRIPTIONS", "CURRENT ISSUE", "ARCHIVE", and "SEARCH". The main content area is titled "2014 NAR Database Summary Paper Category List" and lists various database categories. A callout box highlights the URL <http://www.oxfordjournals.org/nar/database/c/>.

Oxford Journals | Life Sciences | Nucleic Acids Research | Database Summary Paper Categories

www.oxfordjournals.org/nar/database/cat/3

Oxford Journals | Life Sciences | Nucleic Acids Research | Database Summary Pa... The 2014 Nucleic Acids Research Database Issue and an updated NAR online M...

OXFORD JOURNALS CONTACT US MY BASKET MY ACCOUNT

Nucleic Acids Research

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRENT ISSUE ARCHIVE SEARCH

Oxford Journals > Life Sciences > Nucleic Acids Research > Database Summary Paper Categories

2014 NAR Database Summary Paper Category List

- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- Human and other Vertebrate Genomes
- Human Genes and Diseases
 - CancerResource
 - Protein Mutant Database
 - General human genetics databases
 - General polymorphism databases
 - Cancer gene databases
 - Gene-, system- or disease-specific databases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases
- Organelle databases
- Plant databases
- Immunological databases
- Cell biology

Compilation Paper
Category List
Alphabetical List
Category/Paper List
Search Summary Papers

<http://www.oxfordjournals.org/nar/database/c/>

Compilation Paper
Category List
Alphabetical List
Category/Paper List
Search Summary Papers

Oxford University Press is not responsible for the content of external internet sites

Major Molecular Databases

The most popular bioinformatics databases focus on:

- Biomolecular sequence (e.g. [GenBank](#), [UniProt](#))
- Biomolecular structure (e.g. [PDB](#))
- Vertebrate genomes (e.g. [Ensemble](#))
- Small molecules (e.g. [PubChem](#))
- Biomedical literature (e.g. [PubMed](#))

There are also many popular “*boutique*” databases for:

- Classifying protein families, domains and motifs (e.g. [PFAM](#), [PROSITE](#))
- Specific organisms (e.g. [WormBase](#), [FlyBase](#))
- Specific proteins of biomedical importance (e.g. [KinaseDB](#), [GPCRDB](#))
- Specific diseases, mutations (e.g. [OMIM](#), [HGMD](#))
- Specific fields or methods of study (e.g. [GOA](#), [IEDB](#))

Major Molecular Databases

The most popular bioinformatics databases focus on:

- Biomolecular sequence (e.g. [GenBank](#), [UniProt](#))
- Biomolecular structure (e.g. [PDB](#))
- Vertebrate genomes (e.g. [Ensemble](#))
- Small molecules (e.g. [PubChem](#))
- Biomedical literature (e.g. [PubMed](#))

There are also many "niche" databases for:

- Classifying protein families, domains and motifs (e.g. [PFAM](#), [PROSITE](#))
- Specific organisms (e.g. [WormBase](#), [FlyBase](#))
- Specific proteins of biomedical importance (e.g. [KinaseDB](#), [GPCRDB](#))
- Specific diseases, mutations (e.g. [OMIM](#), [HGMD](#))
- Specific fields or methods of study (e.g. [GOA](#), [IEDB](#))

See Online: ["Handout Major Databases.pdf"](#)

Primary, secondary & composite databases

Bioinformatics databases can be usefully classified into *primary*, *secondary* and *composite* according to their data source.

- **Primary databases** (or archival databases) consist of data derived experimentally.
 - ▶ **GenBank**: NCBI's primary nucleotide sequence database.
 - ▶ **PDB**: Protein X-ray crystal and NMR structures.
- **Secondary databases** (or derived databases) contain information derived from a primary database.
 - **RefSeq**: non redundant set of curated reference sequences primarily from GenBank
 - **PFAM**: protein sequence families primarily from UniProt and PDB
- **Composite databases** (or metadatabases) join a variety of different primary and secondary database sources.
 - **OMIM**: catalog of human genes, genetic disorders and related literature
 - **GENE**: molecular data and literature related to genes with extensive links to other databases.

GENBANK & REFSEQ:
NCBI'S NUCLEOTIDE SEQUENCE
DATABASES

What is GenBank?

- GenBank is NCBI's primary **nucleotide only** sequence database
 - ▶ Archival in nature - reflects the state of knowledge at time of submission
 - ▶ Subjective - reflects the submitter point of view
 - ▶ Redundant - can have many copies of the same nucleotide sequence
- GenBank is actually three collaborating international databases from the US, Japan and Europe
 - ▶ GenBank (US)
 - ▶ DNA Database of Japan (DDBJ)
 - ▶ European Nucleotide Archive (ENA)

GenBank sequence record

The screenshot shows the NCBI GenBank sequence record for Homo sapiens kinesin family member 5A (KIF5A), mRNA. The accession number NM_004984 is highlighted in red. The record includes fields such as LOCUS, DEFINITION, ACCESSION, VERSION, KEYWORDS, SOURCE, ORGANISM, REFERENCE, and JOURNAL. The ACCESSION field is highlighted in red.

GenBank flat file format has defined fields including unique identifiers such as the **ACCESSION** number.

This same general format is used for other sequence database records too.

Side node: Database accession numbers

Database **accession numbers** are strings of letters and numbers used as **identifying labels** for sequences and other data within databases

▶ Examples (all for retinol-binding protein, RBP4):

X02775 NT_030059	GenBank genomic DNA sequence Genomic contig	DNA
N91759.1 NM_006744	An expressed sequence tag (1 of 170) RefSeq DNA sequence (from a transcript)	RNA
NP_007635 AAC02945 Q28369 1KT7	RefSeq protein GenBank protein UniProtKB/SwissProt protein Protein Data Bank structure record	Protein
PMID: 12205585	PubMed IDs identify articles at NCBI/NIH	Literature

GenBank sequence record

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

www.ncbi.nlm.nih.gov/nuccore/NM_004984.2

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide (KIF5A) AND "Homo sapiens" Search Limits Advanced Help

Display Settings: GenBank Send:

Homo sapiens kinesin family member 5A (KIF5A), mRNA

NCBI Reference Sequence: NM_004984.2

[FASTA](#) [Graphics](#)

Go to:

LOCUS NM_004984 3897 bp mRNA linear PRI 10-JAN-2014

DEFINITION Homo sapiens kinesin family member 5A (KIF5A), mRNA.

ACCESSION NM_004984

VERSION NM_004984.2 GI:45446748

KEYWORDS RefSeq.

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 3897)

AUTHORS Kawaguchi, K.

TITLE Role of kinesin-1 in the pathogenesis of SPG10, a rare form of hereditary spastic paraplegia

JOURNAL Neuroscientist 19 (4), 336-344 (2013)

PUBMED [22785106](#)

REMARK GeneRIF: A review of the mechanism of pathogenesis involved in spastic paraplegia type 10 when KIF5A is inactivated by mutations. Review article

REFERENCE 2 (bases 1 to 3897)

AUTHORS Prots, I., Veber, V., Brey, S., Campioni, S., Buder, K., Riek, R., Bohm, K.J. and Winner, B.

TITLE alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay

JOURNAL J. Biol. Chem. 288 (30), 21742-21754 (2013)

PUBMED [23744071](#)

Change region shown

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Articles about the KIF5A gene

[alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay \[J Biol Chem. 2013\]](#)

[Molecular motor KIF5A is essential for GABA\(A\) receptor transport, a \[Neuron. 2012\]](#)

[Systems-wide analysis of ubiquitylation dynamics reveals a key role \[Nat Cell Biol. 2012\]](#)

See all..

Pathways for the KIF5A gene

[Peptide hormone metabolism](#)

[MHC class II antigen presentation](#)

GenBank sequence record

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

www.ncbi.nlm.nih.gov/nuccore/NM_004984.2

NCBI Resources How To Sign in to NCBI

Nucleotide (KIF5A) AND "Homo sapiens" Search

Display Settings: GenBank Send: Change region shown

Homo sapiens kinesin family member 5A (KIF5A), mRNA

NCBI Reference Sequence: NM_004984.2

FASTA Graphics ← Can set different display formats here

Go to:

LOCUS NM_004984 3897 bp mRNA linear PRI 10-JAN-2014

DEFINITION Homo sapiens kinesin family member 5A (KIF5A), mRNA.

ACCESSION NM_004984

VERSION NM_004984.2 GI:45446748

KEYWORDS RefSeq.

SCURCE Homo sapiens (human)

ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 3897)

AUTHORS Kawaguchi, K.

TITLE Role of kinesin-1 in the pathogenesis of SPG10, a rare form of hereditary spastic paraplegia

JOURNAL Neuroscientist 19 (4), 336-344 (2013)

PUBMED [22785106](#)

REMARK GeneRIF: A review of the mechanism of pathogenesis involved in spastic paraplegia type 10 when KIF5A is inactivated by mutations. Review article

REFERENCE 2 (bases 1 to 3897)

AUTHORS Prots, I., Veber, V., Brey, S., Campioni, S., Buder, K., Riek, R., Bohm, K.J. and Winner, B.

TITLE alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay

JOURNAL J. Biol. Chem. 288 (30), 21742-21754 (2013)

PUBMED [23744071](#)

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

Articles about the KIF5A gene

[alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay \[J Biol Chem. 2013\]](#)

[Molecular motor KIF5A is essential for GABA\(A\) receptor transport, a \[Neuron. 2012\]](#)

[Systems-wide analysis of ubiquitylation dynamics reveals a key role \[Nat Cell Biol. 2012\]](#)

See all..

Pathways for the KIF5A gene

Peptide hormone metabolism

MHC class II antigen presentation

FASTA sequence record

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

www.ncbi.nlm.nih.gov/nucleotide/45446748?report=fasta

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Search Limits Advanced Help

Display Settings: FASTA Send: Change region shown Customize view

Homo sapiens kinesin family member 5A (KIF5A), mRNA

NCBI Reference Sequence: NM_004984.2

[GenBank](#) [Graphics](#)

```
>gi|45446748|ref|NM_004984.2| Homo sapiens kinesin family member 5A (KIF5A), mRNA
ACGCCAGGTCGCCCCATCCCGCTGCCGACAGAGAGACAGCGCGCCCGGCCCTGCTCCCCAGGCTT
CGCCCGGGGGCCCTCAACTCTGTCCAGAGACTGAGCACTGTCCCTCCGCTCGGCCCTCTGCTGACAGC
CCTCTCCTCTGGAGCACACACCACCCCTGCAGCCCAAGAAAGAGTCCAGCCCCACGCCGGCTACCACCAT
GGCGGAGACCAACAGAAATGTAGCATCAAGGTGCTCTGCCGATTCGGCCCTGAACCAGGCTGAGATT
CTGCCGGGGAGACAAGTTCATCCCATTTTCCAAAGGGGACGACAGCGCTGTTATTGGGGGGGAGCCATATG
TTTTTGACCCTGTATTCCCCCAACACGACTCAAGAGCAAGTTTATCATGCATGTGCCATGCAGATTGT
CAAGATCTCCTTCTGCTTACAATGCCACCATTTTTCTTATGGACAGACATCCTCAGCGAAACACAT
ACCATCGAGGGGAAAGCTGCACGACCCTCAGCTGNTGGAAATCATTCCTCGAATTGCCCGAGACATCTCA
ACCACATCTACTCCATGGATGAGAACCTTGAGTTCACATCAAGGTTTCTTACTTTGAAATTTACCTGGA
CAAAATTCGTGACCTCTGGATGTGCCAAGACAAATCTGTCCGTGCACGAGGACAAGRACCGGGTGCCA
TTTGTCAAGGGTTGTACTGAAACCTTTGTGTCCAGCCCGAGGAGATTCTGGATGTGATGAAAGGGA
AATCAAACTCCTCATGTGCTCTCACCAACATGAATCAACACAGCTCTCCGAGCCACAGCATCTTCTCAT
CAACATCAAGCAGGAGAACATGGAAACGGAGCAGAAGCTCAGTGGSAAGCTGTATCTGCTGAGCTGGCA
GGGAGTGAGAAGGTCAGCAAGACTGGAGCAGAGGGAGCCGTGCTGAGCAGGCAAGAAATATCAACAAGT
CACTGTGAGCTCTGGCAATGTGATCTCCGCACCTGGCTGAGGGCACTAAAAGCTATGTTCCATATCGTGA
CAGCAAAATGACAAAGGATTCTCCAGGACTCTCTCGGGGAAAACCTGCCGGACGACTATGTTTATCTGTTGC
TCACCATCCAGTTATAATGATGCACAGACCAAGTCCACCCCTGATGTTTGGGCAGCGGGCAAGACCATTA
AGAACACTGCCCTCAGTAAATTTGGAGTTGACTGCTGACAGCTGGAAGAAATAATGAGAAGGACAGGA
GAASACAAAGGGCCAGAAAGGAGACGATTTGCCAAGCTGAGAGGCTGAGCTGAGCCGGTGGCCCAATGGAGAG
AATGTGCTGAGACAGAGCGCTGGCTGGGGAGGAGGCAGCCCTGGGAGCCGAGCTCTGTGAGGAGRCCC
CTGTGAATGACAACTCATCCATCGTGGTGCCTACGCGCCCGAGGAGCGGCAGAAATACGAGGAGGAGAT
CCGCCCTCTCTATAAGCACCTTCACCAAAAGGATGATGAAATCAACCAACAAAGCCAACTCATACAAAG
CTCAAGCAGCAAAATGCTGCAACCAAGAGACTGCTGCTGTCCACCCGAGGACAAACGAGAAGCTCCAGC
GGGAGCTGAGCCACCTGCAATCAGAGAACGATGCCGCTAAGGATGAGGTGAAGSAAGTCTGACAGGCCCT
GGAGGAGCTGGCTGTGAACTATGACCAGAACTCCAGGAGGTGGAGGAGAGAGCCAGCAGAACCAGCTT
CTGTTGATGAGCTGTCTCAGAAAGGTGGCCACCATGCTGTCCCTGAGTCTGAAATGACAGCAGCTACAGG
AGGTCAGTCCACACACCGCAAAACCAATTTGCTGAGCTGCTGAACGCGCTGATGAAGGATCTGAGCGGCTT
```

FASTA sequence files consist of records where each record begins with a “>” and header information on that same line. Each subsequent line of the record is sequence information.

This format is commonly used by sequence analysis programs.

Pathways for the KIF5A gene
Peptide hormone metabolism
MHC class II antigen presentation

GenBank 'graphics' sequence record

The screenshot displays the NCBI GenBank 'graphics' sequence record for the Homo sapiens kinesin family member 5A (KIF5A), mRNA (NM_004984.2). The browser address bar shows the URL: www.ncbi.nlm.nih.gov/nuccore/45446748?report=graph. The page title is "Homo sapiens kinesin family member 5A (KIF5A), mRNA" and the NCBI Reference Sequence is NM_004984.2. The main content area shows a graphical representation of the mRNA sequence, including the exon-intron structure, the protein-coding region, and various annotations. The sequence is displayed in a green bar, and the exon-intron structure is shown as a series of black boxes with arrows indicating the direction of transcription. The protein-coding region is shown as a red bar. The annotations include the KIF5A gene, the KIF5A protein, and various domains such as the ATP binding site, microtubule-binding site, and microtubule interacting site. The right sidebar contains several sections: "Analyze this sequence" (Run BLAST, Pick Primers, Highlight Sequence Features), "Articles about the KIF5A gene" (c-Synuclein oligomers impair neuronal microtubule-kinesin interp [J Biol Chem. 2013], Molecular motor KIF5A is essential for GABA(A) receptor transport, a [Neuron. 2012], Systems-wide analysis of ubiquitylation dynamics reveals a key r [Nat Cell Biol. 2012]), "Pathways for the KIF5A gene" (Peptide hormone metabolism, MHC class II antigen presentation, Dopaminergic synapse), and "Reference sequence information" (RefSeq alternative splicing, See the other reference mRNA sequence splice variant for the KIF5A gene).

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Search Limits Advanced Help

Display Settings: Graphics Send:

Homo sapiens kinesin family member 5A (KIF5A), mRNA

NCBI Reference Sequence: NM_004984.2

[GenBank](#) [FASTA](#) [Link To This Page](#) [Feedback](#)

Genes - Exon

Genes

KIF5A

NP_004975.2

KIF5A

KIF5A_KIF5

ATP binding site [c...]

acetyla... Microtubule-binding

microtubule interact...

STS Markers

D12S1839

Analyze this sequence

- Run BLAST
- Pick Primers
- Highlight Sequence Features

Articles about the KIF5A gene

- c-Synuclein oligomers impair neuronal microtubule-kinesin interp [J Biol Chem. 2013]
- Molecular motor KIF5A is essential for GABA(A) receptor transport, a [Neuron. 2012]
- Systems-wide analysis of ubiquitylation dynamics reveals a key r [Nat Cell Biol. 2012]

See all..

Pathways for the KIF5A gene

- Peptide hormone metabolism
- MHC class II antigen presentation
- Dopaminergic synapse

See all..

Reference sequence information

- RefSeq alternative splicing
- See the other reference mRNA sequence splice variant for the KIF5A gene

GenBank sequence record, cont.

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

www.ncbi.nlm.nih.gov/nuccore/NM_004984.2

NCBI Resources How To Sign in to NCBI

Nucleotide (KIF5A) AND "Homo sapiens" Search

Display Settings: GenBank Send:

Homo sapiens kinesin family member 5A (KIF5A), mRNA

NCBI Reference Sequence: NM_004984.2

[FASTA](#) [Graphics](#)

Go to:

LOCUS	NM_004984	3897 bp	mRNA	linear	PRI 10-JAN-2014
DEFINITION	Homo sapiens kinesin family member 5A (KIF5A), mRNA.				
ACCESSION	NM_004984				
VERSION	NM_004984.2 GI:45446748				
KEYWORDS	RefSeq.				
SCURCE	Homo sapiens (human)				
ORGANISM	<u>Homo sapiens</u> Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.				
REFERENCE	1 (bases 1 to 3897)				
AUTHORS	Kawaguchi, K.				
TITLE	Role of kinesin-1 in the pathogenesis of SPG10, a rare form of hereditary spastic paraplegia				
JOURNAL	Neuroscientist 19 (4), 336-344 (2013)				
PUBMED	22785106				
REMARK	GeneRIF: A review of the mechanism of pathogenesis involved in spastic paraplegia type 10 when KIF5A is inactivated by mutations. Review article				
REFERENCE	2 (bases 1 to 3897)				
AUTHORS	Prots, I., Veber, V., Brey, S., Campioni, S., Buder, K., Riek, R., Bohm, K.J. and Winner, B.				
TITLE	alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay				
JOURNAL	J. Biol. Chem. 288 (30), 21742-21754 (2013)				
PUBMED	23744071				

Change region shown

Customize view

Analyze this sequence

- Run BLAST
- Pick Primers
- Highlight Sequence Features
- Find in this Sequence

Articles about the KIF5A gene

- [alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay \[J Biol Chem. 2013\]](#)
- [Molecular motor KIF5A is essential for GABA\(A\) receptor transport, a \[Neuron. 2012\]](#)
- [Systems-wide analysis of ubiquitylation dynamics reveals a key role for KIF5A \[Nat Cell Biol. 2012\]](#)

See all..

Pathways for the KIF5A gene

- Peptide hormone metabolism
- MHC class II antigen presentation

GenBank sequence record, cont.

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

www.ncbi.nlm.nih.gov/nuccore/45446748?report=genbank&to=3897#feature_45446748

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

FEATURES	Location/Qualifiers	OMIM
source	1..3897 /organism="Homo sapiens" /mol_type="mRNA" /db_xref="taxon:9606" /chromosome="12" /map="12q13.13"	Probe Protein PubMed PubMed (RefSeq)
gene	1..3897 /gene="KIF5A" /gene_synonym="D12S1689; MY050; NKHC; SP310" /note="kinesin family member 5A" /db_xref="GeneID:3796" /db_xref="HGNC:6323" /db_xref="HPRD:09108" /db_xref="MIM:602821"	
exon	1..337 /gene="KIF5A" /gene_synonym="D12S1689; MY050; NKHC; SP310" /inference="alignment:Splice:1.39.8"	
misc_feature	134..136 /gene="KIF5A" /gene_synonym="D12S1689; MY050; NKHC; SP310" /note="upstream in-frame stop codon"	
CDS	209..3307 /gene="KIF5A" /gene_synonym="D12S1689; MY050; NKHC; SP310" /note="kinesin, heavy chain, neuron-specific; KIF5A variant protein; neuronal kinesin heavy chain; kinesin heavy chain neuron-specific 1" /codon_start=1 /product="kinesin heavy chain isoform 5A" /protein_id="NP_004975.2" /db_xref="GI:45446749" /db_xref="CCDS:CCDS8945.1" /db_xref="GeneID:3796" /db_xref="HGNC:6323" /db_xref="HPRD:09108" /db_xref="MIM:602821" /translation="MAETNNECSIKVLCRFRPLNQAEILRGDKFIFIFQGDDSVVIGG KPYVFDRVFPNNTTQEQVYHACAMQIVKDVLAGYNGTIFAYGQTSSGKTHMECKLHD PQLMGIIPRIARDIPNHIYSMDENLEPHIKVSYFEIYLDKIRDLLDVTKTNLSVHEDK NRVFPVKGCTERFVSSPEEILDVIDEGKSNRHVAVTNMNEHSSRSHSIFLINIKQENM ETEOKLSGKLYLVDLAGSEKVSKTGAEAVLDEAKNINKSLSALGNVISALAECTKSY VPYRDSKMTRILQDSLGGNCRTTMFICCPSSYNDAETKSTLMFGQRAKTIKNTASVN	

The **FEATURES** section contains annotations including a conceptual translation of the nucleotide sequence.

Recent activity [Turn Off](#) [Clear](#)

- Homo sapiens kinesin family member 5A (KIF5A), mRNA Nucleotide
- (kinesin) AND "Homo sapiens"[porgn] (1351) Nucleotide
- kinesin (37064) Nucleotide

[See more..](#)

GenBank sequence record, cont.

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

www.ncbi.nlm.nih.gov/nuccore/45446748?report=genbank&to=3897#sequence_45446748

Homo sapiens kinesin family member 5A (KIF5A), mRNA - Nucleotide - NCBI

/gene_synonym="D12S1889; MY050; NKHC; SPG10"
/standard_name="D12S1889"
/db_xref="UniSTS:48006"

ORIGIN

```
1 aagcccagggt ogcccgcgato ccgctgcoogo aggagagaga oagcgcgooo oggcccotgot
61 ccccaggcctt ngcccggggcg cccctnaactc tgtcccncaga gactgagnac ctgtcctcng
121 cctcggcctc tgcctgagagc cctctcctct ggagcacaca ccaccctgc agcccgaaga
181 gagtcccagc cccacgcggg ctaccaccat ggccgagacc aacaacgaat gtagcatcaa
241 ggtgctctgc cgattccggc ccctgaacce ggcctgagatt ctgcggggag ecaagttcat
301 oooaattht oaaaggggaog acagogtctg tttggggggg aagccatctg tttttgaocg
361 tgtattccc ccaaacacga ctcaagagca agtttatcat gcctgtgcca tgcagattgt
421 caaagatgtc cttgctggct acaatggcac ctttttggct tatggacaga catcctcagg
481 gaaaacacat accatggagg gaaagctgca cgacctcag ctgatgggaa tcattcctcg
541 aattgcccga gacatcttca accacatctc ctccatggat gagaacctg agttccacat
601 caagtttct taatttgaaa ttaacctgga caaaattogt gacctctgg atgtgaccaa
661 gacaaatctg tccgtgcaog aggacaagaa ccgggtgcca tttgtcaagg gttgtactga
721 acgctttgtg tccagcccgg aggagattct ggatgtgatt gatgaagga aatcaaatcg
781 tcatgtggct gtcaccaaca tgaatgaaca cagctctogg agccacagca tcttctcat
841 caacatcaag caggagaaca tggaaacgga gcagaagctc agtgggaagc tgtatctggc
901 ggacctgga gggagtgaga aggtcagcaa gactggagaa gaggagagcg tgcctggaog
961 ggcaaaagt atcaacaagt cactgtcagc tctgggcaat gtgatctcng cactggctga
1021 gggcactaaa agctatgttc catatctgta cagcaaatg acaaggattc tccaggactc
1081 tctcggggga aactgcccga cgaactatgt catctgttgc tcaccatca gttataatga
1141 tgcagagacc aagtccacc tcatgtttgg gcagcgggca aagaccatta egaacactgc
1201 otoagtaat ttggagttga ctgctgagca gtggaagaag aaatatgaga aggagaagga
1261 gaagacaaag gccacagaag agacgattgc gaagctggag gctgagctga gccgggtggc
1321 caatggagag aatgtgcctg agacagagcg cctggctggg gaggaggcag ccctgggagc
1381 cgagctctgt gaggagacc ctgtgaatga caactcatcc atcgtggctg gcctcggcc
1441 cgaggagcgg cagaatacag aggaggagat ccgcccctc tataagcagc ttgacgacaa
1501 ggatgatgaa atcaaccaac aaagocaaat catagagaag otaagcagc aatgotgga
1561 ccaggagag ctgctggctt ccaccogag agacaacgag aaggtccagc gggagctgag
1621 ccacctgcaa tcagagaacg atgcccctaa ggatgagctg aaggaagtgc tgcaggcct
1681 ggaggagctg gctgtgaaat atgaccagaa gtcccaggag gtggaggaga agagccagca
1741 gaaccagctt ctgggtggtg agctgtctca gaaggtggcc accatgctgt ccctggagtc
1801 tgagttgag oggctaacag aggtcagtg acocagoga aaaogaattg ctgaggtgot
1861 gaaagggctg atgaaggatc tgagcagatt cagtgtcatt gtgggcaacg gggagattaa
1921 gctgccagtg gagatcagtg gggccatcga ggaggagttc actgtggccc gactctacat
1981 cagcaaatc aaatcagaag tcaagtctgt ggtcaagcgg tgcggcagc tggagaacct
2041 ccaggtggag tctcccgcga agatggaagt gaccggcggg gagctctcat cctgcccagc
2101 cctcactctc cagcatgag ccaagatccg ctgcttaac gaatacatgc agagctgga
2161 gctaaagaag cggcacctgg aagagtccta tgactccttg agcagatgagc tggccaagct
2221 ccaggcccag gaaactgtgc atgaagtggc cctgaaggac aaggagcctg acactcagga
2281 tgcagatgaa gtgaagaagg ctctggagct gcagatggag agtcccggg aggcccacaa
2341 ccggcagctg gcccggtcc gggacgagat caacgagaag cagaagacca ttgatgagct
```

The actual sequence entry starts after the word **ORIGIN**

RefSeq: NCBI's Derivative Sequence Database

- RefSeq entries are hand curated best representation of a transcript or protein (in their judgement)
- Non-redundant for a given species although alternate transcript forms will be included if there is good evidence

- Experimentally verified transcripts and proteins
accession numbers begin with “NM_” or “NP_”
- Model transcripts and proteins based on bioinformatics predictions with little experimental support
accession numbers begin with “XM_” or “XP_”
- RefSeq also contains contigs and chromosome records

UNIPROT:

THE PREMIER PROTEIN SEQUENCE
DATABASE

UniProt: Protein sequence database

UniProt is a comprehensive, high-quality resource of protein sequence and functional information

- UniProt comprises four databases:

1. **UniProtKB** (Knowledgebase)

Containing **Swiss-Prot** and **TrEMBL** components

(these correspond to hand curated and automatically annotated entries respectively)

2. **UniRef** (Reference Clusters)

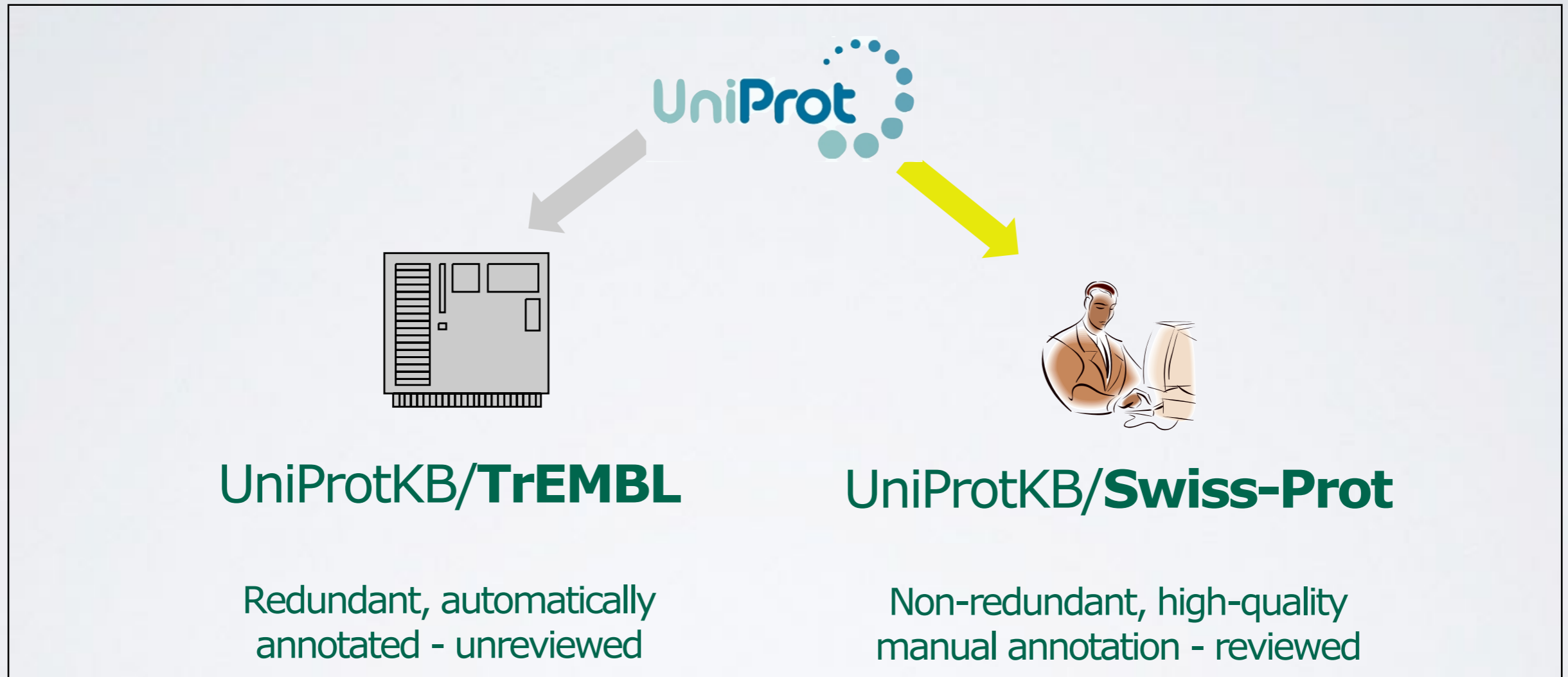
Filtered version of UniProtKB at various levels of sequence identity

e.g. UniRef90 contains sequences with a maximum of 90% sequence identity to each other

3. **UniParc** (Archive) with database cross-references to source.

4. **UniMES** (Metagenomic and Environmental Sequences)

The two sides of UniProtKB



★ Unreviewed, UniProtKB/TrEMBL **Q9N0H9** (Q9N0H9_EQUAS)

★ Reviewed, UniProtKB/Swiss-Prot **P38398** (BRCA1_HUMAN)

Indicators of which part of UniProt an entry belongs to include the color of the stars and the ID

The main information added to a UniProt/Swiss-Prot entry

```

10      20      30      40      50      60
MVGEMETKEK PKPTPDYLMQ LMNDKMLSS LPNFCGIFNH LERLLDEEIS RVRKDMYNDT

70      80      90     100     110     120
LNGSTEKRSÄ ELPDAVGPIV QLQEKLYVPV KEYPDFNFVG RILGPRGLTÄ KQLEAETGCK

130     140     150     160     170     180
IMVVRGKSMR D...ONRG KPNWEHLNED LHVLLTVEDA ONPÄEIKLKR AVEEVKLLLV

190     200     210     220     230     240
PAAEGEDSLK KMQLMELAIL NG...RDANIK SPALAFSLAA TAQAAPRIIT GPAPVLPAA

250     260     270     280     290     300
LRTPTPAGPT IMPLIRQIQT AVMPNGTPHP TAAIVPPGPE AGLIYTPYEV PYTLAPATSI

310     320     330     340
LEYPLEPSGV LGAVATKVRÄ HDMRVHPYQR IVTADRAATG N
    
```

Sequence

- [1] "The quaking gene product necessary in embryogenesis and myelination combines features of RNA binding and signal transduction proteins."
Ebersole T.A., Chen Q., Justice M.J., Artzt K.
Nat. Genet. 12:260-265(1996) [PubMed: 8589716] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 3), INVOLVEMENT IN QKV, TISSUE SPECIFICITY, MUTAGENESIS C
- [2] "Genomic org..."
Kondo T., Furuta T., Mitsuhashi K., Ebersole T.A., Shichiri M., Wu J., Artzt K., Yamamura K., Abe K.
Mamm. Genome 10:662-669(1999) [PubMed: 10384037] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA / MRNA] (ISOFORMS 2; 3; 4 AND 7), ALTERNATIVE SPLICING (ISOFORM 1).
Strain: 129/J.

References

General annotation (Comments)	Info
Function	RNA-binding protein that plays a central role in myelination. Also required for visceral endoderm function and blood vessel development. Binds to the 5'-NACUAAAY-N(1,20)-UAAAY-3' RNA core sequence. Acts by regulating pre-mRNA splicing, mRNA export, mRNA stability and protein translation, as well as cellular processes including apoptosis, cell cycle, glial cell fate and development. Required to protect and promote stability of mRNAs such as MBP and CDKN1B to promote oligodendrocyte differentiation. Participates in mRNA transport by regulating the nuclear export of MBP mRNA. Isoform 1 is involved in regulation of mRNA splicing of MAG pre-mRNA by acting as a negative regulator of MAG exon 12 alternative splicing. Isoform 3 can induce apoptosis, while heterodimerization with other isoforms result in nuclear translocation of isoform 3 and suppression of apoptosis. Isoform 4 acts as a translational repressor for GLUT1. May also play a role in smooth muscle development.
Subunit structure	Homodimer. Does not require RNA to homodimerize. Able to heterodimerize with BICC1.
Subcellular location	Cytoplasm; Nucleus. Note=isoform 1 localizes predominantly in the nucleus and at lower level in cytoplasm. It shuttles between the cytoplasm and the nucleus. Isoform 3 localizes predominantly in the cytoplasm and at much lower level in nucleus. Isoform 4 localizes both in the cytoplasm and nucleus.
Tissue specificity	Highly expressed in neurogenesis cells. Expressed in oligodendrocytes and astrocytes in the central nervous system as well as Schwann cells. Expressed in the mesodermal site of developing blood islands. Expressed in brain, lung, heart and testis.
Developmental stage	Neuronal differentiation. By contrast, neural progenitors located in specific sub-domains of the vz maintain expression as they differentiate and migrate away into the emerging nervous system. These have characteristics consistent with the acquisition of a glial rather than neuronal fate (at protein level). First detected in the neuroepithelium of the head folds at E7.5. Expression is strongly present ventrally in the ventral brain and neurula of E8.5 and E9.5 and in the brain and spinal cord. Isoform 1 is expressed in early embryos, while isoform 3 and isoform 4 are expressed in late embryos.
Post-translational modification	Tyrosine phosphorylation. Phosphorylation of tyrosine residues is essential for protein-protein interactions and protein activity, affecting transport and/or stabilization of MBP mRNA. The level of Tyr phosphorylation in the developing myelin is highest in the first postnatal week (P7). During the vigorous accumulation of MBP mRNA between P1 and P-20, phosphorylation in the developing myelin drastically declined. By the end of the fourth postnatal week (P28), phosphorylation is reduced approximately 90%.
Involvement in disease	Defects in Qki are the cause of quakinglike (qkv). Qkv is a spontaneous mutation resulting in hypomyelination of the central and peripheral nervous systems. Mutant mice develop normally until postnatal day 10 when they display rapid tremors or 'quaking' that is especially pronounced in hindlimbs and experience convulsive tonic-clonic seizures as they mature. Mice with qkv specifically lack isoform 3.

Literature Annotations

Cell cycle	Regulation of cell proliferation Traceable author statement. Source: UniProtKB
DNA damage	
DNA repair	Regulation of transcription from RNA polymerase II promoter Traceable author statement. Source: Protinc
Fatty acid biosynthesis	Regulation of transcription from RNA polymerase III promoter Traceable author statement. Source: UniProtKB
Lipid synthesis	
Nucleus	Response to estrogen stimulus Inferred from direct assay. Source: UniProtKB
Polymorphism	BRCA1-BARD1 complex Inferred from direct assay. Source: UniProtKB
Disease mutation	
Repeat	
Zinc-finger	Gamma-tubulin ring complex Non-traceable author statement. Source: UniProtKB
DNA-binding	
Metal-binding	DNA binding Traceable author statement. Source: Protinc
Zinc	
Anti-oncogene	Androgen receptor binding Non-traceable author statement. Source: UniProtKB
Phosphorylation	Enzyme binding Inferred from physical interaction. Source: UniProtKB
3D-structure	

Ontologies



Isoform 1 (identifier: Q9QYS9-1)

Also known as Qk1-5;

This isoform has been chosen as the 'canonical' sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

Isoform 2 (identifier: Q9QYS9-2)

Also known as Qk1-7;

The sequence of this isoform differs from the canonical sequence as follows:

312-341 GAVATKVRRHDMRVHPYQRIVTADRAATGN → WLSQRKAKNSRTLTEPSSDLNLTNA

Isoform 3 (identifier: Q9QYS9-3)

Also known as Qk1-7;

The sequence of this isoform differs from the canonical sequence as follows:

312-341 GAVATKVRRHDMRVHPYQRIVTADRAATGN → EWIEPVMVDISAH

Sequence variants

Protein names

Protein quaking

Also known as:

Mqkl

Nomenclature

Gene names

Name: Qki

Synonyms: Qk, Qk1, Qka1

Molecule processing

Chain 1 – 341 341 Protein quaking

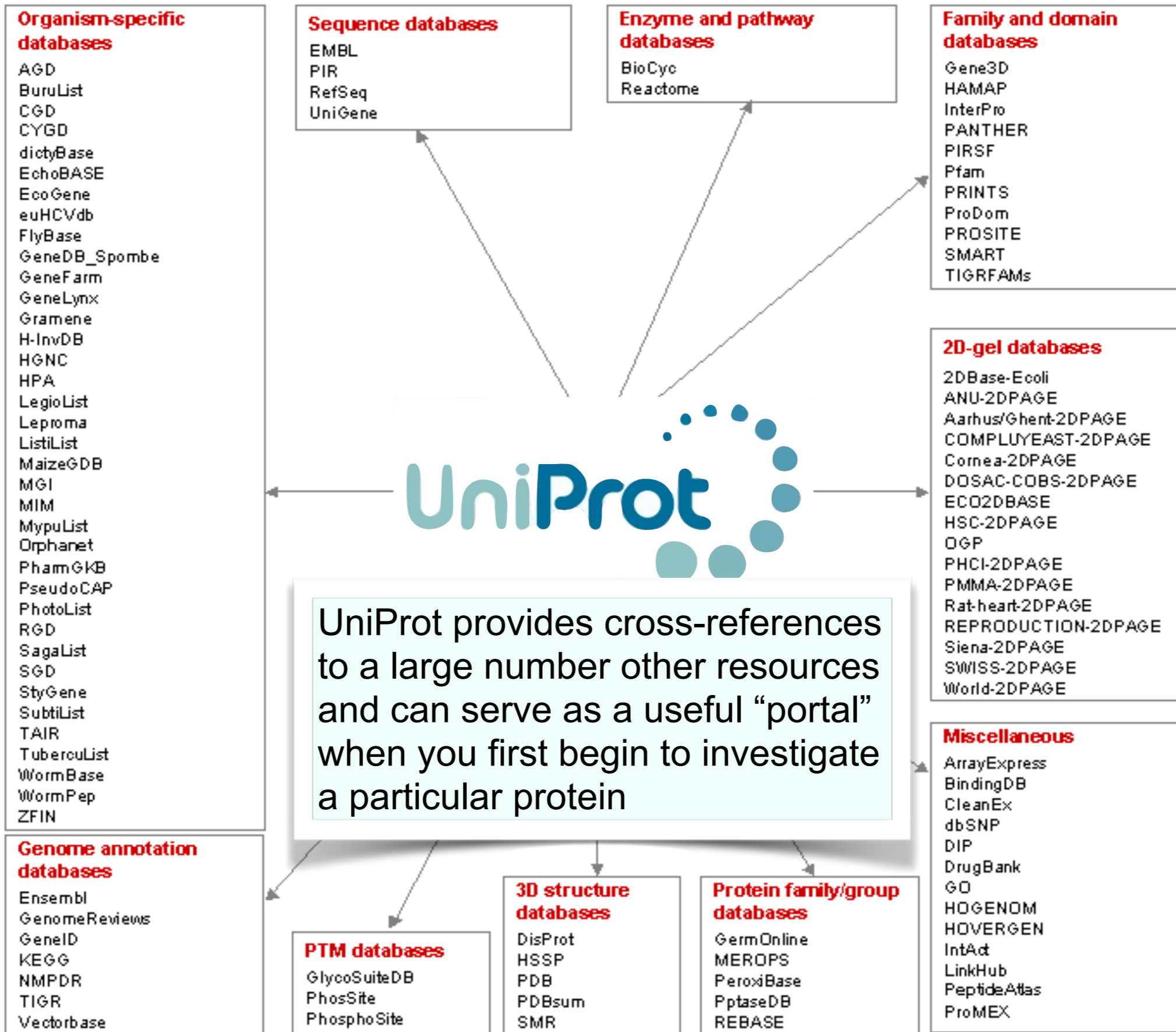
Regions

Domain 87 – 153 67 KH

Motif 276 – 300 25

Motif 324 – 330 7 Nuclear localization signal

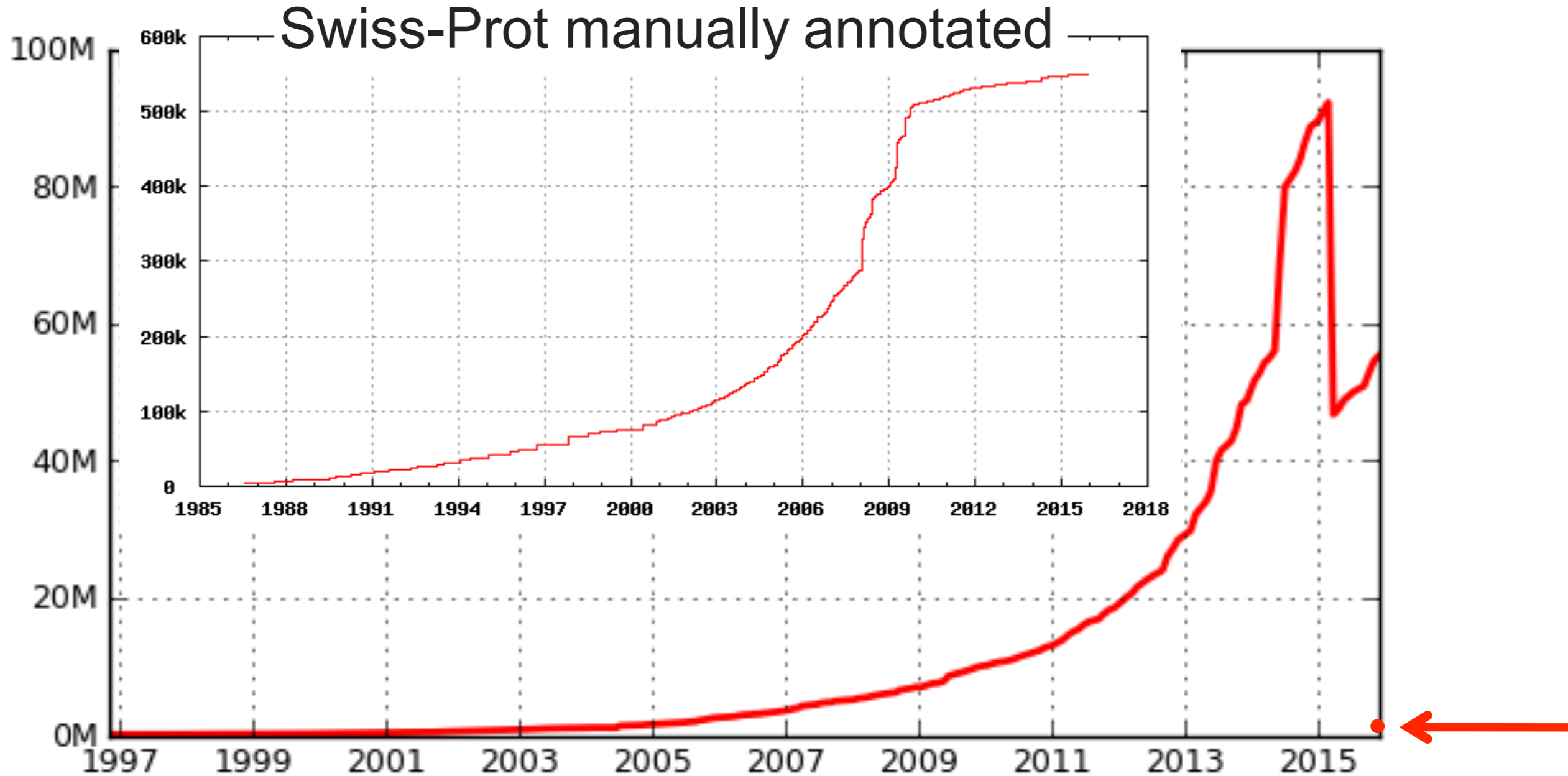
Sequence features



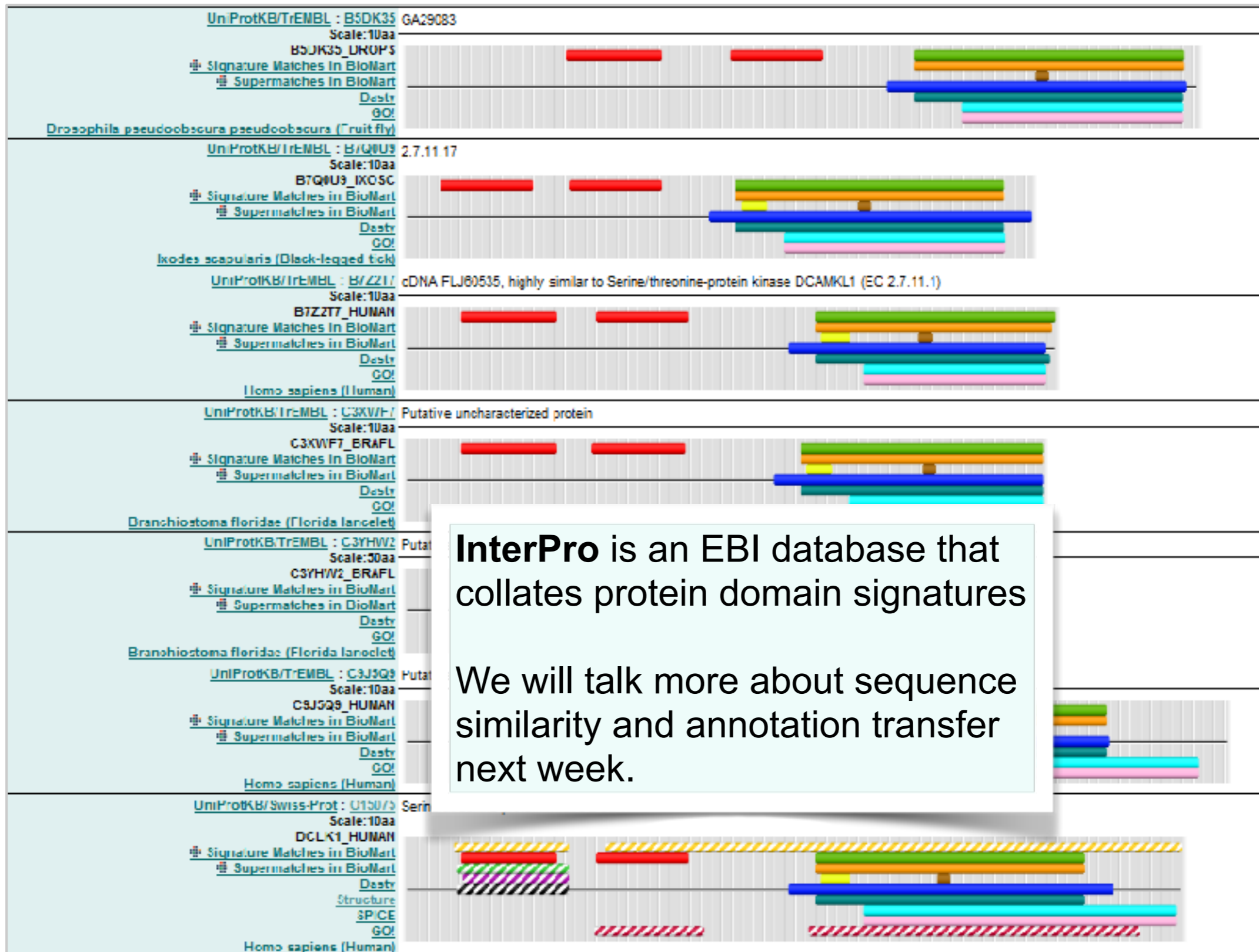
UniProt/Swiss-Prot vs UniProt/TrEMBL

- UniProtKB/Swiss-Prot is a **non-redundant** database with one entry per protein
- UniProtKB/TrEMBL is a **redundant** database with one entry per translated ENA entry (ENA is the EBI's equivalent of GenBank)
 - ▶ Therefore TrEMBL can contain multiple entries for the same protein
 - ▶ Multiple UniProtKB/TrEMBL entries for the same protein can arise due to:
 - Erroneous gene model predictions
 - Sequence errors (Frame shifts)
 - Polymorphisms
 - Alternative start sites
 - Isoforms
 - OR because the same sequence was submitted by different people

Side note: Automatic Annotation (sharing the wealth)



Same domain composition = same function = annotation transfer



DATABASE VIGNETTE

You have just come out a seminar about gastric cancer and one of your co-workers asks:

“What do you know about that ‘Kras’ gene the speaker kept taking about?”

You have some recollection about hearing of ‘Ras’ before. How would you find out more?

- Google?
- Library?
- **Bioinformatics databases at NCBI and EBI!**

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

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

The image shows a screenshot of the National Center for Biotechnology Information (NCBI) website. The browser's address bar displays www.ncbi.nlm.nih.gov/. The NCBI logo and navigation menu are visible at the top. A search bar is present with the text "All Databases" and a dropdown menu showing "ras", which is highlighted with a red rectangular box. A blue "Search" button is located to the right of the search bar. A diagonal banner with red text reads "Hands on demo (or see following slides)".

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 and the National Library of Medicine provide access to biomedical and health information and health by providing access to biomedical and health information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News](#)

Get Started

- [Data](#): Get NCBI data or software
- [Tutorials](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

Genotypes and Phenotypes

Data from Genome Wide Association studies that link genes and diseases. See study variables, protocols, and analysis.

NCBI Announcements

RefSeq release 69 available on...

The full RefSeq release 69 is not available on the FTP site with 74 records describing 52,376,489...

Search NCBI databases

[Help](#)

ras



Search

About 2,978,774 search results for "ras"

Literature

Books	1,677	books and reports
MeSH	402	ontology used for PubMed indexing
NLM Catalog	223	books, journals and more in the NLM Collections
PubMed	54,672	scientific & medical abstracts/citations
PubMed Central	96,114	full-text journal articles

Health

ClinVar	759	human variations of clinical significance
dbGaP	120	genotype/phenotype interaction studies
GTR	1,879	genetic testing registry

Genes

EST	3,985	expressed sequence tag sequences
Gene	87,165	collected information about gene loci
GEO DataSets	3,732	functional genomics studies
GEO Profiles	1,622,789	gene expression and molecular abundance profiles
HomoloGene	696	homologous gene sets for selected organisms
PopSet	2,254	sequence sets from phylogenetic and population studies
UniGene	4,770	clusters of expressed transcripts

Proteins



Gene [Save search](#) [Advanced](#) [Help](#)

[Show additional filters](#)

Display Settings: Tabular, 20 per page, Sorted by Relevance

Send to:

[Hide sidebar >>](#)

Filters: [Manage Filters](#)

▼ Top Organisms [\[Tree\]](#)

- Homo sapiens (1126)**
- Mus musculus (823)
- Rattus norvegicus (625)
- Oreochromis niloticus (533)
- Neolamprologus brichardi (507)
- All other taxa (82019)

[More...](#)

Find related data

Database:

Search details

ras[All Fields] AND alive[property]

Did you mean ras as a gene symbol?
Search Gene for [ras](#) as a symbol.

<< First < Prev Page 1 of 4282 Next > Last >>

Results: 1 to 20 of 85633

i Filters activated: Current only. [Clear all](#) to show 87165 items.

Name/Gene ID	Description	Location	Aliases
<input type="checkbox"/> ras ID: 19412	resistance to audiogenic seizures [<i>Mus musculus</i> (house mouse)]		asr
<input type="checkbox"/> ras ID: 43873	raspberry [<i>Drosophila melanogaster</i> (fruit fly)]	Chromosome X, NC_004354.4 (10744502..10749097)	Dmel_CG1799, CG11485, CG1799, DmelCG1799, EP(X)1093,

[Clear all](#)

Gene sources

- Genomic
- Mitochondria
- Organelles
- Plasmids
- Plastids

Categories

- Alternatively spliced
- Annotated genes
- Non-coding
- Protein-coding
- Pseudogene

Sequence content

- CCDS
- Ensembl
- RefSeq

Gene

Gene

(ras) AND "Homo sapiens"[porgn: __txid9606]

Search

Help

Show additional filters

Display Settings: Tabular, 20 per page, Sorted by Relevance

Send to:

Hide sidebar >>

Filters: Manage Filters

Clear all

Results: 1 to 20 of 1126 << First < Prev Page 1 of 57 Next > Last >>

Filters activated: Current only. Clear all to show 1499 items.

Gene sources

Genomic

Categories

Alternatively spliced

Annotated genes

Non-coding

Protein-coding

Pseudogene

Sequence content

CCDS

Ensembl

RefSeq

Status

clear

Current only

Chromosome locations

Name/Gene ID	Description	Location	Aliases
<input type="checkbox"/> NRAS ID: 4893	neuroblastoma RAS viral (v-ras) oncogene homolog [Homo sapiens (human)]	Chromosome 1, NC_000001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
<input type="checkbox"/> KRAS ID: 3845	Kirsten rat sarcoma viral oncogene homolog [Homo sapiens (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, KI-RAS1, KRAS2, NS, NS3, RASK2

Find related data

Database:

Select

Find items

Search details

ras[All Fields] AND "Homo sapiens"[porgn] AND alive[property]

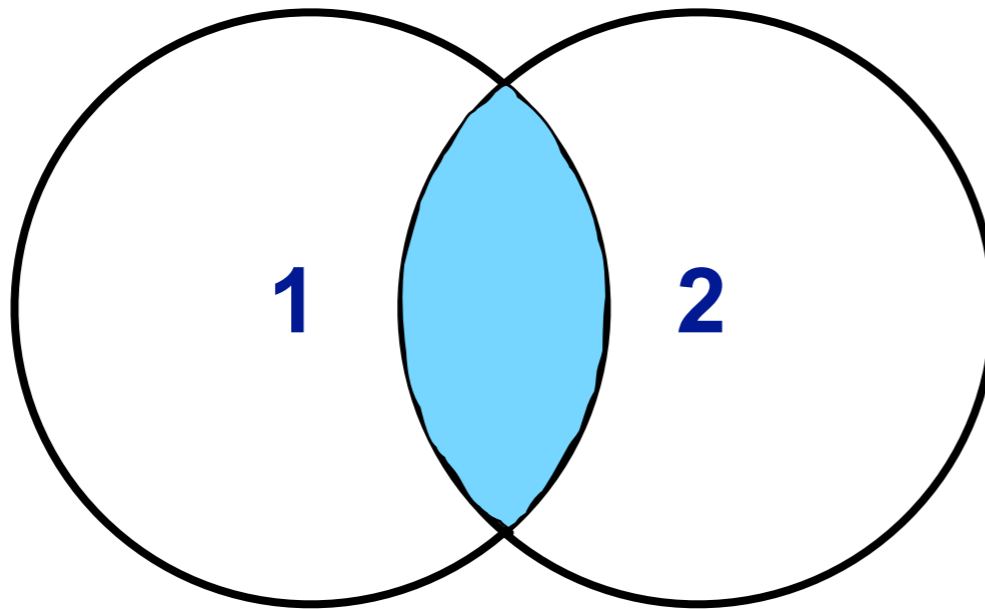
Search

See more...

Recent activity

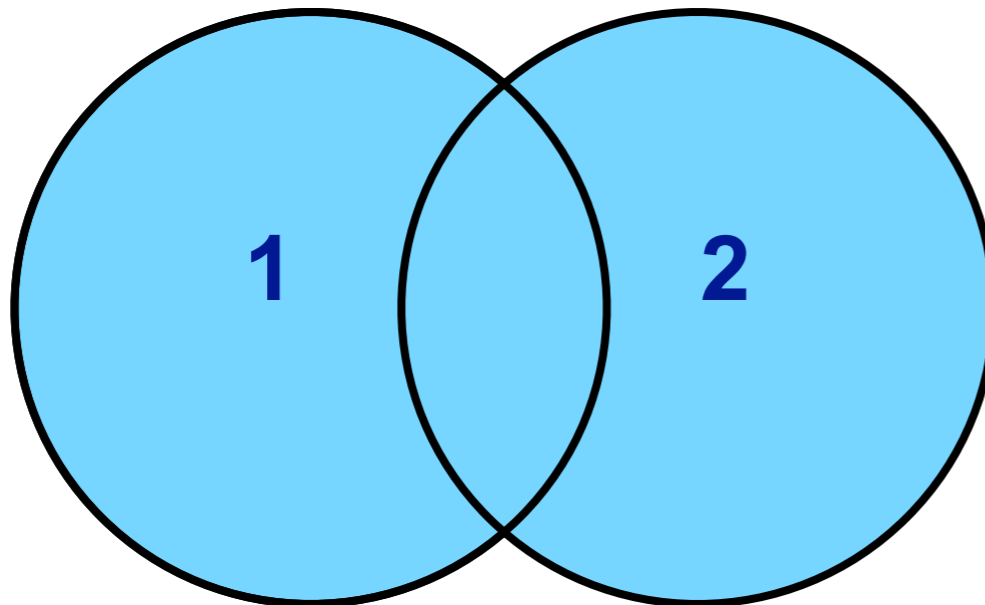
Turn Off Clear

1 AND 2



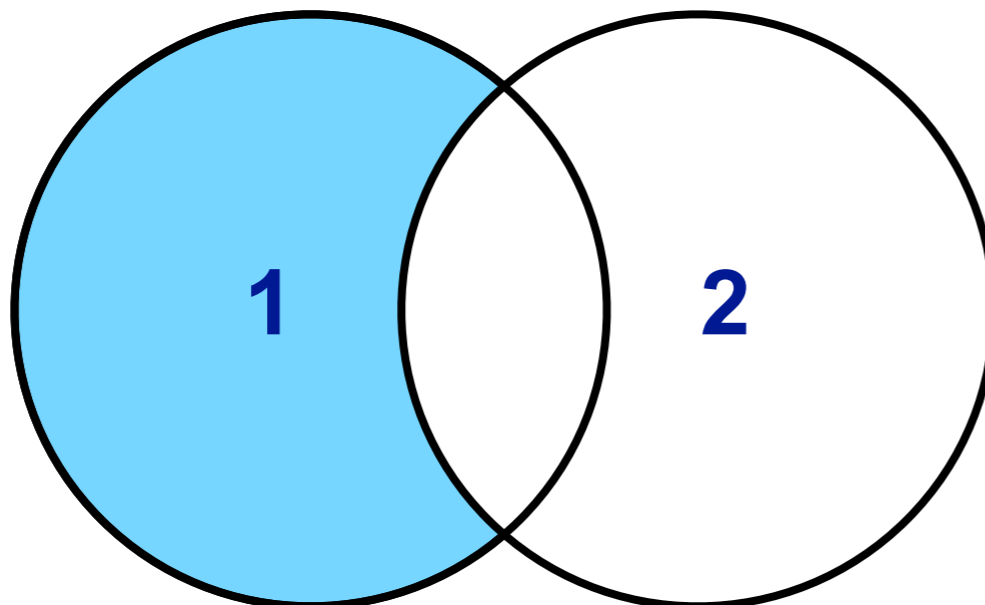
**ras AND disease
(1185 results)**

1 OR 2



**ras OR disease
(134,872 results)**

1 NOT 2



**ras NOT disease
(84,448 results)**

Gene (ras) AND "Homo sapiens"[porgn: __txid9606] Save search Advanced Help

Show additional filters

Display Settings: Tabular, 20 per page, Sorted by Relevance Send to:

Hide sidebar >>

Filters: Manage Filters

Clear all

Results: 1 to 20 of 1126 << First < Prev Page 1 of 57 Next > Last >>

Filters activated: Current only. Clear all to show 1499 items.

- Gene sources
- Genomic
- Categories
 - Alternatively spliced
 - Annotated genes
 - Non-coding
 - Protein-coding
 - Pseudogene
- Sequence content
 - CCDS
 - Ensembl
 - RefSeq
- Status
- Current only
- Chromosome locations

Name/Gene ID	Description	Location	Aliases
<input type="checkbox"/> NRAS ID: 4893	neuroblastoma RAS viral (v-ras) oncogene homolog [<i>Homo sapiens</i> (human)]	Chromosome 1, NC_000001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
<input type="checkbox"/> KRAS ID: 3845	Kirsten rat sarcoma viral oncogene homolog [<i>Homo sapiens</i> (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, KI-RAS1, KRAS2, NS, NS3, RASK2

Find related data

Database:

Search details

ras[All Fields] AND "Homo sapiens"[porgn] AND alive[property]

Recent activity

Gene [Advanced](#) [Help](#)

[Display Settings:](#) Full Report [Send to:](#)

KRAS Kirsten rat sarcoma viral oncogene homolog [*Homo sapiens* (human)]

Gene ID: 3845, updated on 4-Jan-2015

Summary

Official Symbol KRAS provided by HGNC
Official Full Name Kirsten rat sarcoma viral oncogene homolog provided by HGNC
Primary source [HGNC:HGNC:6407](#)
See related [Ensembl:ENSG00000133703](#); [HPRD:01817](#); [MIM:190070](#); [Vega:OTTHUMG00000171193](#)
Gene type protein coding
RefSeq status REVIEWED
Organism [Homo sapiens](#)
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Hominidae; Homo
Also known as NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

- Table of contents
- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Phenotypes
- Variation
- HIV-1 interactions
- Pathways from BioSystems
- Interactions
- General gene information
Markers, Related pseudogene(s), Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)



www.ncbi.nlm.nih.gov/gene/3845

NCBI Resources How To Sign in to NCBI

Gene Search Help

Display Settings Hide sidebar >>

Example Questions:
 What chromosome location and what genes are in the vicinity?

Table of contents

- Summary
- Genomic context**
- Genomic regions, transcripts, and products
- Bibliography
- Phenotypes
- Variation
- HIV-1 interactions
- Pathways from BioSystems
- Interactions
- General gene information
 - Markers, Related pseudogene(s), Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences

Summary

Official Symbol KRAS provided by HGNC

Official Full Name Kirsten rat sarcoma viral oncogene homolog provided by HGNC

Primary source [HGNC:HGNC:6407](#)

See related [Ensembl:ENSG00000133703](#); [HPRD:01817](#); [MIM:190070](#); [Vega:OTTHUMG00000171193](#)

Gene type protein coding

RefSeq status REVIEWED

Organism [Homo sapiens](#)

Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Hominidae; Homo

Also known as NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

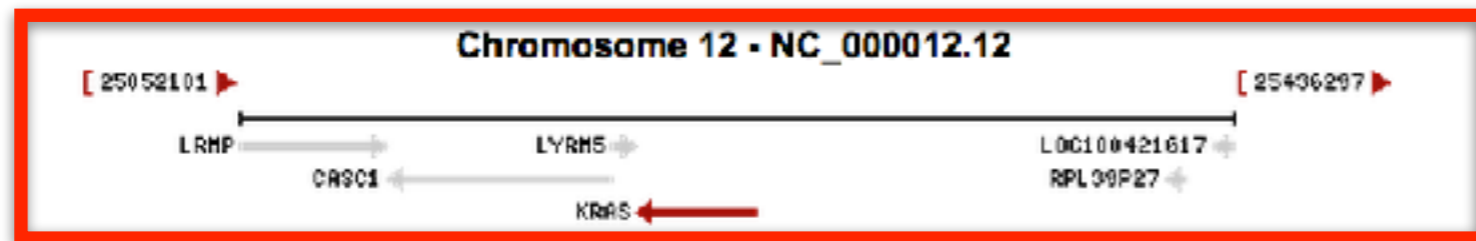
Genomic context

Location: 12p12.1

Exon count: 6

See KRAS in [Epigenomics](#), [MapViewer](#)

Annotation release	Status	Assembly	Chr	Location
106	current	GRCh38 (GCF_000001405.26)	12	NC_000012.12 (25205246..25250923, complement)
105	previous assembly	GRCh37.p13 (GCF_000001405.25)	12	NC_000012.11 (25358180..25403870, complement)



Genomic regions, transcripts, and products

Go to [reference sequence details](#)

Genomic Sequence: NC_000012.12 chromosome 12 reference GRCh38 Primary Assembly

Go to nucleotide: [Graphics](#) [FASTA](#) [GenBank](#)

- [BioAssay by Target \(Summary\)](#)
- [BioAssay, by Gene target](#)
- [BioAssays, RNAi Target, Active](#)
- [BioAssays, RNAi Target, Tested](#)
- [BioProjects](#)
- [BioSystems](#)
- [Books](#)
- [CCDS](#)
- [ClinVar](#)
- [Conserved Domains](#)
- [dbVar](#)
- [EST](#)
- [Full text in PMC](#)
- [Full text in PMC_nucleotide](#)
- [Gene neighbors](#)
- [Genome](#)
- [GEO Profiles](#)
- [GTR](#)
- [HomoloGene](#)
- [Map Viewer](#)
- [MedGen](#)
- [Nucleotide](#)

www.ncbi.nlm.nih.gov/gene/3845

NCBI Resources How To Sign in to NCBI

Gene

Search Help

Hide sidebar >>

Table of contents

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Phenotypes
- Variation
- HIV-1 interactions
- Pathways from BioSystems
- Interactions
- General gene information**
Markers, Related pseudogene(s), Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences

Example Questions:
 What 'molecular functions', 'biological processes', and 'cellular component' information is available?

Display Settings

KRAS **Ki**
(human)]

Gene ID: 3845

Summary

Official Symbol KRAS provided by HGNC

Official Full Name Kirsten rat sarcoma viral oncogene homolog provided by HGNC

Primary source [HGNC:HGNC:6407](#)

See related [Ensembl:ENSG00000133703](#); [HPRD:01817](#); [MIM:190070](#); [Vega:OTTHUMG00000171193](#)

Gene type protein coding

RefSeq status REVIEWED

Organism [Homo sapiens](#)

Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

Also known as NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-

Gene Ontology Provided by GOA

Function	Evidence Code	Pubs
GDP binding	IEA	
GMP binding	IEA	
GTP binding	IEA	
LRR domain binding	IEA	
protein binding	IPI	PubMed
protein complex binding	IDA	PubMed

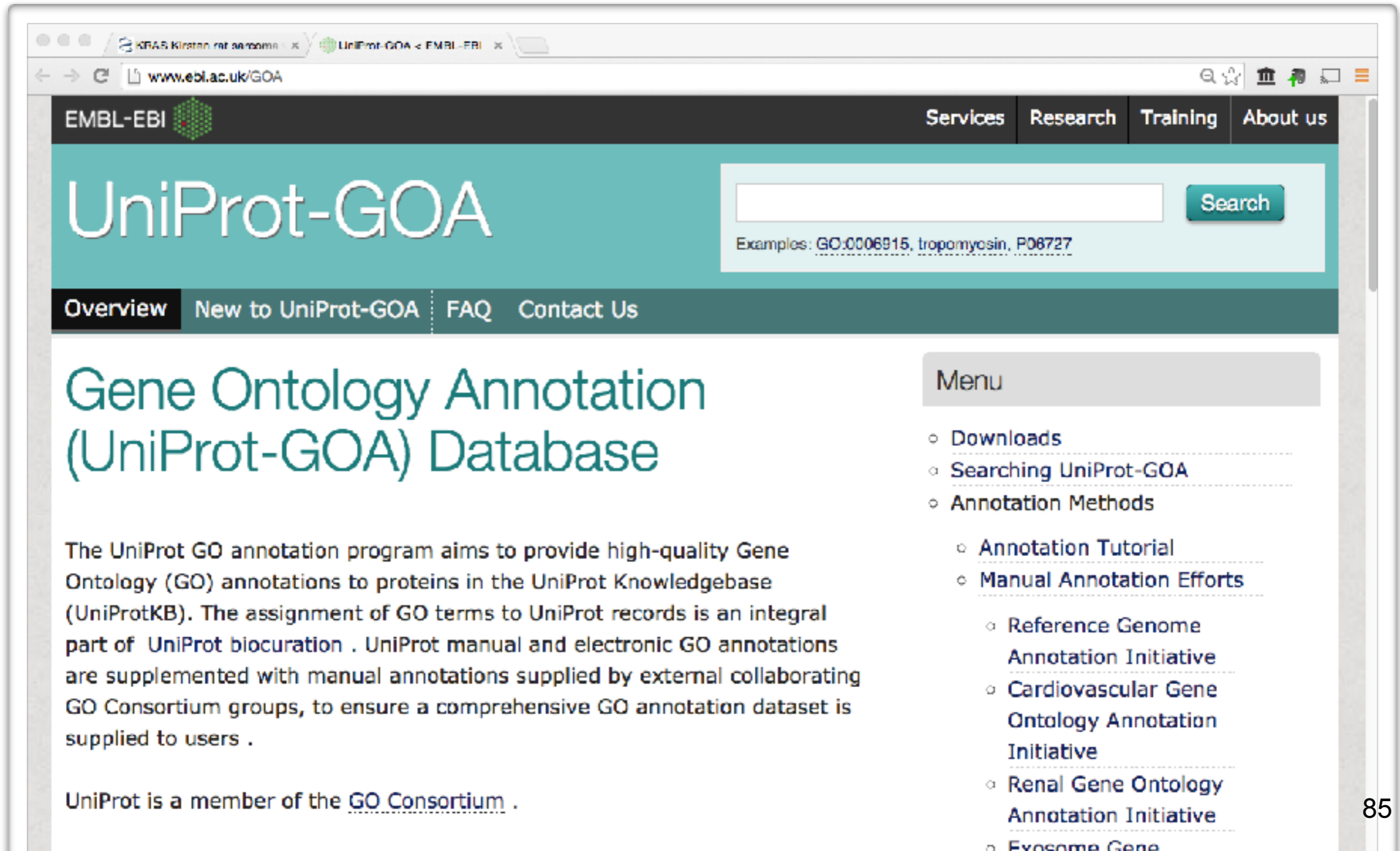
Items 1 - 25 of 33 < Prev Page 1 of 2 Next >

Process	Evidence Code	Pubs
Fc-epsilon receptor signaling pathway	TAS	
GTP catabolic process	IEA	
MAPK cascade	TAS	
Ras protein signal transduction	TAS	
actin cytoskeleton organization	IEA	
activation of MAPKK activity	TAS	
axon guidance	TAS	
blood coagulation	TAS	



GO: Gene Ontology

GO provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data



The screenshot shows the UniProt-GOA website interface. At the top, there is a navigation bar with links for Services, Research, Training, and About us. Below this is a search bar with a search button and examples of search terms: GO:0006915, tropomyosin, P08727. The main heading is "Gene Ontology Annotation (UniProt-GOA) Database". A menu on the right lists various resources: Downloads, Searching UniProt-GOA, Annotation Methods, Annotation Tutorial, Manual Annotation Efforts, Reference Genome Annotation Initiative, Cardiovascular Gene Ontology Annotation Initiative, Renal Gene Ontology Annotation Initiative, and Exosome Gene.

EMBL-EBI [Services](#) [Research](#) [Training](#) [About us](#)

UniProt-GOA

[Search](#)

Examples: [GO:0006915](#), [tropomyosin](#), [P08727](#)

[Overview](#) [New to UniProt-GOA](#) [FAQ](#) [Contact Us](#)

Gene Ontology Annotation (UniProt-GOA) Database

The UniProt GO annotation program aims to provide high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB). The assignment of GO terms to UniProt records is an integral part of UniProt biocuration. UniProt manual and electronic GO annotations are supplemented with manual annotations supplied by external collaborating GO Consortium groups, to ensure a comprehensive GO annotation dataset is supplied to users.

UniProt is a member of the [GO Consortium](#).

Menu

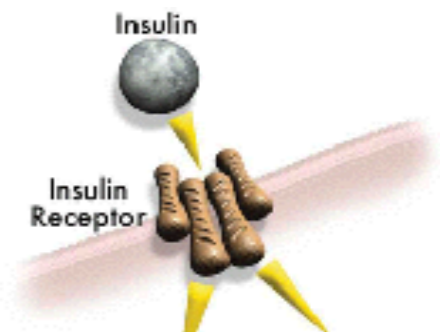
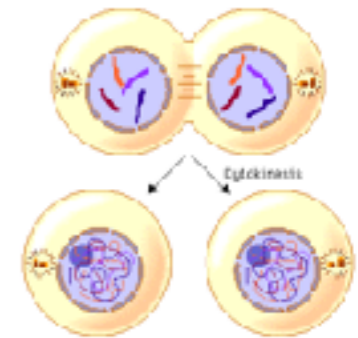
- [Downloads](#)
- [Searching UniProt-GOA](#)
- [Annotation Methods](#)
- [Annotation Tutorial](#)
- [Manual Annotation Efforts](#)
 - [Reference Genome Annotation Initiative](#)
 - [Cardiovascular Gene Ontology Annotation Initiative](#)
 - [Renal Gene Ontology Annotation Initiative](#)
 - [Exosome Gene](#)

Why do we need Ontologies?

- Annotation is essential for capturing the understanding and knowledge associated with a sequence or other molecular entity
- Annotation is traditionally recorded as “free text”, which is easy to read by humans, but has a number of disadvantages, including:
 - ▶ Difficult for computers to parse
 - ▶ Quality varies from database to database
 - ▶ Terminology used varies from annotator to annotator
- Ontologies are annotations using standard vocabularies that try to address these issues
- GO is integrated with UniProt and many other databases including a number at NCBI

GO Ontologies

- There are three ontologies in GO:
 - ▶ **Biological Process**
A commonly recognized series of events
e.g. cell division, mitosis,
 - ▶ **Molecular Function**
An elemental activity, task or job
e.g. kinase activity, insulin binding
 - ▶ **Cellular Component**
Where a gene product is located
e.g. mitochondrion, mitochondrial
membrane



www.ncbi.nlm.nih.gov/gene/3845#gene-ontology

Gene Ontology Provided by GOA

Function	Evidence Code	Pubs
GDP binding		
GMP binding		
GTP binding		
LRR domain binding		
protein binding		
protein complex binding		

Process

Code	Pubs
Fc-epsilon receptor signaling pathway	TAS
GTP catabolic process	IEA
MAPK cascade	TAS
Ras protein signal transduction	TAS
actin cytoskeleton organization	IEA
activation of MAPKK activity	TAS
axon guidance	TAS
blood coagulation	TAS

The 'Gene Ontology' or GO is actually maintained by the EBI so lets switch or link over to UniProt also from the EBI.

⋮ Scroll down to
▼ **UniProt** link

UniProt will detail much more information for protein coding genes such as this one

The screenshot shows the NCBI Gene page for X01669.1. The browser address bar displays www.ncbi.nlm.nih.gov/gene/3345#gene-ontology. The page header includes the genomic coordinates X01669.1 and CAA25828.1, and a pagination indicator for items 1-25 of 43 on page 1 of 2. A table of links is present, with the UniProtKB link highlighted in red:

Protein Accession	Links	
	GenPept Link	UniProtKB Link
P01116.1	GenPept	UniProtKB/Swiss-Prot:P01116

Below the table is an 'Additional links' section. At the bottom of the page, there is a navigation menu with the following categories:

- GETTING STARTED**
 - NCBI Education
 - NCBI Help Manual
 - NCBI Handbook
 - Training & Tutorials
- RESOURCES**
 - Chemicals & Bioassays
 - Data & Software
 - DNA & RNA
 - Domains & Structures
 - Genes & Expression
 - Genetics & Medicine
 - Genomes & Maps
 - Homology
 - Literature
 - Proteins
 - Sequence Analysis
 - Taxonomy
- POPULAR**
 - PubMed
 - Bookshelf
 - PubMed Central
 - PubMed Health
 - BLAST
 - Nucleotide
 - Genome
 - SNP
 - Gene
 - Protein
 - PubChem
- FEATURED**
 - Genetic Testing Registry
 - PubMed Health
 - GenBank
 - Reference Sequences
 - Gene Expression Omnibus
 - Map Viewer
 - Human Genome
 - Mouse Genome
 - Influenza Virus
 - Primer-BLAST
 - Sequence Read Archive
- NCBI INFORMATION**
 - About NCBI
 - Research at NCBI
 - NCBI News
 - NCBI FTP Site
 - NCBI on Facebook
 - NCBI on Twitter
 - NCBI on YouTube

A red arrow points to the UniProt link with the text 'Scroll down to UniProt link'.

UniProt will detail much more information for protein coding genes

P01116 - RAS_HUMAN

Protein | **GTPase KRas**
Gene | **KRAS**
Organism | *Homo sapiens (Human)*
Status | Reviewed - - Experimental evidence at protein levelⁱ

BLAST Align Retrieve/ID Mapping Help Contact

Basket

Display None

- FUNCTION
- NAMES & TAXONOMY
- SUBCELL. LOCATION
- PATHOL./BIOTECH
- PTM / PROCESSING
- EXPRESSION
- INTERACTION
- STRUCTURE
- FAMILY & DOMAINS
- SEQUENCES (2)
- CROSS-REFERENCES

BLAST Align Format Add to basket History Feedback Help video

Functionⁱ

Ras proteins bind GDP/GTP and possess Intrinsic GTPase activity. Plays an Important role in the regulation of cell proliferation (PubMed:23698361, PubMed:22711838).

Enzyme regulationⁱ

Alternates between an inactive form bound to GDP and an active form bound to GTP. Activated by a guanine nucleotide-exchange factor (GEF) and inactivated by a GTPase-activating protein (GAP). Interaction with SOS1 promotes exchange of bound GDP by GTP.

Regions

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Nucleotide binding ⁱ	10 – 18	9	GTP			
Nucleotide binding ⁱ	29 – 35	7	GTP			
Nucleotide binding ⁱ	59 – 60	2	GTP			



UniProtKB

Advanced



BLAST Align Retrieve/ID Mapping

Help Contact

Basket

P01116 - RASK_HUMAN

- Protein: **GTPase KRas**
- Gene: **KRAS**
- Organism: *Homo sapiens (Human)*
- Status: Reviewed -

Display None

- FUNCTION
- NAMES & TAXONOMY
- SUBCELL LOCATION
- PATHOL/BIOTECH
- PTM / PROCESSING
- EXPRESSION
- INTERACTION
- STRUCTURE
- FAMILY & DOMAINS
- SEQUENCES (2)
- CROSS-REFERENCES

Function

Ras proteins bind GDP/GTP and possess intrinsic GTPase activity. Plays an important role in the regulation of cell proliferation (PubMed:23698361, PubMed:22711838).

Enzyme regulation

Alternates between an inactive form bound to GDP and an active form bound to GTP. Activated by a guanine nucleotide-exchange factor (GEF) and inactivated by a GTPase-activating protein (GAP). Interaction with SOS1 promotes exchange of bound GDP by GTP.

Regions

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Nucleotide binding ⁱ	10 – 18	9	GTP			
Nucleotide binding ⁱ	29 – 35	7	GTP			
Nucleotide binding ⁱ	59 – 60	2	GTP			



Example Questions:
 What positions in the protein are responsible for GTP binding?

Example Questions:

What variants of this enzyme are involved in gastric cancer and other human diseases?

www.uniprot.org/uniprot/P01116

Display None

- FUNCTION
- NAMES & TAXONOMY
- SUBCELL LOCATION
- PATHOL/BIOTECH**
- PTM / PROCESSING
- EXPRESSION
- INTERACTION
- STRUCTURE
- FAMILY & DOMAINS
- SEQUENCES (2)
- CROSS-REFERENCES
- PUBLICATIONS
- ENTRY INFORMATION
- MISCELLANEOUS
- SIMILAR PROTEINS

[▲ Top](#)

Pathology & Biotech

Involvement in disease¹

LEUKEMIA, ACUTE MYELOGENOUS (AML)

[MIM:601626]: A subtype of acute leukemia, a cancer of the white blood cells. AML is a malignant disease of bone marrow characterized by maturational arrest of hematopoietic precursors at an early stage of development. Clonal expansion of myeloid blasts occurs in bone marrow, blood, and other tissue. Myelogenous leukemias develop from changes in cells that normally produce neutrophils, basophils, eosinophils and monocytes. [1 Publication](#)

Note: The disease is caused by mutations affecting the gene represented in this entry.

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Natural variant ¹	10 – 10		1 G → GG in one individual with AML; expression in 3T3 cell causes cellular transformation; expression in COS cells activates the Ras-MAPK signaling pathway; lower GTPase activity; faster GDP dissociation rate. 1 Publication		VAR_034601	

LEUKEMIA, JUVENILE MYELOMONOCYTIC (JMML)

[MIM:607785]: An aggressive pediatric myelodysplastic syndrome/myeloproliferative disorder characterized by malignant transformation in the hematopoietic stem cell compartment with proliferation of differentiated progeny. Patients have splenomegaly, enlarged lymph nodes, rashes, and hemorrhages. **Note:** The disease is caused by mutations affecting the gene represented in this entry.

NOONAN SYNDROME 3 (NS3)

[MIM:609942]: A form of Noonan syndrome, a disease characterized by short stature, facial dysmorphic features such as hypertelorism, a downward eyeslant and low-set posteriorly rotated ears, and a high incidence of congenital heart

Example Questions:

Are high resolution protein structures available to examine the details of these mutations?

The screenshot shows the UniProt website for protein P01116 (KRAS). The 'Structure' tab is selected in the left sidebar. The main content area displays the 'Secondary structure' as a bar chart and a table of '3D structure databases'. The table lists various X-ray crystallography structures with their resolutions and positions. A red arrow points to the bottom of the table.

Select the link destinations:	Entry	Method	Resolution (Å)	Chain	Positions	PDBsum
<input checked="" type="radio"/> PDBe ⁱ	1D8D	X-ray	2.00	P	178-188	[*]
<input type="radio"/> RCSB PDB ⁱ	1D8E	X-ray	3.00	P	178-188	[*]
<input type="radio"/> PDBj ⁱ	1KZO	X-ray	2.20	C	169-173	[*]
	1KZP	X-ray	2.10	C	169-173	[*]
	3GFT	X-ray	2.27	A/B/C/D/E/F	1-164	[*]
	4DSN	X-ray	2.03	A	2-164	[*]
	4DSO	X-ray	1.85	A	2-164	[*]
	4EPR	X-ray	2.00	A	1-164	[*]
	4EPT	X-ray	2.00	A	1-164	[*]
	4EPV	X-ray	1.35	A	1-164	[*]
	4EPW	X-ray	1.70	A	1-164	[*]
	4EPX	X-ray	1.76	A	1-164	[*]
	4EPY	X-ray	1.80	A	1-164	[*]
	4L8G	X-ray	1.52	A	1-169	[*]
	4LDJ	X-ray	1.15	A	1-164	[*]
	4LPK	X-ray	1.50	A/B	1-169	[*]

Example Questions:

What is known about the protein family, its species distribution, number in humans and residue-wise conservation, etc... ?

www.uniprot.org/uniprot/P01116

Display **None**

- FUNCTION
- NAMES & TAXONOMY
- SUBCELL LOCATION
- PATHOL/BIOTECH
- PTM / PROCESSING
- EXPRESSION
- INTERACTION
- STRUCTURE
- FAMILY & DOMAINS**
- SEQUENCES (2)
- CROSS-REFERENCES
- PUBLICATIONS
- ENTRY INFORMATION
- MISCELLANEOUS
- SIMILAR PROTEINS

OrthoDBⁱ E01
PhylomeDBⁱ P01
TreeFamⁱ TF3

Family and domain databases

Gene3D ⁱ	3.40.50.300. 1 hit.
InterPro ⁱ	IPR027417. P-loop_NTPase. IPR005225. Small_GTP-bd_dom. IPR001806. Small_GTPase. IPR020849. Small_GTPase_Ras. [Graphical view]
PANTHER ⁱ	PTHR24070. PTHR24070. 1 hit.
Pfam ⁱ	PF00071. Ras. 1 hit. [Graphical view]
PRINTS ⁱ	PR00449. RASTRNSFRMNG.
SMART ⁱ	SM00173. RAS. 1 hit. [Graphical view]
SUPFAM ⁱ	SSF52540. SSF52540. 1 hit.
TIGRFAMs ⁱ	TIGR00231. small_GTP. 1 hit.
PROSITE ⁱ	PS51421. RAS. 1 hit. [Graphical view]


PFAM is one of the best protein family databases

Sequences (2)ⁱ

Sequence statusⁱ: Complete.
Sequence processingⁱ: The displayed sequence is further processed into a mature form.
This entry describes **2** isoformsⁱ produced by **alternative splicing**. [Align](#)

Example Questions:

What is known about the protein family, its **species distribution**, number in humans and residue-wise conservation, etc... ?

EMBL-EBI  HOME | SEARCH

Family: Ras (PF00071)

332 architectures 21243 sequences 30 interactions 1006 species 663 structures

- Summary
- Domain organisation
- Clan
- Alignments
- HMM logo
- Trees
- Curation & model
- Species**
- Interactions
- Structures

Jump to...

Summary: Ras family

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

[Wikipedia: Ras subfamily](#) [Wikipedia: Ras superfamily](#) [Pfam](#) [InterPro](#)

This is the Wikipedia entry entitled "[Ras subfamily](#)". [More...](#)

Ras subfamily [Edit Wikipedia article](#)

This article is about p21/Ras protein. For the p21/waf1 protein, see p21.

Ras is the name given to a family of related proteins which is ubiquitously expressed in all cell lineages and organs. All Ras protein family members belong to a class of protein called *small GTPase*, and are involved in transmitting signals within cells (cellular signal transduction). Ras is the prototypical member of the Ras superfamily of proteins, which are all related in 3D structure and regulate diverse cell behaviours.

The name 'Ras' is an abbreviation of 'Rat sarcoma', reflecting the way the first members of the protein family were discovered. The name ras is also used to refer to the family of genes encoding those proteins.


When Ras is 'switched on' by incoming signals, it subsequently switches on other proteins, which ultimately turn on genes involved in cell growth, differentiation and survival. As a result, mutations in ras genes can lead to the production of permanently activated Ras proteins. This can cause unintended and overactive signalling inside the cell, even in the absence of incoming signals.

Because these signals result in cell growth and division, overactive Ras signalling can ultimately lead to cancer.^[2] The 3 Ras genes in humans (HRAS, KRAS, and NRAS) are the most common oncogenes in human cancer; mutations that permanently activate Ras are found in 20% to 25% of all human tumors and up to 90% in certain types of cancer (e.g., pancreatic cancer).

^[2] For this reason, Ras inhibitors are being studied as a treatment for cancer, and other diseases with Ras overexpression.

Contents [\[hide\]](#)

- History
- Structure
- Function
 - 3.1 Activation and deactivation
 - 3.2 Membrane attachment
- Members
- Ras in cancer
 - 5.1 Inappropriate activation
 - 5.2 Constitutively active Ras

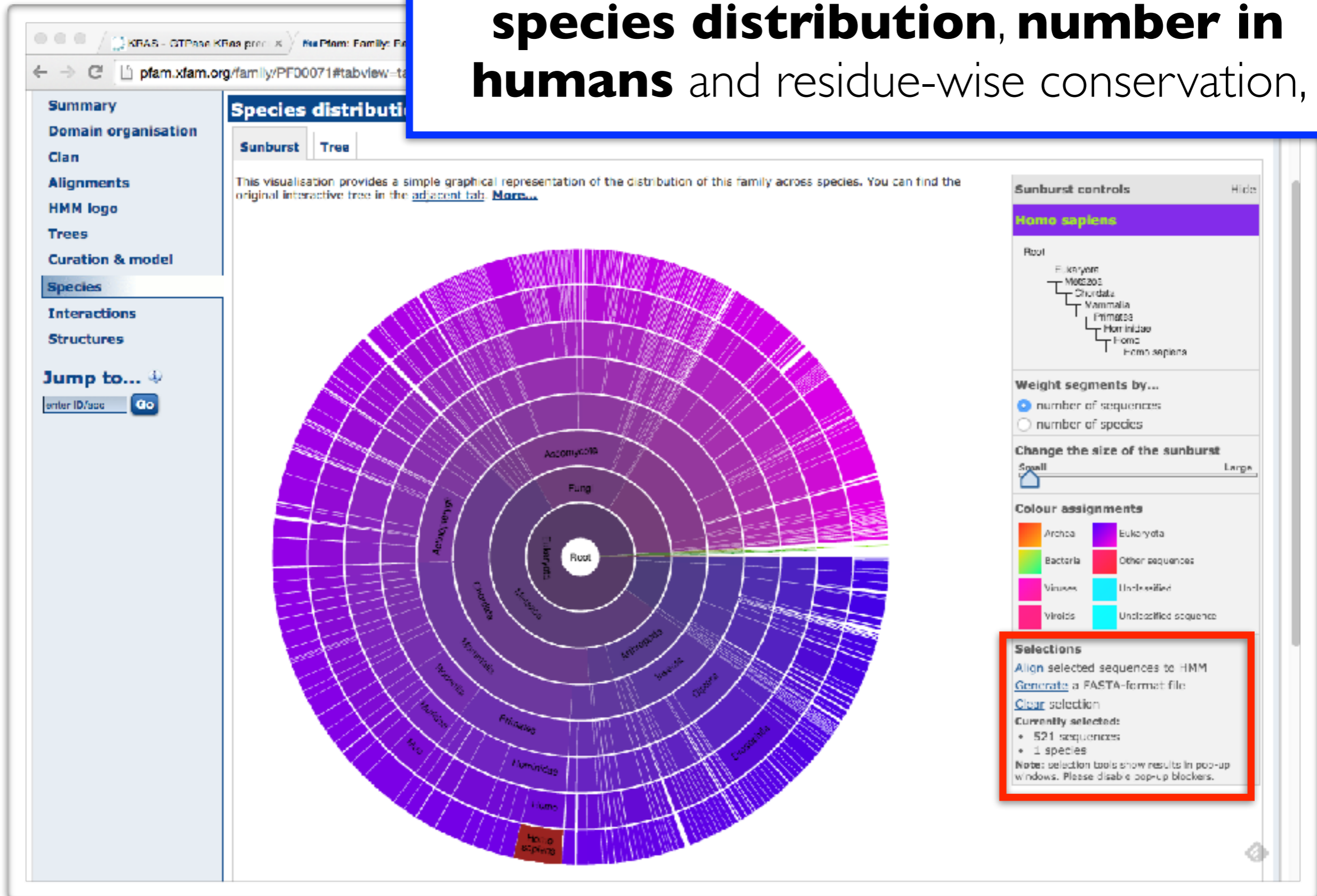


H-Ras structure PDB 121p, surface colored by conservation in Pfam seed alignment: gold, most conserved; dark cyan, least conserved.

Identifiers	
Symbol	Ras
Pfam	PF00071 E
InterPro	IPR013753 E
PROSITE	PDOC00017 E
SCOP	5p21 E
SUPERFAMILY	5p21 E

Example Questions:

What is known about the protein family, its **species distribution**, **number in humans** and residue-wise conservation,



Example Questions:

What is known about the protein family, its species distribution, number in humans and **residue-wise conservation**, etc... ?

The image displays a composite of three browser windows from the Pfam database. The top window shows the 'Species distribution' page for protein family PF00071, featuring a phylogenetic tree and a 'Sunburst controls' panel. The middle window shows the 'Alignment for selected sequences' page, displaying a multiple sequence alignment with a color-coded conservation scale. The bottom window shows the 'Sunburst controls' panel in more detail, including a 'Selections' section with a red border.

Species distribution controls:

- can find the
- Sunburst controls (Hide)
- Homo sapiens**
- Root
- Phylogenetic tree showing relationships between Eukaryota, Metazoa, Chordata, Mammalia, Primates, Hominoidea, and Homo sapiens.
- Weight segments by...
 - number of sequences
 - number of species
- Change the size of the sunburst (Small to Large)
- Colour assignments:
 - Archaea (Orange)
 - Bacteria (Green)
 - Viruses (Pink)
 - Vireids (Light Blue)
 - Eukaryota (Purple)
 - Other sequences (Red)
 - Unclassified (Cyan)
 - Unclassified sequence (Light Blue)
- Selections** (highlighted in red):
 - Align selected sequences to HMM
 - Generate a FASTA-format file
 - Clear selection
 - Currently selected:
 - + 521 sequences
 - + 1 species
 - Note: selection tools show results in pop-up windows. Please disable pop-up blockers.

Alignment for selected sequences:

EMBL-EBI logo

Alignment for selected sequences

Currently showing rows 1 to 30 of 536 rows in this alignment. Show 30 rows of alignment

PF11234/16-178	..KLVVGGSCGVNSALTI.....Q.....FM.....Y..D..EF..V....E..DYEPRK..-AD...SYRKKVLD.....
PF11112/5-160	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
Q14068/38-204	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
Q98883/7-173	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
PF15153/5-178	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
Q00194/11-183	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
Q15907/13-174	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
PF10114/5-166	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
PF11153/10-171	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
PF55040/77-241	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
PF55042/93-253	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
PF01116/5-160	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
Q98097/21-182	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
Q9ULC3/11-171	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
Q14807/15-177	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
Q9NKS7/7-202	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
Q98082/35-201	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
Q95905/9-174	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
PF11149/10-175	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
Q9ULN5/65-227	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
PF57735/14-175	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
PF11153/11-183	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
PF11111/5-166	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
PF11233/16-177	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
Q9UL25/21-182	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
Q9NP72/10-171	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
Q98004/10-171	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
Q90626/7-168	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
Q9UBK7/23-179	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....
PF11157/14-179	..KLVVGGSCGVNSALTI.....Q.....LI.....Q..N..HF..V....D..EYDPTI..-ED...SYRKKVLD.....



There are 18 pages in this alignment. Show page 1

Download this alignment.

Close window

Example Questions:

What is known about the protein family, its species distribution, number in humans and **residue-wise conservation**, etc... ?

EMBL-EBI  HOME | SEARCH | BROWSE | FTP | HELP | ABOUT  keyword search Go

Family: Ras (PF00071)

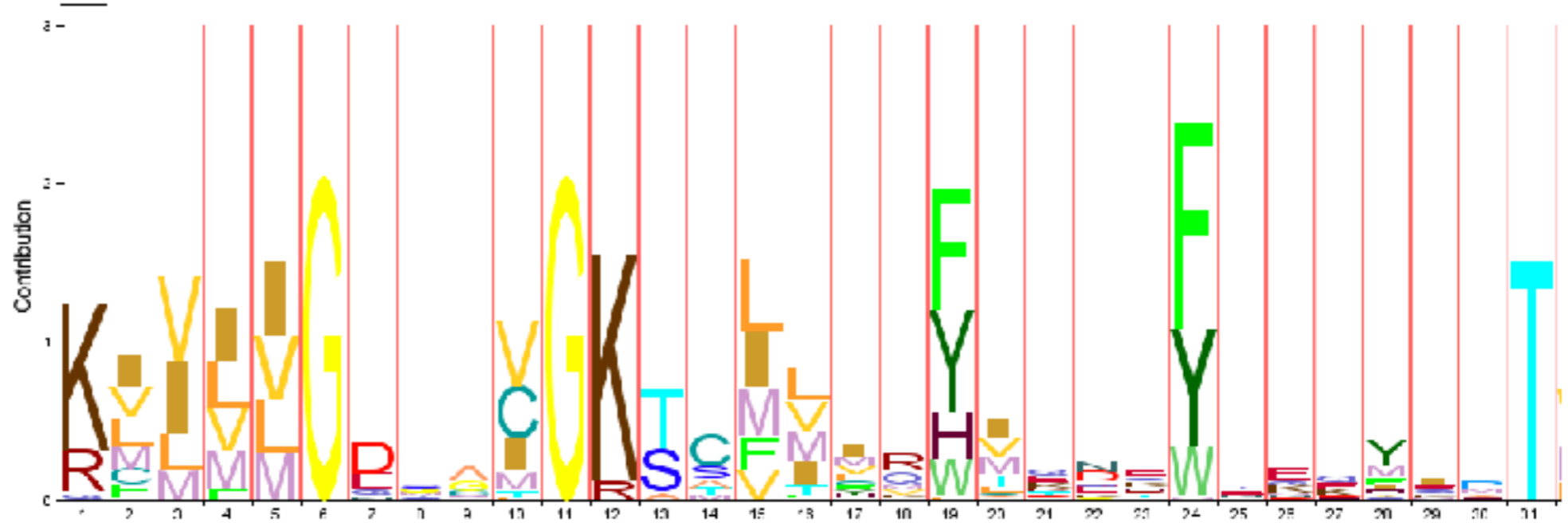
332 architectures 21243 sequences 30 interactions 1006 species 553 structures

- Summary
- Domain organisation
- Clan
- Alignments
- HMM logo**
- Trees
- Curation & model
- Species
- Interactions
- Structures

Jump to... enter ID/acc

HMM logo






HMM logos is one way of visualising profile HMMs. Logos provide a quick overview of the properties of an HMM in a graphical form. You can see a more detailed description of HMM logos and find out how you can interpret them here? [More...](#)



Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.
European Molecular Biology Laboratory

Family: *Kinesin* (PF00225)

⌂ Loading page components (1 remaining)...

 126 architectures
  4150 sequences
  6 interactions
  248 species
  114 structures

Summary

Domain organisation

Clans

Alignments

HMM logo

Trees

Curation & models

Species

Interactions

Structures

Jump to...

Interactions

There are **6** interactions for this family. [More...](#)

[Tubulin](#)
[Tubulin_C](#)

[Tubulin_C](#)

[Kinesin](#)

[Tubulin](#)

[Kinesin](#)

Questions or comments: pfam@janelia.hhmi.org

Howard Hughes Medical Institute

Family: *Kinesin* (PF00225)

 126 architectures
  4150 sequences
  6 interactions
  248 species
  114 structures

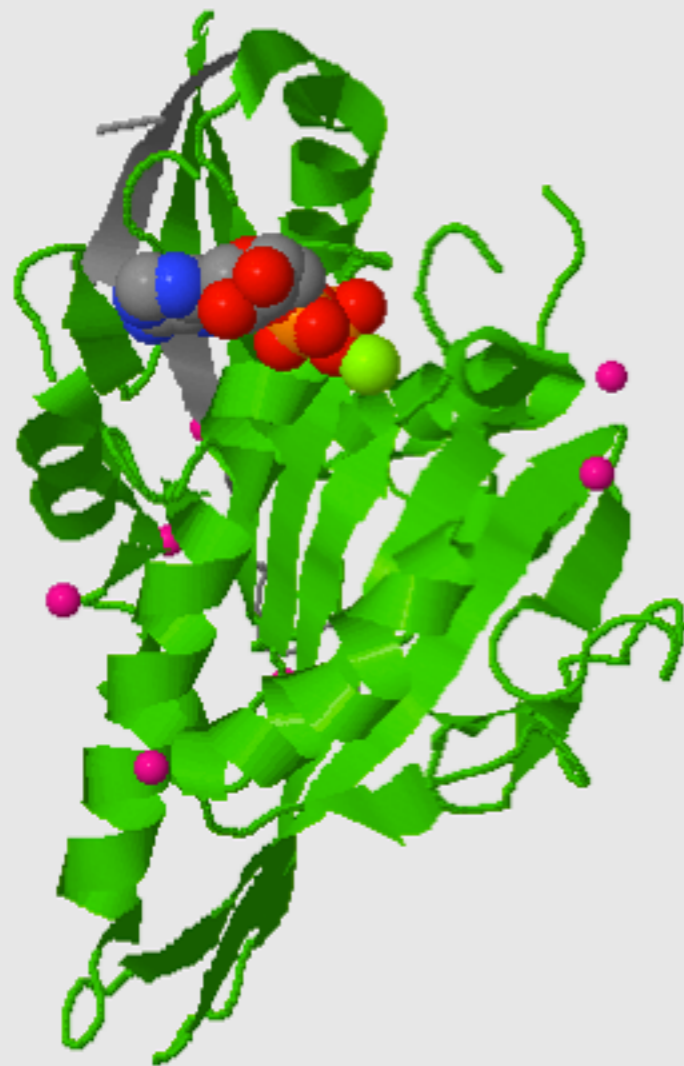
[Summary](#)
[Domain organisation](#)
[Clans](#)
[Alignments](#)
[HMM logo](#)
[Trees](#)
[Curation & models](#)
[Species](#)
[Interactions](#)
[Structures](#)
[Jump to...](#)

Structures

For those sequences which have a structure in the [Protein DataBank](#), we use the mapping between [UniProt](#), PDB and Pfam coordinate systems from the [PDBer](#) group, to allow us to map Pfam domains onto UniProt sequences and three-dimensional protein structures. The table below shows the structures on which the **Kinesin** domain has been found.

UniProt entry	UniProt residues	PDB ID	PDB chain ID	PDB residues	View	
A8BKD1_GIALA	11 - 335	2vvq	A	11 - 335	Jmol AstexViewer SPICE	
			B	11 - 335	Jmol AstexViewer SPICE	
CENPE_HUMAN	12 - 329	1t5c	A	12 - 329	Jmol AstexViewer SPICE	
			B	12 - 329	Jmol AstexViewer SPICE	
KAR3_YEAST	392 - 723		1f9t	A	392 - 723	Jmol AstexViewer SPICE
			1f9u	A	392 - 723	Jmol AstexViewer SPICE
			1f9v	A	392 - 723	Jmol AstexViewer SPICE
			1f9w	A	392 - 723	Jmol AstexViewer SPICE
			1f9w	B	392 - 723	Jmol AstexViewer SPICE
			3kar	A	392 - 723	Jmol AstexViewer SPICE
KI13B_HUMAN	11 - 352	3qbj	A	11 - 352	Jmol AstexViewer SPICE	
			B	11 - 352	Jmol AstexViewer SPICE	
			C	11 - 352	Jmol AstexViewer SPICE	
			1ii6	A	24 - 359	Jmol AstexViewer SPICE
			1ii6	B	24 - 359	Jmol AstexViewer SPICE
			1q0b	A	24 - 359	Jmol AstexViewer SPICE
				B	24 - 359	Jmol AstexViewer SPICE
			1x88	A	24 - 359	Jmol AstexViewer SPICE
				B	24 - 359	Jmol AstexViewer SPICE
			A	24 - 359	Jmol AstexViewer SPICE	

PDB entry 3bfm



Jmol

PDB			UniProt			Pfam family	Colour
Chain	Start	End	ID	Start	End		
A	49	368	KIF22_HUMAN	49	368	Kinesin (PF00225)	

SUMMARY

- Bioinformatics is computer aided biology.
- Bioinformatics deals with the collection, archiving, organization, and interpretation of a wide range of biological data.
- There are a large number of primary, secondary and tertiary bioinformatics databases.
- The NCBI and EBI are major online bioinformatics service providers.
- Introduced GenBank, RefSeq, UniProt, PDB databases as well as a number of 'boutique' databases including PFAM and OMIM.
- Introduced the notion of *controlled vocabularies* and *ontologies*.
- Described the use of ENTREZ and BLAST for searching databases.

HOMEWORK

- ☑ Complete the **initial course questionnaire**:
<http://tinyurl.com/bioinf525-questions>
- ☑ Check out the “**Background Reading**” material online:
[PDF1 \(bioinformatics review\)](#),
[PDF 2 \(bioinformatics challenges\)](#).
- ☑ Complete the **lecture 1.1 homework questions**:
<http://tinyurl.com/bioinf525-quiz1>

THANK YOU

The text "THANK YOU" is displayed in a large, bold, sans-serif font. Each letter is a different color: T (green), H (blue), A (black), N (magenta), K (blue), Y (green), O (black), and U (black). Below each letter is a small number from 1 to 8. The letters are overlapping: the 'H' overlaps the 'T', the 'N' overlaps the 'A', the 'K' overlaps the 'N', the 'Y' overlaps the 'K', the 'O' overlaps the 'Y', and the 'U' overlaps the 'O'. The numbers 1-8 are positioned below the letters: 1 under T, 2 under H, 3 under A, 4 under N, 5 under K, 6 under Y, 7 under O, and 8 under U.

ADDITIONAL DATABASES OF NOTE
(SLIDES FOR YOUR REFERENCE)

ENTREZ & BLAST:

TOOLS FOR SEARCHING AND ACCESSING
MOLECULAR DATA AT NCBI

Entrez: Integrated search of NCBI databases

The image shows the NCBI Entrez homepage. On the left is a navigation menu with categories like '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', and 'Variation'. A search bar is at the top right with a 'Search' button. A dropdown menu is open, listing various databases such as 'All Databases', 'PubMed', 'Protein', 'Nucleotide', 'GSS', 'EST', 'Structure', 'Genome', 'BioProject', 'BioSample', 'BioSystems', 'Books', 'Conserved Domains', 'Clone', 'dbGaP', 'dbVar', 'Epigenomics', 'Gene', 'GEO DataSets', 'GEO Profiles', 'HomoloGene', 'MeSH', 'NCBI Web Site', 'NLM Catalog', 'OMIA', 'OMIM', 'PMC', 'PopSet', 'Probe', 'Protein Clusters', 'PubChem BioAssay', 'PubChem Compound', 'PubChem Substance', 'PubMed Health', 'SNP', 'SRA', 'Taxonomy', 'ToolKit', 'ToolKitAll', 'UniGene', and 'UniSTS'. A central text box states: 'Entrez is available from the main NCBI homepage or from the homepage of individual databases'. On the right, there are sections for 'Popular Resources' (PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, PubChem) and 'NCBI Announcements' (BI's April Newsletter is on the Bookshelf, Information about May's Discovery Workshop, the new GTR and Assembly, New Filter Sidebar will be added to PubMed, A Filter Sidebar will be added soon to the PubMed result pages, DELTA BLAST - more sensitive protein searching, Domain Enhanced Lookup Time Accelerated BLAST (DELTA-BLAST)).

NCBI
National Center for
Biotechnology Information

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

✓ All Databases
PubMed
Protein
Nucleotide
GSS
EST
Structure
Genome
BioProject
BioSample
BioSystems
Books
Conserved Domains
Clone
dbGaP
dbVar
Epigenomics
Gene
GEO DataSets
GEO Profiles
HomoloGene
MeSH
NCBI Web Site
NLM Catalog
OMIA
OMIM
PMC
PopSet
Probe
Protein Clusters
PubChem BioAssay
PubChem Compound
PubChem Substance
PubMed Health
SNP
SRA
Taxonomy
ToolKit
ToolKitAll
UniGene
UniSTS

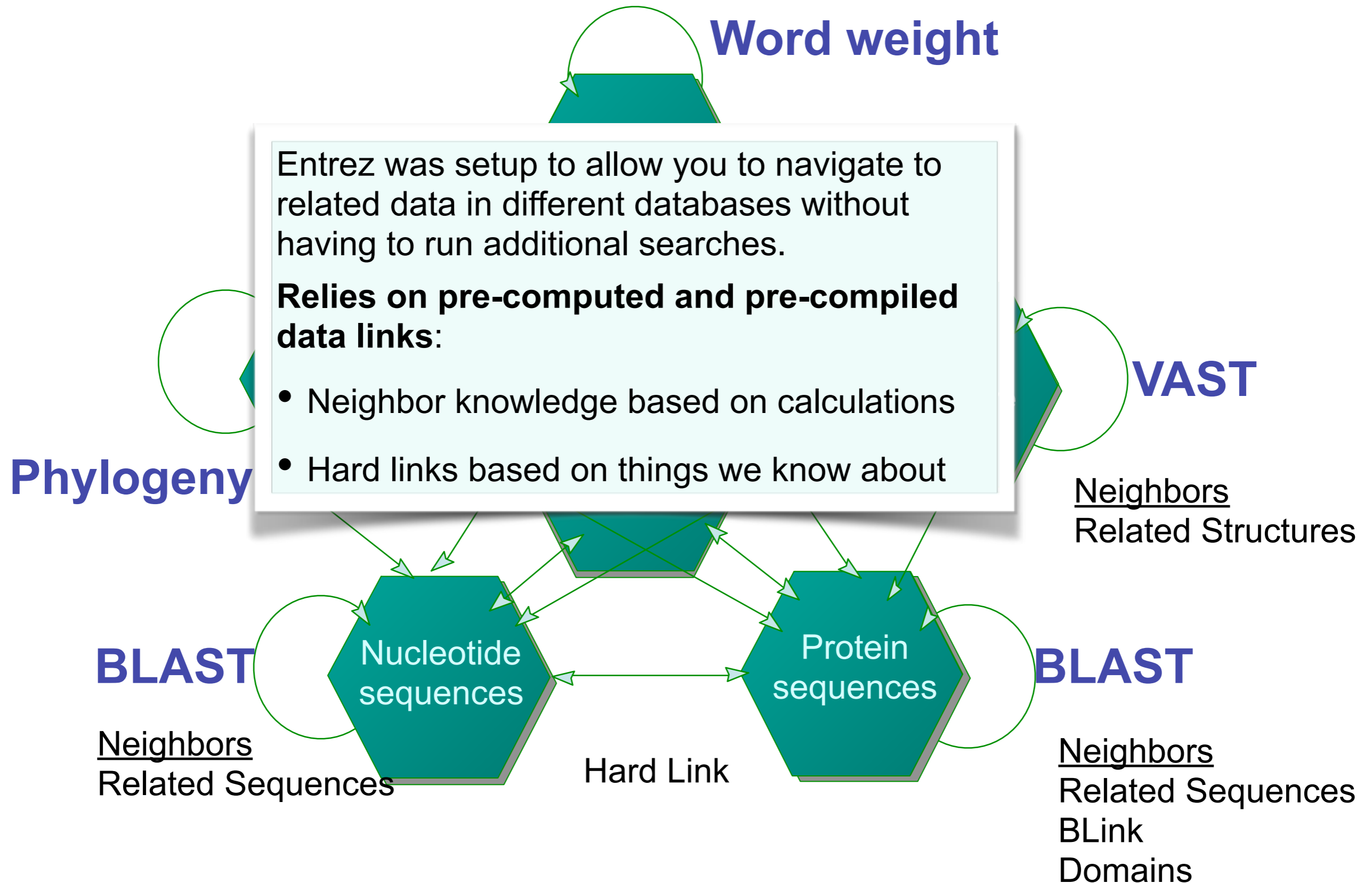
Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
About NCBI | Mission | Organization | Research | RSS Feeds

Popular Resources
PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI Announcements
NCBI's April Newsletter is on the Bookshelf
04 May 2012
Information about May's Discovery Workshop, the new GTR and Assembly
New Filter Sidebar will be added to PubMed
03 May 2012
A Filter Sidebar will be added soon to the PubMed result pages. The useful
DELTA BLAST - more sensitive protein searching
30 Apr 2012
Domain Enhanced Lookup Time Accelerated BLAST (DELTA-BLAST)
More...

Entrez is available from the main NCBI homepage or from the homepage of individual databases

Entrez: navigating across databases



Global Entrez Query: All NCBI Databases

The screenshot shows the NCBI Global Entrez Query interface. The search term "ras" has been entered, resulting in approximately 2,978,774 search results. A list of integrated databases is displayed, including Literature, Books, MeSH, NLM Catalog, PubMed, PubMed Central, Health, ClinVar, dbGaP, GTR, EST, GEO Profiles, HomoloGene, PopSet, UniGene, and Proteins. A URL box highlights the main query page, and a text box states: "The Entrez system: 38 (and counting) integrated databases".

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

The Entrez system: 38 (and counting) integrated databases

Database	Count	Description
Literature		
Books	1,000	books and book chapters
MeSH	402	ontology used for PubMed indexing
NLM Catalog	223	information about gene
PubMed	54,672	clinical studies
PubMed Central	96,114	and molecular
Health		
ClinVar	759	human variations of clinical significance
dbGaP	120	genotype/phenotype interaction studies
GTR	1,879	genetic testing registry
EST	3,985	sequences
GEO Profiles	1,022,789	abundance profiles
HomoloGene	696	homologous gene sets for selected organisms
PopSet	2,254	sequence sets from phylogenetic and population studies
UniGene	4,770	clusters of expressed transcripts
Proteins		

Search Results

Nucleotide [Save search](#) [Limits](#) [Advanced](#) [Help](#)

Display Settings: Summary, 20 per page, Sorted by Default order

Found 2324 nucleotide sequences. Nucleotide (35) EST (2289)

Send to:

Filter your results:

- All (35)
- Bacteria (0)
- INSDC (GenBank) (27)
- mRNA (32)
- RefSeq (8)

[Manage Filters](#)

Top Organisms [Tree](#)

- Danio rerio (29)
- Ictalurus furcatus (6)

Find related data

Database:

Search details

```
("Danio rerio"[Organism] OR zebrafish[All Fields]) AND creatine kinase[All Fields]
```

[See more...](#)

Recent activity

Results: 1 to 20 of 35 << First < Prev Page 1 of 2 Next > Last >>

- [Danio rerio creatine kinase, muscle b \(ckmb\), mRNA](#)
1. 1,463 bp linear mRNA
Accession: NM_001105683.1 GI: 157787180
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- [Danio rerio zgc:63663 \(zgc:63663\), mRNA](#)
2. 2,478 bp linear mRNA
Accession: NM_200614.1 GI: 41055386
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- [Danio rerio creatine kinase, muscle b \(ckmb\), mRNA](#)
3. 1,552 bp linear mRNA
Accession: NM_130932.1 GI: 18858426
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- [Danio rerio creatine kinase, mitochondrial 2 \(sarcomeric\), mRNA \(cDNA clone MGC:198091 IMAGE:9039080\), complete cds](#)
4. 1,296 bp linear mRNA
Accession: BC171364.1 GI: 213624628
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- [Danio rerio creatine kinase, mitochondrial 2 \(sarcomeric\), mRNA \(cDNA clone MGC:172259 IMAGE:8798676\), complete cds](#)
5. 1,400 bp linear mRNA
Accession: BC154617.1 GI: 159155933
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

Discovery Column (sort, filter, link)

Advanced: Search Builder

Nucleotide Advanced Search Builder

zebrafish[Organism] AND "creatine kinase"[Title]

[Clear](#)

Helps build complex fielded queries

Organism [Show index list](#)

AND [Hide index list](#)

- creatine kinase (749)
- creatine kinase 1 (6)
- creatine kinase 2 (2)
- creatine kinase b (30)
- creatine kinase b gene (3)
- creatine kinase b mrna (3)
- creatine kinase b pseudogene 1 (1)
- creatine kinase b subunit (1)
- creatine kinase brain (43)
- creatine kinase chain b (1)

[Previous 200](#)

[Next 200](#)

[Refresh index](#)

AND [Show Index list](#)

or [Add to history](#)

Items from search history can be included / combined / modified

History

[Clear history](#)

Search	Add to builder	Query	Items found	Time
#7	Add	Search zebrafish[organism] AND actin[title]	71	12:41:16
#4	Add	Search zebrafish actin	1288	12:40:07
#1	Add	Search zebrafish creatine kinase	34	12:39:02

Complex Query Results

Display Settings: (v) Summary, 20 per page, Sorted by Default order

Send to: (v) Filter your results:

Results: 6

- [Danio rerio creatine kinase, brain a \(ckba\), mRNA](#)
 - 1,481 bp linear mRNA
Accession: NM_001077163.1 GI: 116004536
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- [Danio rerio creatine kinase, mitochondrial 1 \(ckmt1\), nuclear gene encoding mitochondrial protein, mRNA](#)
 - 2
- [Danio rerio creatine kinase, muscle a \(ckma\), mRNA](#)
 3. 1,552 bp linear mRNA
Accession: NM_130932.1 GI: 18858426
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- [Danio rerio creatine kinase, mitochondrial 2 \(sarcomeric\) \(ckmt2\), nuclear gene encoding mitochondrial protein, mRNA](#)
 4. 1,401 bp linear mRNA
Accession: NM_200697.1 GI: 41152341
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- [Danio rerio creatine kinase, muscle b \(ckmb\), mRNA](#)
 5. 1,463 bp linear mRNA
Accession: NM_001105683.1 GI: 157787180
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)
- [Danio rerio creatine kinase, brain b \(ckbb\), mRNA](#)
 6. 1,459 bp linear mRNA
Accession: NM_173222.1 GI: 27545192
[GenBank](#) [FASTA](#) [Graphics](#) [Related Sequences](#)

Filter your results:

- All (6)
- Bacteria (0)
- INSDC (GenBank) (0)
- [mRNA \(6\)](#)
- [RefSeq \(6\)](#)

[Manage Filters](#)

Analyze these sequences

Run BLAST

Find related data

Database:

Search details

```
("Danio rerio"[Organism] AND "creatine kinase"[Title]) AND "refseq"[Filter]
```

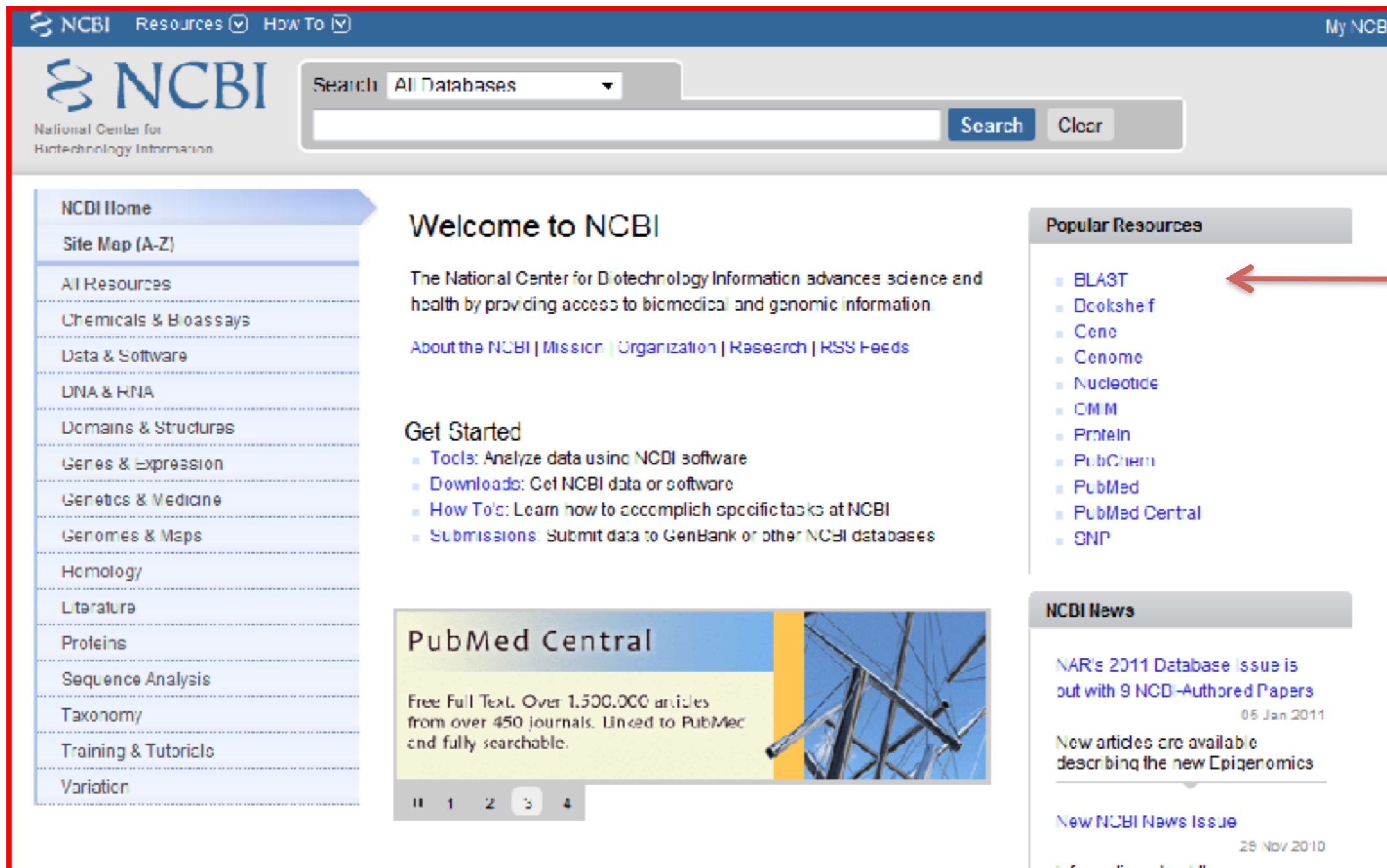
[See more...](#)

Recent activity

("Danio rerio"[Organism] AND "creatine kinase"[Title]) AND "refseq"[Filter] AND mrna[Filter]

BLAST is a very important tool available from the NCBI Homepage

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



The screenshot shows the NCBI homepage with a search bar at the top. The search bar contains the text "Search All Databases" and has "Search" and "Clear" buttons. Below the search bar is a navigation menu with "NCBI Home" selected. The main content area is titled "Welcome to NCBI" and includes a description of the center's mission and a "Get Started" section with links to Tools, Downloads, How To's, and Submissions. A "PubMed Central" banner is also visible. On the right side, there is a "Popular Resources" section with a list of links, including "BLAST", "Bookshelf", "Gene", "Genome", "Nucleotide", "OMIM", "Protein", "PubChem", "PubMed", "PubMed Central", and "SNP". A red arrow points to the "BLAST" link. Below this is an "NCBI News" section with two news items: "NAR's 2011 Database Issue is out with 9 NCI-Authored Papers" (dated 05 Jan 2011) and "New articles are available describing the new Epigenomics".

NCBI Resources How To My NCBI

NCBI
National Center for
Biotechnology Information

Search All Databases Search Clear

NCBI Home
Site Map (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 advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

PubMed Central

Free Full Text. Over 1,500,000 articles from over 450 journals. Linked to PubMed and fully searchable.

1 2 3 4

Popular Resources

- BLAST
- Bookshelf
- Gene
- Genome
- Nucleotide
- OMIM
- Protein
- PubChem
- PubMed
- PubMed Central
- SNP

NCBI News

NAR's 2011 Database Issue is out with 9 NCI-Authored Papers
05 Jan 2011

New articles are available describing the new Epigenomics

New NCBI News Issue
28 Nov 2010

BLAST – Basic Local Alignment Search Tool

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

The screenshot shows the NCBI BLAST website interface. At the top, there is a navigation bar with links for Home, Recent Results, Saved Strategies, and Help. Below this, the main content area is divided into several sections. On the left, there is a section for 'BLAST Assembled RefSeq Genomes' with a list of species including Human, Mouse, Rat, Arabidopsis thaliana, Oryza sativa, Bos taurus, Danio rerio, Drosophila melanogaster, Gallus gallus, Pan troglodytes, Microbes, and Apis mellifera. Below this is the 'Basic BLAST' section, which lists various BLAST programs such as nucleotide blast, protein blast, blastx, tblastn, and tblastx, each with a brief description and a list of algorithms. On the right side, there is a 'News' section with a link to 'New WGS BLAST page' and a 'Tip of the Day' section with a link to 'How to do Batch BLAST jobs'. The overall layout is clean and organized, with a blue header and a white background for the main content.

BLAST performs sequence similarity searches of query sequences vs sequence databases. We will cover this in detail in the next lecture.

NCBI Metadatabases

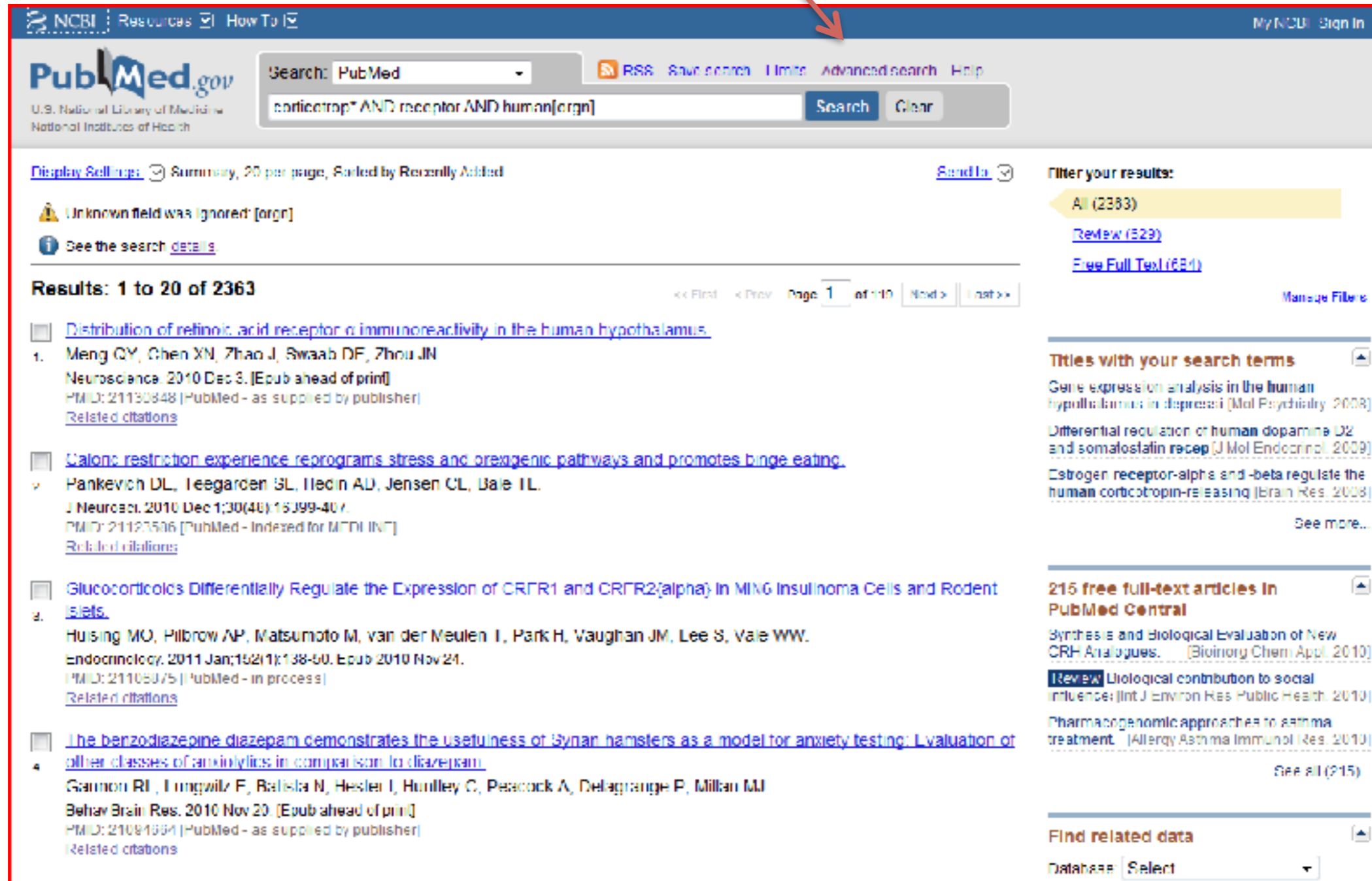
- **Gene**
 - ▶ molecular data and literature related to genes
- **HomoloGene**
 - ▶ automated collection of homologous genes from selected eukaryotes
- **Taxonomy**
 - ▶ access to NCBI data through source organism taxonomic classification
- **PubChem**
 - ▶ small organic molecules and their biological activities
- **BioSystems**
 - ▶ biochemical pathways and processes linked to NCBI genes, gene products, small molecules, and structures

PubMed

- Curated database of biomedical journal articles
- Data records are annotated with MeSH terms (Medical Subject Headings)
- Contract workers actually read all of the articles and classify them with the MeSH terms
- PubMed entries contain article abstracts
- PubMed Central contains full journal articles, but the majority are not freely re-distributable

PubMed results

Limits and Advanced search can be used to refine searches



The screenshot displays the PubMed.gov search interface. At the top, the search bar contains the query "corticotrop* AND receptor AND human[orgn]". The search results page shows 2363 results. The first four results are listed, each with a checkbox, a title, authors, journal information, and PMID. A red arrow points to the "Limits" link in the top navigation bar, which is used to refine search results.

NCBI Resources How To MyNCBI Sign In

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

Search: PubMed
RSS Save search Limits Advanced search Help

Search Clear

Display Settings Summary, 20 per page, Sorted by Recently Added Send To

Unknown field was ignored: [orgn]
See the search details

Results: 1 to 20 of 2363
Page 1 of 110

1. [Distribution of retinoic acid receptor \$\alpha\$ immunoreactivity in the human hypothalamus.](#)
Meng QY, Chen XN, Zhao J, Swaab DF, Zhou JN
Neuroscience. 2010 Dec 3. [Epub ahead of print]
PMID: 21130848 [PubMed - as supplied by publisher]
[Related citations](#)

2. [Caloric restriction experience reprograms stress and orexigenic pathways and promotes binge eating.](#)
Pankovich DL, Teegarden SL, Hedin AD, Jensen CL, Bale TL.
J Neurosci. 2010 Dec 1;30(48):16399-407.
PMID: 21127506 [PubMed - Indexed for MEDLINE]
[Related citations](#)

3. [Glucocorticoids Differentially Regulate the Expression of CRFR1 and CRFR2\(alpha\) in MIN6 Insulinoma Cells and Rodent Islets.](#)
Huisling MO, Pilbrow AP, Matsumoto M, van der Meulen T, Park H, Vaughan JM, Lee S, Vale WW.
Endocrinology. 2011 Jan;152(1):138-50. Epub 2010 Nov 24.
PMID: 21105075 [PubMed - in process]
[Related citations](#)

4. [The benzodiazepine diazepam demonstrates the usefulness of Syrian hamsters as a model for anxiety testing: Evaluation of other classes of anxiolytics in comparison to diazepam.](#)
Gannon RI, Jungwilt F, Balista N, Hessler J, Humley C, Pascock A, Delagrange P, Millan MJ
Behav Brain Res. 2010 Nov 20. [Epub ahead of print]
PMID: 21091661 [PubMed - as supplied by publisher]
[Related citations](#)

Filter your results:
All (2363)
[Review \(523\)](#)
[Free Full Text \(681\)](#)
Manage Filters

Titles with your search terms
Gene expression analysis in the human hypothalamus in depression [Mol Psychiatry. 2008]
Differential regulation of human dopamine D2 and somatostatin receptor [J Mol Endocrinol. 2009]
Estrogen receptor-alpha and -beta regulate the human corticotropin-releasing [Brain Res. 2008]
See more...

215 free full-text articles in PubMed Central
Synthesis and Biological Evaluation of New CRH Analogues. [Bioinorg Chem Appl. 2010]
[Review](#) Biological contribution to social influence [Int J Environ Res Public Health. 2010]
Pharmacogenomic approaches to asthma treatment. [Allergy Asthma Immunol Res. 2010]
See all (215)

Find related data
Database Select

Small molecule databases have been added at NCBI

<http://pubchem.ncbi.nlm.nih.gov/>

The screenshot displays the PubChem website interface. At the top, there is a navigation bar with dropdown menus for "Databases", "Deposition", "Services", and "Help more". The main header features the "PubChem" logo. Below the logo are three buttons: "BioAssay", "Compound", and "Substance", each with a corresponding icon. A search bar is located below these buttons, with an "Advanced search" button to its right. A central banner reads "Chemical structure search | BioActivity analysis". A news box below the banner states: "New More than 2.5 million structures from the IBM BAO (Business Analytics and Optimization) strategic IP insight platform (SIIP) are now available in PubChem. See more.. and related news." To the right of the news box is a "more ..." link with a RSS icon. On the far right, a vertical sidebar contains various tools and services, each with an icon: "Bioactivity summary", "Bioactivity datatable", "Bioactivity structure-activity", "Chemical structure search", "3D conformer viewer", "Chemical structure clustering", "Deposition gateway", "Structure download", "Bioassay download", and "PubChem FTP".

Databases ▾ Deposition Services ▾ Help more ▾

PubChem

BioAssay ? Compound ? Substance ?

 Advanced search

[Chemical structure search](#) | [BioActivity analysis](#)

New More than 2.5 million structures from the **IBM BAO** (Business Analytics and Optimization) strategic IP insight platform (SIIP) are now available in PubChem. See [more..](#) and related news.

[more ...](#)

[Write to Helpdesk](#) | [Disclaimer](#) | [Privacy Statement](#) | [Accessibility](#) | [Data Citation Guidelines](#)
National Center for Biotechnology Information
NLM | NIH | HHS

- Bioactivity summary
- Bioactivity datatable
- Bioactivity structure-activity
- Chemical structure search
- 3D conformer viewer
- Chemical structure clustering
- Deposition gateway
- Structure download
- Bioassay download
- PubChem FTP

HomoloGene - Homologous genes from different organisms <http://www.ncbi.nlm.nih.gov/homologene>

The screenshot shows the NCBI HomoloGene website. At the top, there is a search bar with "HomoloGene" selected and a "Go" button. Below the search bar are navigation tabs for "Limits", "Preview/Index", "History", "Clipboard", and "Details". The main content area features a description of HomoloGene, a table of statistics for Release 65, and sections for "What's New" and "Related Resources".

HomoloGene Release 65 Statistics

Initial numbers of genes from complete genomes, numbers of genes placed in a homology group, and the numbers of groups for each species.

Species	Number of Genes		HomoloGene groups
	Input	Grouped	
Homo sapiens	19,943*	18,981	18,431
Pan troglodytes	25,096	16,050	15,980
Canis familiaris	19,766	16,708	15,951
Bos taurus	22,049	18,180	16,224
Mus musculus	25,388	21,766	19,005
Rattus norvegicus	21,991	19,229	17,473
Gallus gallus	17,959	13,142	11,905
Danio rerio	25,690*	21,084	14,067
Drosophila melanogaster	13,027*	9,282	7,749
Anopheles gambiae	12,460	8,867	7,541
Caenorhabditis elegans	20,132*	8,678	4,810
Schizosaccharomyces pombe	5,043	3,226	2,936
Saccharomyces cerevisiae	5,880	4,861	4,370
Kluyveromyces lactis	5,335	4,459	4,382
Eremothecium gossypii	4,722	3,928	3,884
Magnaporthe grisea	12,832	7,330	6,399
Neurospora crassa	9,821*	6,287	6,144
Arabidopsis thaliana	27,000*	19,961	11,242

What's New

HomoloGene release 65 includes updated annotations for the following species: Homo sapiens (NCBI release 37.2), Danio rerio (NCBI release 4.1), Drosophila melanogaster (NCBI release 9.3), Caenorhabditis elegans (NCBI release 9.1), Arabidopsis thaliana (NCBI release 9.1).

Related Resources

Entrez Genomes
A collection of complete genome sequences that includes more than 1000 viruses and over hundred microbes

- Archaea
- Bacteria
- Eukaryota

Online Mendelian Inheritance in Man – OMIM

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

The screenshot displays the OMIM website interface. At the top, the NCBI logo is on the left, and the OMIM logo with the text "Online Mendelian Inheritance in Man" and "Johns Hopkins University" is on the right. A navigation bar includes links for "All Databases", "PubMed", "Nucleotide", "Protein", "Genome", "Structure", "PMC", and "OMIM". A search bar contains the text "Search OMIM" and "for" followed by a dropdown arrow, with "Go" and "Clear" buttons. Below the search bar are buttons for "Limits", "Preview/Index", "History", "Clipboard", and "Details".

On the left side, there is a sidebar with the following sections:

- Entrez**
- OMIM**
 - Search OMIM
 - Search Gene Map
 - Search Morbid Map
- Help**
 - OMIM Help
 - How to Link
- FAQ**
 - Numbering System
 - Symbols
 - How to Print
 - Citing OMIM
 - Download
- OMIM Facts**
 - Statistics
 - Update Log

The main content area features a list of instructions:

- Enter one or more search terms.
- Use **Limits** to restrict your search by search field, chromosome, and other criteria
- Use **Index** to browse terms found in OMIM records
- Use **History** to retrieve records from previous searches, or to combine searches.

Below this is a purple header for "OMIM® - Online Mendelian Inheritance in Man".

The main text reads: "Welcome to OMIM®, Online Mendelian Inheritance in Man®. OMIM is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes. The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 12,000 genes. OMIM focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources."

A second paragraph states: "This database was initiated in the early 1960s by Dr. Victor A. McKusick as a catalog of mendelian traits and disorders, entitled Mendelian Inheritance in Man (MIM). Twelve book editions of MIM were published between 1966 and 1998. The online version, OMIM, was created in 1985 by a collaboration between the National Library of Medicine and the William H. Welch Medical Library at Johns Hopkins. It was made generally available on the

OMIM is essentially a set of reviews of human genes, gene function and phenotypes. Includes causative mutations where known.

The NCBI Bookshelf includes many well known molecular biology texts.

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

The screenshot shows the NCBI Bookshelf website. At the top left is the NCBI logo. The main header features the word "Bookshelf" in a large, bold, red font. Below this is a navigation bar with tabs for "All Databases", "PubMed", "Nucleotide", "Protein", "Genome", "Structure", "PMC", and "Taxonomy". A search bar is located below the navigation bar, with the text "Search Books" and a dropdown menu set to "for". There are "Go" and "Clear" buttons next to the search bar. Below the search bar are several tabs: "Limits", "Preview/Index", "History", "Clipboard", and "Details". On the left side, there is a blue sidebar with links for "Introduction", "Quick Start Guide", "Help", "Information for Authors and Publishers", "What's New", "FAQ", "My NCBI", and "Privacy Policy". The main content area is enclosed in a red border and contains the following text: "The Bookshelf is a growing collection of biomedical books that can be searched directly by typing a concept into the textbox above and selecting 'Go'. Try one of these searches: cell cycle control, immunodeficiency, protein evolution." Below this is a section titled "New on the Bookshelf:" with a list of books, each with a small thumbnail image and a title: "Health United States, 2009" (National Center for Health Statistics, 2010), "Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis" (Cambridge University Press, 2007), "Probe Reports from the Molecular Libraries Program" (National Center for Biotechnology Information, 2010), "StemBook" (Harvard Stem Cell Institute, 2008-), and "VA Evidence-based Synthesis Program Reports" (Department of Veterans Affairs, 2007-).

GEO: Gene Expression Omnibus

- Gene expression data (mostly from microarrays but also RNA-seq data, 2 methods for measuring RNA levels)

Query browse and download data sets

The screenshot displays the GEO website interface. At the top, there is the NCBI logo on the left and the GEO logo (Gene Expression Omnibus) on the right. Below the logos are navigation links: HOME, SEARCH, SITE MAP, GEO Publications, FAQ, MIAME, and Email GEO. A status bar indicates 'Not logged in | Login'. The main content area features a description of GEO as a public functional genomics data repository. Below this, there are two main sections: 'GEO navigation' and 'Site contents'. The 'GEO navigation' section is divided into 'QUERY' and 'BROWSE' categories. The 'QUERY' category includes 'DataSets', 'Gene profiles', 'GEO accession', and 'GEO BLAST', each with a search input field and a 'GO' button. The 'BROWSE' category includes 'DataSets' and 'GEO accessions', with 'GEO accessions' further subdivided into 'Platforms', 'Samples', and 'Series'. The 'Site contents' section on the right lists 'Public data' (Platforms: 8,246, Samples: 514,893, Series: 20,827), 'Documentation' (Overview, FAQ, Find, Submission guide, Linking & citing, Journal citations, Construct a Query, Programmatic access, DataSet clusters, GEO announce list, Data disclaimer, GEO staff), and 'Query & Browse' (Repository browser, Submitters). A 'Submitter login' link is located at the bottom left of the main content area.

- **Series** - (GSExxx) is an original submitter-supplied record that summarizes a study. May contain multiple individual **Samples** (GSMxxx).

Platform(s) (1) [GPL4091](#) Agilent-014693 Human Genome CGH Microarray 244A (Feature number version)

Samples (4) [GSM495808](#) Aspc1 Cell Line
[GSM495809](#) JH39 Xenograft
[GSM495810](#) JH21 Xenograft

Download family

Download family	Format
SOFT formatted family file(s)	SOFT ?
MINiML formatted family file(s)	MINiML ?
Series Matrix File(s)	TXT ?

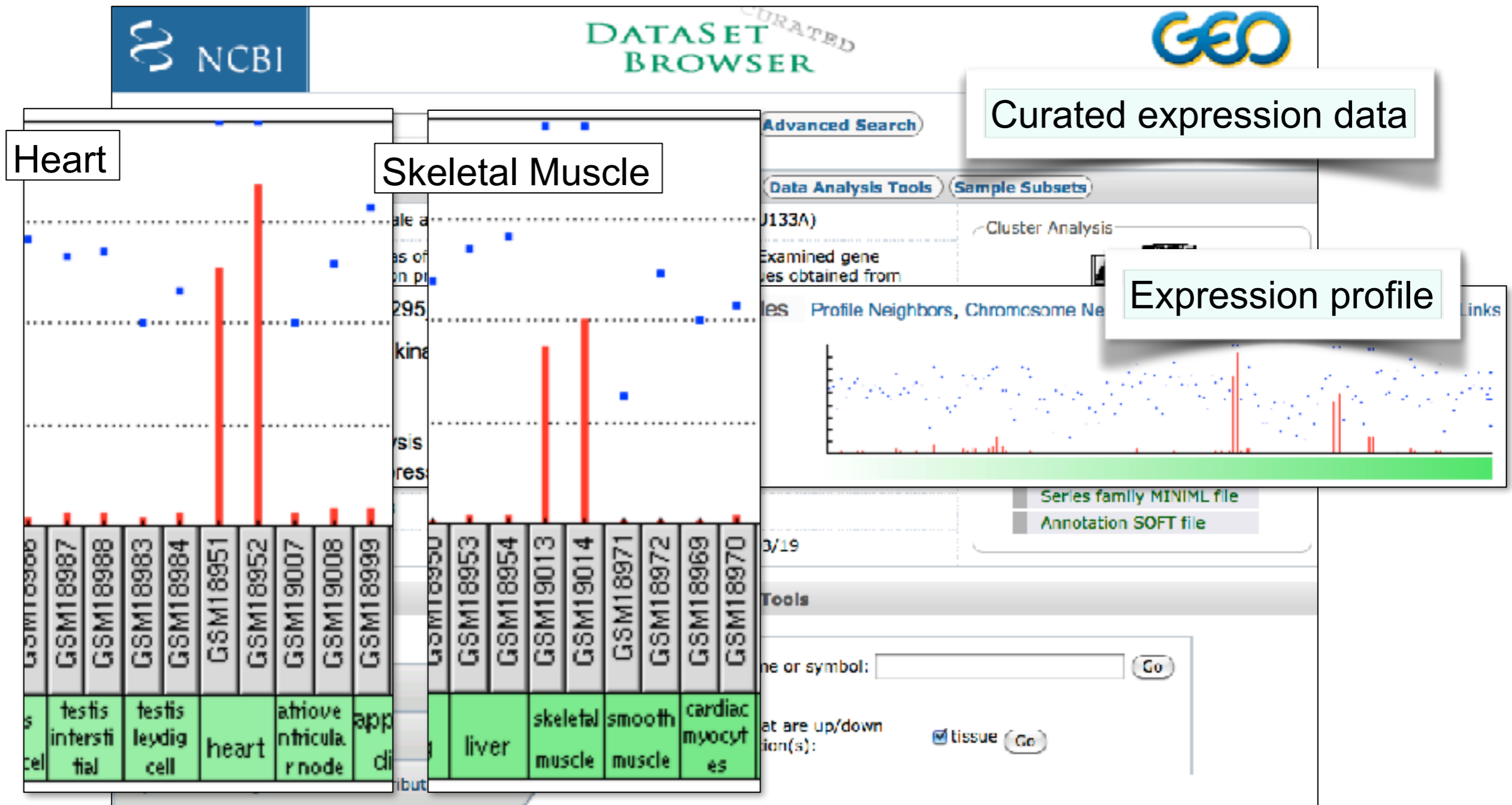
Supplementary file	Size	Download	File type/resource
GSE19852_RAW.tar	170.7 Mb	(ftp) (http)	TAR (of TXT)

Raw data provided as supplementary file
 Processed data included within Sample Table

NLM | NIH | [GEO Help](#) | [Disclaimer](#) | [Section 508](#)

uter. Significance Analysis of Microarrays (SAM) identified 92 genes differentially expressed by strain. Interestingly, several members of the solute carrier family of amino acid transporters, genes involved in amino acid synthesis and development, and amiloride-sensitive epithelial sodium channel gene were induced by strain. These results were confirmed by quantitative real-time polymerase chain reaction (qRT-PCR). Thus, this study identifies genes induced by strain that may be important for amino acid signaling pathways, protein

- **DataSets** - (GDSxxx) are curated collections of selected Samples that are biologically and statistically comparable



QuickGO is a fast web-based browser of the Gene Ontology and Gene Ontology annotation data

The screenshot shows the QuickGO web interface in a browser window. The address bar displays www.ebi.ac.uk/QuickGO/. The page features a search bar with the placeholder text "Enter Text Here" and a "Find" button. A navigation menu includes "Databases", "Tools", "Research", "Training", "Industry", "About Us", and "Help". A sidebar on the left lists "QuickGO", "Help", "Reference", "FAQs", "Video tutorials", "Downloads", "geneontology.org", "UniProt-GOA project", and "Web Services". The main content area includes a "QuickGO" heading, a description: "QuickGO is a fast web-based browser for [Gene Ontology](#) terms and annotations, which is provided by the [UniProt-GOA project](#) at the [EBI](#).", and a search bar with the placeholder "Click for example search" and a "Search!" button. Below the search bar are icons for "Web Services", "Dataset", and "Term Basket: 0". Three informational panels are visible: "Search and Filter GO annotation sets" (Extensive filters are available from this page to allow the generation of specific subsets of GO annotations, mapped to sequence identifiers of your choice.), "Investigate GO slims" (GO slims are lists of GO terms that have been selected from the full set of terms available from the Gene Ontology project. GO slims can be used to generate a focused view of part of the GO, or with annotation data they can be used to see how a set of proteins/genes can be broadly categorized (using annotation data and the relationships that exist between terms in the ontologies). Further information on GO slims can be found at the [GO Consortium web site](#).), and "View the history of changes to GO" (This page allows you to view the changes to GO, optionally filtered by date, term identifier, or type of change.). On the right side, there are sections for "QuickGO News" (19 August 2011 - Changes to the Term Basket, 14 June 2011 - New term history displays, 20 April 2011 - Display improvements) and "QuickGO Tips" (Tutorial).

GO annotation in UniProt

An example UniProt entry for hemoglobin beta (HBB_human, P68871) with GO annotation displayed.

The screenshot shows the UniProt entry for Hemoglobin subunit beta (HBB_human, P68871). The page is titled "Hemoglobin subunit beta - Homo sapiens (Human)" and displays the Gene Ontology (GO) annotations. The annotations are organized into three main categories: Biological process, Cellular component, and Molecular function. Each category lists specific GO terms with their corresponding sources and evidence codes.

Category	GO Term	Source	Evidence
Biological process	bicarbonate transport	Reactome	Traceable author statement
	blood coagulation	Reactome	Traceable author statement
	hydrogen peroxide catabolic process	BHF-UCL	Inferred from direct assay (PubMed 13740759)
	nitric oxide transport	UniProtKB	Non-traceable author statement (PubMed 8292032)
	positive regulation of cell death	BHF-UCL	Inferred from direct assay (PubMed 13740759)
	positive regulation of nitric oxide biosynthetic process	UniProtKB	Non-traceable author statement (PubMed 7565123)
	protein heterooligomerization	BHF-UCL	Inferred from direct assay (PubMed 13740759)
	regulation of blood pressure	UniProtKB-KW	Inferred from electronic annotation
	regulation of blood vessel size	UniProtKB-KW	Inferred from electronic annotation
	renal absorption	UniProtKB	Inferred from mutant phenotype (PubMed 15465053, PubMed 10374555)
Cellular component	endocytic vesicle lumen	Reactome	Traceable author statement
	extracellular region	Reactome	Traceable author statement
	haptoglobin-hemoglobin complex	BHF-UCL	Inferred from direct assay (PubMed 13740759)
	hemoglobin complex	UniProtKB	Non-traceable author statement (Ref.33, Ref.72)
Molecular function	heme binding	InterPro	Inferred from electronic annotation

GO annotation in UniProt

An example UniProt entry for hemoglobin beta (HBB_human, P68871) with GO annotation displayed.

The screenshot displays a web browser window showing the UniProt entry for Hemoglobin subunit beta (HBB_human, P68871) with its GO annotation. The browser address bar shows the URL: www.uniprot.org/uniprot/P68871. The page title is "Hemoglobin subunit beta - Homo sapiens (Human)".

The main content area shows the QuickGO interface for the GO term "GO:0020037 heme binding". The QuickGO logo and tagline "A fast browser for Gene Ontology terms and annotations." are visible. The page includes a search bar and navigation links for "Web Services", "Dataset", and "Term Basket: 0".

The "Term Information" tab is selected, displaying the following details:

- ID:** GO:0020037
- Name:** heme binding
- Ontology:** Molecular Function
- Definition:** Interacting selectively and non-covalently with heme, any compound of iron complexed in a porphyrin (tetrapyrrole) ring.
- GONUTS:** GO:0020037 Wiki Page

Below the term information, there are tabs for "Synonyms", "Annotation Guidance", "Cross-Ontology Relations", and "Cross-references". The "Synonyms" tab is active, showing a table of synonyms:

Type	Synonym
exact	haem binding

Additional text explains that synonyms are alternative words or phrases closely related in meaning to the term name, with indication of the relationship between the name and synonym given by the synonym scope. Click on the icon for more details.

At the bottom of the page, there is a footer with the text: "Please send comments, suggestions or bug reports to goa@ebi.ac.uk. Click here for details of how to cite UniProt-GOA and QuickGO." The page number "11ms" is visible in the bottom right corner.

DAVID: a online tool for assessing GO term enrichment in gene lists

The screenshot shows the DAVID Bioinformatics Resources 6.7 website. The browser address bar displays david.abcc.ncifcrf.gov/home.jsp. The page header includes the DAVID logo and the text "DAVID Bioinformatics Resources 6.7 National Institute of Allergy and Infectious Diseases (NIAID), NIH". A navigation menu contains links for Home, Start Analysis, Shortcut to DAVID Tools, Technical Center, Downloads & APIs, Term of Service, Why DAVID?, and About Us. A dropdown menu under "Shortcut to DAVID Tools" lists: Functional Annotation (with sub-links for Clustering, Chart, and Table), Gene Functional Classification, Gene ID Conversion, Gene Name Batch Viewer, and NIAID Pathogen Annotation Browser. The main content area features a search bar, a "What's Important in DAVID?" section with links to release notes and requirements, and a list of features with checkboxes: Identify enriched terms, Discover enriched terms, Cluster redundant terms, Visualize genes, Display related terms, Search for other functionally related genes not in the list, List interacting proteins, and Explore gene names in batch. A text box in the foreground states: "DAVID allows you to upload lists of genes and search for enriched GO and search for functionally related genes not in your list" with the URL <http://david.abcc.ncifcrf.gov>.

Example output: enriched functions from GO

DAVID: Database for Annotation, Visualization, and Integrat...ID); Science Applications International Corporation (SAIC)

david.abcc.ncifcrf.gov/chartReport.jsp?annot=25

DAVID: Functional Annotation Result Summary

DAVID Bioinformatics Resources 6.7
National Institute of Allergy and Infectious Diseases (NIAID), NIH

Functional Annotation Chart

[Help and Manual](#)

Current Gene List: List_1
Current Background: Homo sapiens
14 DAVID IDs

Options

Rerun Using Options Create Sublist

10 chart records [Download File](#)

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of osteoclast differentiation	RT		2	14.3	2.1E-2	9.9E-1
<input type="checkbox"/>	GOTERM_BP_FAT	response to organic substance	RT		4	28.6	2.9E-2	9.6E-1
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of myeloid leukocyte differentiation	RT		2	14.3	3.9E-2	9.5E-1
<input type="checkbox"/>	GOTERM_BP_FAT	positive regulation of transcription from RNA polymerase II promoter	RT		3	21.4	4.8E-2	9.4E-1
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of myeloid cell differentiation	RT		2	14.3	6.5E-2	9.5E-1
<input type="checkbox"/>	GOTERM_BP_FAT	cartilage development	RT		2	14.3	6.9E-2	9.3E-1
<input type="checkbox"/>	GOTERM_BP_FAT	positive regulation of transcription, DNA-dependent	RT		3	21.4	7.5E-2	9.2E-1
<input type="checkbox"/>	GOTERM_BP_FAT	positive regulation of RNA metabolic process	RT		3	21.4	7.6E-2	8.9E-1
<input type="checkbox"/>	GOTERM_BP_FAT	response to protein stimulus	RT		2	14.3	9.8E-2	9.3E-1
<input type="checkbox"/>	GOTERM_BP_FAT	positive regulation of transcription	RT		3	21.4	1.0E-1	9.1E-1

8 gene(s) from your list are not in the output.

Please cite *Nature Protocols* 2009; 4(1):44 & *Genome Biology* 2003; 4(5):P3 within any publication that makes use of any methods inspired by DAVID.