

# INTRODUCTION TO BIOINFORMATICS

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

Barry Grant  
University of Michigan  
[www.thegrantlab.org](http://www.thegrantlab.org)

BIOINF 525 [http://bioboot.github.io/bioinf525\\_w17/](http://bioboot.github.io/bioinf525_w17/) 10-Jan-2017



Barry Grant, Ph.D.  
[bjgrant@umich.edu](mailto:bjgrant@umich.edu)



Ryan Mills, Ph.D.  
[remills@umich.edu](mailto:remills@umich.edu)



Lauren Jepsen (GSI)  
[ljepsen@umich.edu](mailto:ljepsen@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."

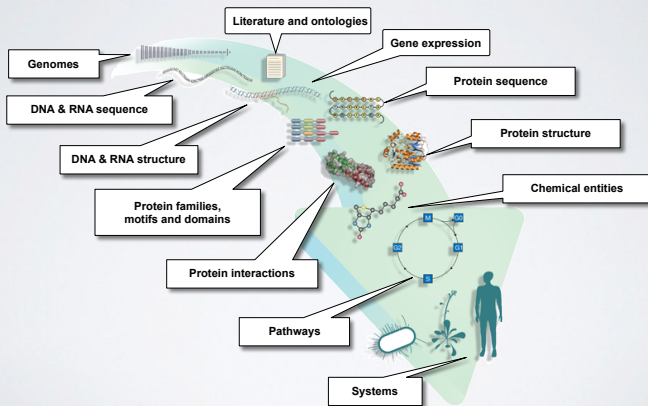
- ... Bioinformatics is a hybrid of biology and computer science
- ... **Bioinformatics is computer aided biology!**

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

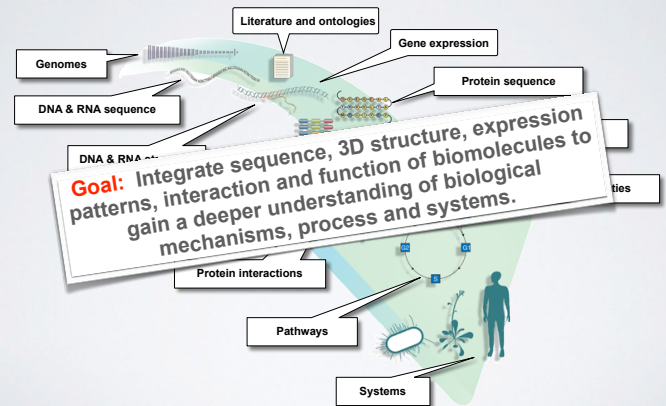
## MORE DEFINITIONS

- ▶ "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/l3gx6b> )

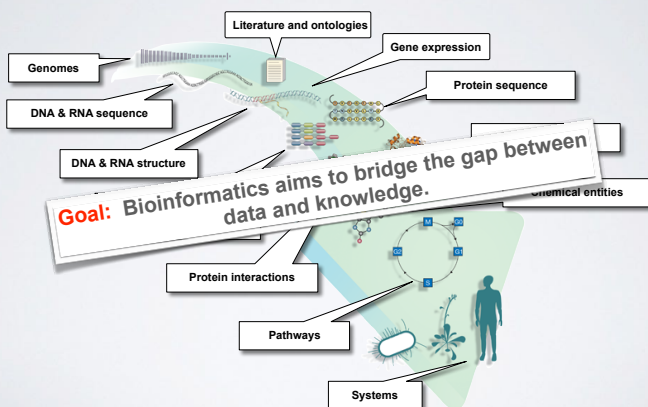
## Major types of Bioinformatics Data



## 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

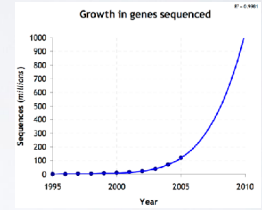
- DNA sequence determines protein sequence.
- Protein sequence determines protein structure.
- Protein structure determines protein function.
- Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.

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

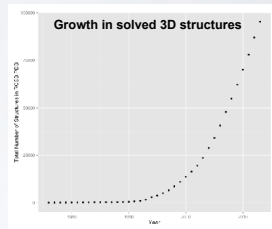


E.G. data from sequencing, structural genomics, microarrays, proteomics, new high throughput assays, etc...

## 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)

Protein BLAST: search protein databases using a protein query

blast.ncbi.nlm.nih.gov/blast.cgi?PROGRAM=blastp&BLAST\_PROGRAMS=blastp&PAGE\_TYPE=blastSearch&SHOW\_DEFAULTS=on&LINK\_LOC=blasthome

**General Parameters**

- Max target sequences: 500
- Short queries:  Automatically adjust parameters for short input sequences
- Expect threshold: 10
- Word size: 3
- Max matches in a query range: 0

**Scoring Parameters**

- Matrix: BLOSUM62
- Gap Costs: Existence: 11 Extension: 1

**Filters and Masking**

- Filter:  Low complexity regions
- Mask:  Mask for lookup table only

**PSI/PHIDELTA BLAST**

- Upload PSSM: Choose File
- PSI-BLAST Threshold: 0.005
- Pseudocount: 0

**Related tools with different terminology**

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

Even Blast has many settable parameters

# Key Online Bioinformatics Resources: NCBI & EBI

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

NCBI National Center for Biotechnology Information

Welcome to NCBI

Popular Resources

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

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

The European Bioinformatics Institute

EBI provides freely available tools, data and services to support research in genomics, bioinformatics and related fields.

<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

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.

Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- HowTo: 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 chemical, and associated biosystems.

NCBI Announcements

New version of Genome Workbench available

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

National Center for Biotechnology Information

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 Biotech and health by providing access information.

Get Started

- Tools: Analyze data using
- Downloads: Get NCBI data
- HowTo: Learn how to acc
- Submissions: Submit data 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 chemical, and associated biosystems.

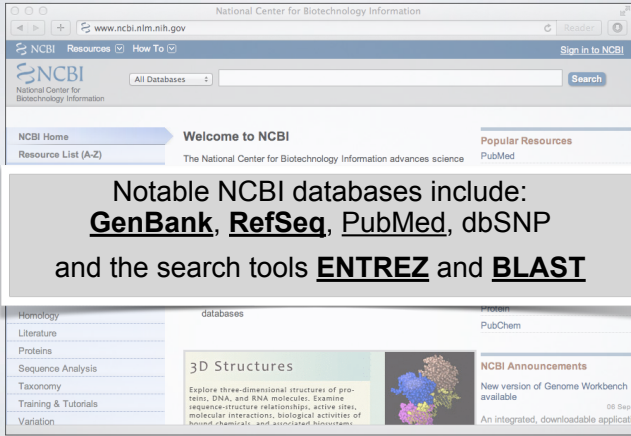
NCBI Announcements

New version of Genome Workbench available

**Popular Resources**

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

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



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

Homology	databases	Protein
Literature		PubChem
Proteins		
Sequence Analysis		
Taxonomy		
Training & Tutorials		
Variation		

3D Structures  
Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemical, and associated biosystems.

NCBI Announcements  
New version of Genome Workbench available  
An integrated, downloadable applica...

## Key Online Bioinformatics Resources: NCBI & EBI

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



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



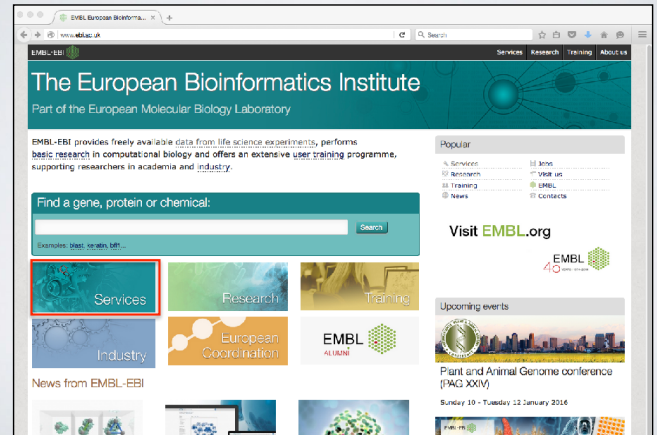
<https://www.ebi.ac.uk>

## 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 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:

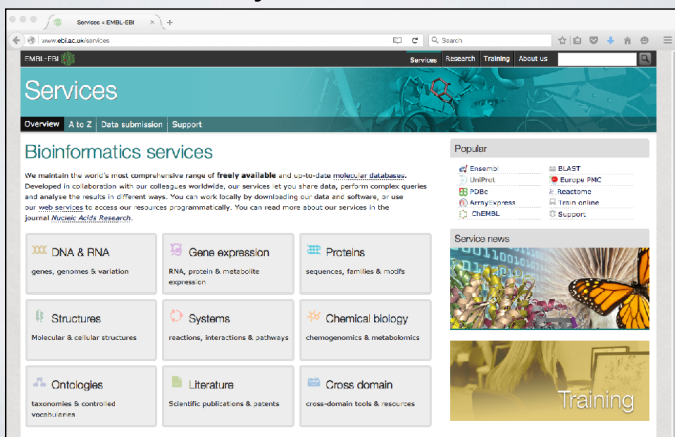
Services Research Training

Industry European Coordination EMBL ALIANCE

News from EMBL-EBI

Upcoming events  
Plant and Animal Genome conference (PAG XXIV)  
Sunday 19 - Tuesday 12 January 2016

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



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 of Molecular Biology Research](#).

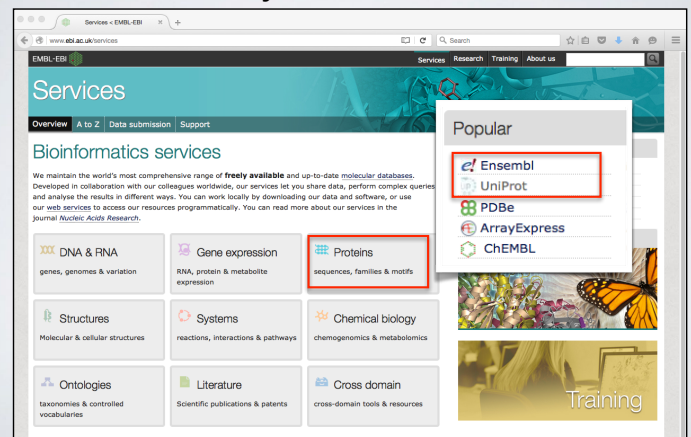
Popular

- Ensembl
- UniProt
- PDB
- ArrayExpress
- CHEMBL
- BLAST
- Europe PMC
- Reactome
- UniProt
- Support

Services news

Training

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



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 of Molecular Biology Research](#).

Popular

- Ensembl
- UniProt
- PDB
- ArrayExpress
- CHEMBL

Services news

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 co-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 generate 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

**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:

Services, Research, Training, Industry, European Coordination, EMBL ALIANCE

Visit EMBL.org

Upcoming events: Plant and Animal Genome conference (PAG XXIV) Sunday 19 - Tuesday 22 January 2016

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

**Train online**

Using sequence similarity searching tools at EMBL-EBI: webinar

Course content

Using sequence similarity searching tools at EMBL-EBI: webinar

Popular

- Train online
- Training
- Find us at...
- Open days and career fairs
- Conference notifications
- EMBL events and reports
- Gene expression events
- Balance for health

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

**Train online | EBI Train online**

Navigation: Databases, Tools, Research, Training, Industry, About Us, Help

Find a course

Browse by subject

- Genes and Genomes
- Gene Expression
- Interactions, Pathways and Networks

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

# 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, Bearnf, BiomagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBYD, 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, Genillesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDb, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIBD, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMG, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, MycDB, NDB, NRSdb, 0-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, BCGD, Bearnf, BiomagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBYD, 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, Genillesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDb, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIBD, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMG, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, MycDB, NDB, NRSdb, 0-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 .....!!!!

There are lots of Bioinformatics Databases  
For an 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

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

## 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](#))

The 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](#))

The 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](#))

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.

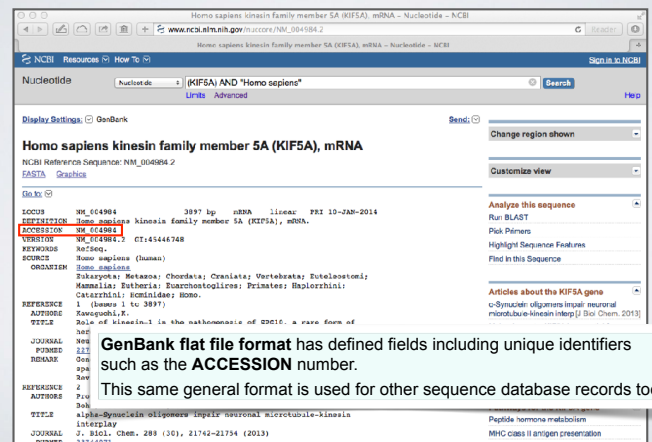
- **Primary databases** (or *archival databases*) consist of data derived experimentally.
  - ▶ **GenBank**: NCBI's primary nucleotide sequence database.
  - ▶ **PDB**: Protein X-ray crystal and NMR structures.
- **Secondary databases** (or *derived databases*) contain information derived from a primary 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



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





# GenBank sequence record

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

NCBI Reference Sequence: NM\_004984.2

FASTA

Can set different display formats here

ICDCU NM\_004984.2 Homo sapiens kinesin family member 5A (KIF5A), mRNA

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

ACCESSION NM\_004984

VERSION NM\_004984.2 GI:45446748

KEYWORDS mRNA.

SOURCE Homo sapiens (human)

ORGANISM Homo sapiens

RAJAYAKA; METAZOA; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Chiroptera; Hominoidea; Homo.

REFERENCE

1 (bases 1 to 3897)

TITLE Role of kinesin-1 in the pathogenesis of DP101, a rare form of hereditary spastic paraplegia

PMID 2171102

PMID 2171102 A review of the mechanism of pathogenesis involved in spastic paraplegia type 10 when KIF5A is inactivated by mutations. Review article

REMARKS

2 (bases 1 to 3897)

1. Puvion, J., Nègre, J.F., Baye, F., Campion, J., Dubet, K., Riab, R., Jahn, G., and Nègre, B. alpha-Dynactin oligomers impede neuronal microtubule-kinesin transport. *J. Biol. Chem.* 289 (30), 21742-21754 (2013)

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

NCBI Reference Sequence: NM\_004984.2

FASTA

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.

ICDCU NM\_004984.2 Homo sapiens kinesin family member 5A (KIF5A), mRNA

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

ACCESSION NM\_004984

VERSION NM\_004984.2 GI:45446748

KEYWORDS mRNA.

SOURCE Homo sapiens (human)

ORGANISM Homo sapiens

RAJAYAKA; METAZOA; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Chiroptera; Hominoidea; Homo.

REFERENCE

1 (bases 1 to 3897)

TITLE Role of kinesin-1 in the pathogenesis of DP101, a rare form of hereditary spastic paraplegia

PMID 2171102

PMID 2171102 A review of the mechanism of pathogenesis involved in spastic paraplegia type 10 when KIF5A is inactivated by mutations. Review article

REMARKS

2 (bases 1 to 3897)

1. Puvion, J., Nègre, J.F., Baye, F., Campion, J., Dubet, K., Riab, R., Jahn, G., and Nègre, B. alpha-Dynactin oligomers impede neuronal microtubule-kinesin transport. *J. Biol. Chem.* 289 (30), 21742-21754 (2013)

Pathways for the KIF5A gene

Peptide hormone metabolism

MHC class II antigen presentation

# GenBank 'graphics' sequence record

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

NCBI Reference Sequence: NM\_004984.2

Graphics

ICDCU NM\_004984.2 Homo sapiens kinesin family member 5A (KIF5A), mRNA

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

ACCESSION NM\_004984

VERSION NM\_004984.2 GI:45446748

KEYWORDS mRNA.

SOURCE Homo sapiens (human)

ORGANISM Homo sapiens

RAJAYAKA; METAZOA; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Chiroptera; Hominoidea; Homo.

REFERENCE

1 (bases 1 to 3897)

TITLE Role of kinesin-1 in the pathogenesis of DP101, a rare form of hereditary spastic paraplegia

PMID 2171102

PMID 2171102 A review of the mechanism of pathogenesis involved in spastic paraplegia type 10 when KIF5A is inactivated by mutations. Review article

REMARKS

2 (bases 1 to 3897)

1. Puvion, J., Nègre, J.F., Baye, F., Campion, J., Dubet, K., Riab, R., Jahn, G., and Nègre, B. alpha-Dynactin oligomers impede neuronal microtubule-kinesin transport. *J. Biol. Chem.* 289 (30), 21742-21754 (2013)

Pathways for the KIF5A gene

Peptide hormone metabolism

MHC class II antigen presentation

Dorsal/ventral axis

RefSeq alternative splicing

# GenBank sequence record, cont.

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

NCBI Reference Sequence: NM\_004984.2

FASTA

ICDCU NM\_004984.2 Homo sapiens kinesin family member 5A (KIF5A), mRNA

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

ACCESSION NM\_004984

VERSION NM\_004984.2 GI:45446748

KEYWORDS mRNA.

SOURCE Homo sapiens (human)

ORGANISM Homo sapiens

RAJAYAKA; METAZOA; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Chiroptera; Hominoidea; Homo.

REFERENCE

1 (bases 1 to 3897)

TITLE Role of kinesin-1 in the pathogenesis of DP101, a rare form of hereditary spastic paraplegia

PMID 2171102

PMID 2171102 A review of the mechanism of pathogenesis involved in spastic paraplegia type 10 when KIF5A is inactivated by mutations. Review article

REMARKS

2 (bases 1 to 3897)

1. Puvion, J., Nègre, J.F., Baye, F., Campion, J., Dubet, K., Riab, R., Jahn, G., and Nègre, B. alpha-Dynactin oligomers impede neuronal microtubule-kinesin transport. *J. Biol. Chem.* 289 (30), 21742-21754 (2013)

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

NCBI Reference Sequence: NM\_004984.2

FASTA

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

ICDCU NM\_004984.2 Homo sapiens kinesin family member 5A (KIF5A), mRNA

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

ACCESSION NM\_004984

VERSION NM\_004984.2 GI:45446748

KEYWORDS mRNA.

SOURCE Homo sapiens (human)

ORGANISM Homo sapiens

RAJAYAKA; METAZOA; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Chiroptera; Hominoidea; Homo.

REFERENCE

1 (bases 1 to 3897)

TITLE Role of kinesin-1 in the pathogenesis of DP101, a rare form of hereditary spastic paraplegia

PMID 2171102

PMID 2171102 A review of the mechanism of pathogenesis involved in spastic paraplegia type 10 when KIF5A is inactivated by mutations. Review article

REMARKS

2 (bases 1 to 3897)

1. Puvion, J., Nègre, J.F., Baye, F., Campion, J., Dubet, K., Riab, R., Jahn, G., and Nègre, B. alpha-Dynactin oligomers impede neuronal microtubule-kinesin transport. *J. Biol. Chem.* 289 (30), 21742-21754 (2013)

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

NCBI Reference Sequence: NM\_004984.2

FASTA

The actual sequence entry starts after the word ORIGIN

ICDCU NM\_004984.2 Homo sapiens kinesin family member 5A (KIF5A), mRNA

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

ACCESSION NM\_004984

VERSION NM\_004984.2 GI:45446748

KEYWORDS mRNA.

SOURCE Homo sapiens (human)

ORGANISM Homo sapiens

RAJAYAKA; METAZOA; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Chiroptera; Hominoidea; Homo.

REFERENCE

1 (bases 1 to 3897)

TITLE Role of kinesin-1 in the pathogenesis of DP101, a rare form of hereditary spastic paraplegia

PMID 2171102

PMID 2171102 A review of the mechanism of pathogenesis involved in spastic paraplegia type 10 when KIF5A is inactivated by mutations. Review article

REMARKS

2 (bases 1 to 3897)

1. Puvion, J., Nègre, J.F., Baye, F., Campion, J., Dubet, K., Riab, R., Jahn, G., and Nègre, B. alpha-Dynactin oligomers impede neuronal microtubule-kinesin transport. *J. Biol. Chem.* 289 (30), 21742-21754 (2013)

Pathways for the KIF5A gene

Peptide hormone metabolism

MHC class II antigen presentation

## 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

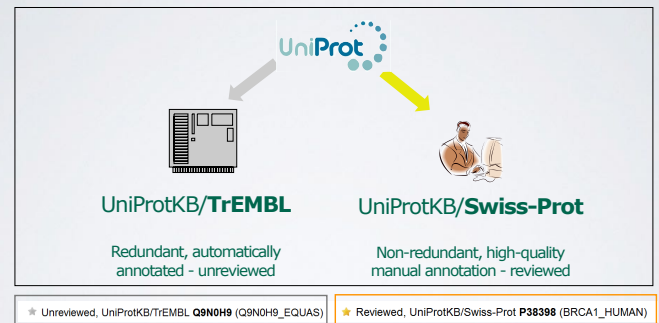
## UniProt: Protein sequence database

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

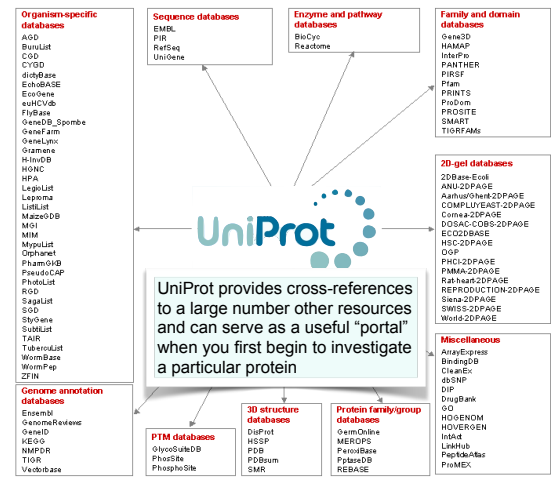
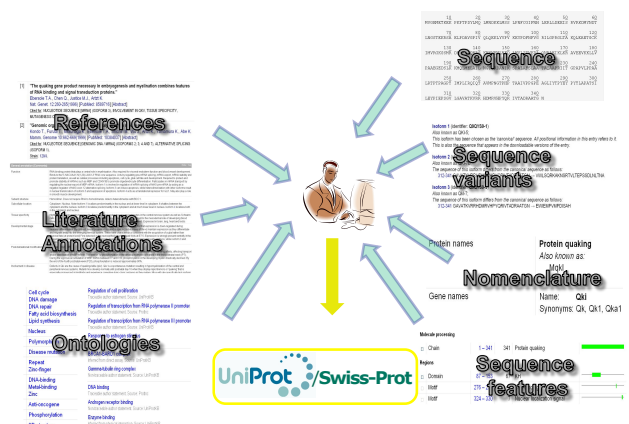
- UniProt comprises four databases:

- 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



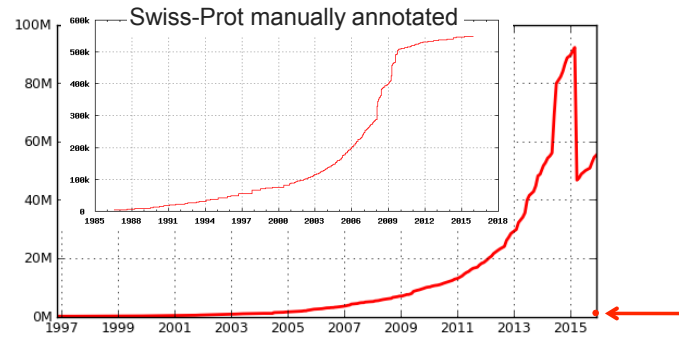
## The main information added to a UniProt/Swiss-Prot entry



## UniProt/Swiss-Prot vs UniProt/TrEMBL

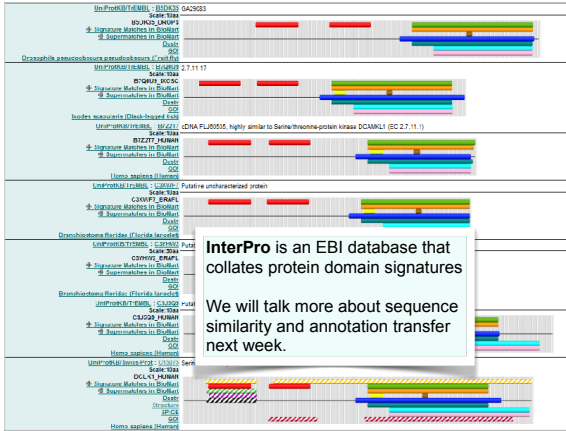
- *UniProtKB/Swiss-Prot* is a **non-redundant** database with one entry per protein
- *UniProtKB/TrEMBL* is a **redundant** database with one entry per translated ENA entry (ENA is the EBI's equivalent of GenBank)
  - › Therefore TrEMBL can contain multiple entries for the same protein
  - › Multiple 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)



71

## Same domain composition = same function = annotation transfer



72

## 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/>

Hands on demo (or see following slides)

Database	Count	Description
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	67,165	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		

75

NCBI Resources How To Sign in to NCBI

Gene  Search

Save search Advanced Help

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

Clear all

Did you mean ras as a gene symbol? Search Gene for ras as a symbol.

Results: 1 to 20 of 85633  
Filters activated: Current only. Clear all to show 87165 items.

Name/Gene ID	Description	Location	Aliases
ras ID: 16412	resistance to autogenic solzuros (Mus musculus (house mouse))		asr
ras ID: 43873	raspberry (Drosophila melanogaster (fruit fly))	Chromosoma X, NC_004354.4 (10744902..10749097)	Dmel_CG1799, CG1796, DmelCG1799, EP(X)1693

Find related data  
Database: Select  
[+][-][x] [All Fields] AND "Homo sapiens" [property]

76

NCBI Resources How To Sign in to NCBI

Gene  Search

Save search Advanced Help

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

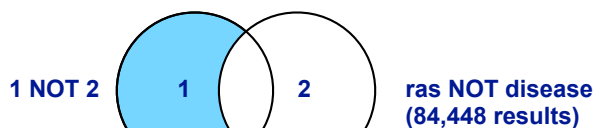
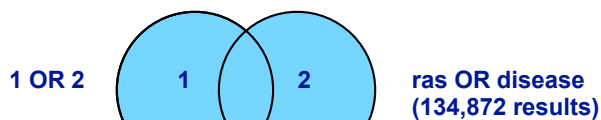
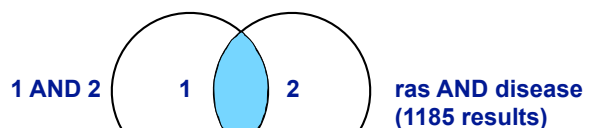
Clear all

Results: 1 to 20 of 1126  
Filters activated: Current only. Clear all to show 1439 items.

Name/Gene ID	Description	Location	Aliases
NRAS ID: 4863	neuroblastoma RAS viral (v-ras) oncogene homolog (Homo sapiens (human))	Chromosome 1, NC_000011.11 (114704464..114716884, complement)	RPS1, 1000E10.2, ALP54, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS ID: 3645	Kirsten rat sarcoma viral oncogene homolog (Homo sapiens (human))	Chromosome 12, NC_000012.12 (25205248..2520923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS, NS3, NS4, NS5

Find related data  
Database: Select  
[+][-][x] [All Fields] AND "Homo sapiens" [porgn] AND alive[property]

77



78

NCBI Resources How To Sign in to NCBI

Gene  Search

Save search Advanced Help

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

Clear all

Results: 1 to 20 of 1126  
Filters activated: Current only. Clear all to show 1439 items.

Name/Gene ID	Description	Location	Aliases
NRAS ID: 4863	neuroblastoma RAS viral (v-ras) oncogene homolog (Homo sapiens (human))	Chromosome 1, NC_000011.11 (114704464..114716884, complement)	RPS1, 1000E10.2, ALP54, CMNS, N-ras, NCMS1, NS6, NRAS
KRAS ID: 3645	Kirsten rat sarcoma viral oncogene homolog (Homo sapiens (human))	Chromosome 12, NC_000012.12 (25205248..2520923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS, NS3, NS4, NS5

Find related data  
Database: Select  
[+][-][x] [All Fields] AND "Homo sapiens" [porgn] AND alive[property]

79

NCBI Resources How To Sign in to NCBI

Gene  Search

Advanced Help

Display Settings: Full Report Send to: Hide sidebar >>

**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:8407  
See related Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTTHUMG00000171193

Gene type protein coding  
RefSeq status REVIEWED  
Organism *Homo sapiens*  
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhina; Catarrhini; Hominoidea; 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 pseudogenes(), Homology, Gene Ontology  
General protein information  
NCBI Reference Sequences (RefSeq)

80

NCBI Resources How To Sign in to NCBI

Gene  Search

Advanced Help

Display Settings: Full Report Send to: Hide sidebar >>

**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:8407  
See related Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTTHUMG00000171193

Gene type protein coding  
RefSeq status REVIEWED  
Organism *Homo sapiens*  
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhina; Catarrhini; Hominoidea; 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 pseudogenes(), Homology, Gene Ontology  
General protein information  
NCBI Reference Sequences (RefSeq)

81

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

Genomic context

Location: 12p12.1  
Exon count: 6

Annotation release	Status	Assembly	Chr.	Location
108	current	GRCv38 (GCF_000001405.26)	12	NC_000012.12 (25205246..25250823, complement)
105	previous assembly	GRCv37.p13 (GCF_000001405.25)	12	NC_000012.11 (25358190..25403870, complement)

Chromosome 12 - NC\_000012.12

Genomic regions, transcripts, and products

Genomic Sequence: NC\_000012.12 chromosome 12 reference GRCv38 Primary Assembly

82

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

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 pseudogenes, Homology, Gene Ontology
- General protein information
- NCBI Reference Sequences (RefSeq)

83

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

Process	Evidence Code	PubS
Ec-rep10n 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	

84

## GO: Gene Ontology

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

UniProt-GOA

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 bioinformatics. 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.

85

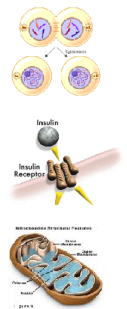
## Why do we need Ontologies?

- 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

86

## 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



87

Gene Ontology Provided by GOA

Function

- GDP binding
- GMP binding
- GTP binding
- LRR domain binding
- protein binding
- protein complex binding

Process

- Epitope receptor signaling pathway
- GTP catalytic process
- MAPK cascade
- Ras protein signal transduction
- actin cytoskeleton organization
- activation of MAPKK activity
- axon guidance
- blood coagulation

Evidence Code

- TAS
- IEA
- TAS
- TAS
- IEA
- TAS
- TAS
- TAS
- TAS

Pubmed

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

genomic X01669.1 CAA25828.1

Items 1 - 25 of 43 < Prev Page 1 of 2 Next >

Protein Accession Links

- GenPept Link
- UniProtKB Link
- UniProtKB/Swiss-Prot:P01116
- GenPept

Additional links

You are here: NCBI > Genes & Expression > Gene

GETTING STARTED

- NCBI Education
- NCBI Help Manual
- NCBI Handbook
- Training & Tutorials

RESOURCES

- Chemicals & Bioassays
- Data & Software
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Histology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy

POPULAR

- PubMed
- Bioconductor
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- Sequence Analysis
- 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

Write to the Help Desk

Scroll down to UniProt link

UniProt will detail much more information for protein coding genes

P01116 - RASK\_HUMAN

Protein: GTPase KRas

Gene: KRAS

Organism: Homo sapiens (Human)

Status: Reviewed - experimental evidence at protein level

Function

Ras proteins bind GDP/GTP and possess intrinsic GTPase activity. Plays an important role in the regulation of cell proliferation (PubMed:23495361, PubMed:22711938). #2 Publications @ Crossref

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. #3 Publications

Regions

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Nucleotide binding <sup>1</sup>	10 - 18	9	GTP @ 2 Publications			
Nucleotide binding <sup>2</sup>	29 - 35	7	GTP @ 2 Publications			
Nucleotide binding <sup>3</sup>	59 - 60	2	GTP @ 2 Publications			

Display None

- FUNCTION
- NAME & TAXONOMY
- SUBCELLULAR LOCATION
- PATROLN/BIOTICH
- PTM/PROCESSING
- EXPRESSION
- INTERACTION
- STRUCTURE
- FAMILY & DOMAINS
- SEQUENCES (3)
- CROSS REFERENCES

Scroll down to UniProt link

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

P01116 - RASK\_HUMAN

Protein: GTPase KRas

Gene: KRAS

Organism: Homo sapiens (Human)

Status: Reviewed - experimental evidence at protein level

Function

Ras proteins bind GDP/GTP and possess intrinsic GTPase activity. Plays an important role in the regulation of cell proliferation (PubMed:23495361, PubMed:22711938). #2 Publications @ Crossref

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. #3 Publications

Regions

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Nucleotide binding <sup>1</sup>	10 - 18	9	GTP @ 2 Publications			
Nucleotide binding <sup>2</sup>	29 - 35	7	GTP @ 2 Publications			
Nucleotide binding <sup>3</sup>	59 - 60	2	GTP @ 2 Publications			

Display None

- FUNCTION
- NAME & TAXONOMY
- SUBCELLULAR LOCATION
- PATROLN/BIOTICH
- PTM/PROCESSING
- EXPRESSION
- INTERACTION
- STRUCTURE
- FAMILY & DOMAINS
- SEQUENCES (3)
- CROSS REFERENCES

Scroll down to UniProt link

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

Pathology & Biotech

Involvement in disease

LEUKEMIA, ACUTE MYELOGENOUS (AML) [MIM:601435]: A subtype of acute leukemia, a cancer of the white blood cells. AML is a malignant disease of bone marrow characterized by maturation arrest of hematopoietic precursors at an early stage of development. Clonal expansion of myeloid blasts occurs in bone marrow, blood, and other tissues. Myelogenous leukemias develop from changes in cells that normally produce neutrophils, basophils, eosinophils and monocytes. #1 Publications

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 <sup>1</sup>	10 - 10	1	G - GC 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 Publications		VAR_034601	
------------------------------	---------	---	--	--	------------	--

LEUKEMIA, MYELOID F. 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 downturned cystic and low-set posteriorly rotated ears, and a high incidence of congenital heart

Display None

- FUNCTION
- NAME & TAXONOMY
- SUBCELLULAR LOCATION
- PATROLN/BIOTICH
- PTM/PROCESSING
- EXPRESSION
- INTERACTION
- STRUCTURE
- FAMILY & DOMAINS
- SEQUENCES (3)
- CROSS REFERENCES
- PUBLICATIONS
- ENTRY INFORMATION
- MISCELLANEOUS
- SIMILAR PROTEINS

Scroll down to UniProt link

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

Structure

Secondary structure

Legend: Helix Turn Beta strand

Show more details

3D structure databases

Entry	Method	Resolution (Å)	Chain	Positions	PDBsum
1D8D	X-ray	2.00	P	178-188	[+]
1D8E	X-ray	3.00	P	178-188	[+]
1K20	X-ray	2.20	C	169-173	[+]
1K2P	X-ray	2.10	C	169-173	[+]
3GTF	X-ray	2.27	A/B/C/D/E/F	1-154	[+]
4D5N	X-ray	2.03	A	2-154	[+]
4D5O	X-ray	1.85	A	2-154	[+]
4E9K	X-ray	2.00	A	1-154	[+]
4EPT	X-ray	2.00	A	1-154	[+]
4EPV	X-ray	1.35	A	1-154	[+]
4EPW	X-ray	1.70	A	1-154	[+]
4EPX	X-ray	1.76	A	1-154	[+]
4EPY	X-ray	1.80	A	1-154	[+]
4L8G	X-ray	1.52	A	1-159	[+]
4LDJ	X-ray	1.15	A	1-154	[+]
4L7K	X-ray	1.50	A/B	1-159	[+]

Display None

- FUNCTION
- NAME & TAXONOMY
- SUBCELLULAR LOCATION
- PATROLN/BIOTICH
- PTM/PROCESSING
- EXPRESSION
- INTERACTION
- STRUCTURE
- FAMILY & DOMAINS
- SEQUENCES (3)
- CROSS REFERENCES
- PUBLICATIONS
- ENTRY INFORMATION
- MISCELLANEOUS
- SIMILAR PROTEINS

Scroll down to UniProt link

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

Family and domain databases

Class3D1 3,40,50,300, 1 hit.  
 HtrpH11 [PR027417, P4loop\_NTPase, [Graphical view]  
 PR002225 Small\_GTP\_bd\_dom. [PR002225, Small\_GTPase, [Graphical view]  
 PR022945 Small\_GTPase\_Ras. [Graphical view]  
 PANTHER1 PTHR24070, PTHR24070, 1 hit.  
 Pfam1 PF00071, Ras, 1 hit. [Graphical view]  
 PRINTS1 PR00449, RASTRNSFRNG. [Graphical view]  
 SMART1 SM01173, RAS, 1 hit. [Graphical view]  
 SUPFAM1 SSF2540, SSF2540, 1 hit. [Graphical view]  
 TIGRFAM1 TIGR00231, small\_GTP\_1 hit. [Graphical view]  
 PROSITE1 PS1421, RAS, 1 hit. [Graphical view]

Sequences (2)

Sequence status: Complete.  
 Sequence processing: The displayed sequence is further processed into a mature form.  
 This entry describes 2 isoforms produced by alternative splicing.

**Pfam** is one of the best protein family databases

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

Family: Ras (PF00071)

Summary: Ras family

Domain organisation

Alignments

HMM logo

Trees

Curator & model

Species

Structures

Jump to...

Interact

Enter ID/acc

126 architectures 4150 sequences 6 interactions 248 species 114 structures

There are 6 interactions for this family. More...

Tubulin Tubulin\_C Kinesin Tubulin Kinesin

Questions or comments: pfam@janelia.hhmi.org  
 Howard Hughes Medical Institute

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

Species distribution

This visualization provides a simple graphical representation of the distribution of the family across species. You can find the original interactive version in the adjacent tab.

Weight segments by...  
 number of sequences  
 number of species

Change the size of the sunburst  
 Larger

Colour assignments

Green Eukarya  
 Red Bacteria  
 Blue Eukarya  
 Yellow Eukarya  
 Purple Eukarya

Selections

Click on selected sequences to view details in Pfam's family file. (Click on the left)

Click on selected sequences to view details in Pfam's family file. (Click on the left)

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

Alignment for selected sequences

Currently displaying 1 to 50 of 500 entries in 1000 alignment blocks. Show 1000 entries of alignment

Jump to...

Interact

Enter ID/acc

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

Family: Ras (PF00071)

Summary

Domain organisation

Alignments

HMM logo

Trees

Curator & model

Species

Interactions

Structures

Jump to...

Interact

Enter ID/acc

126 architectures 4150 sequences 6 interactions 248 species 114 structures

There are 6 interactions for this family. More...

Tubulin Tubulin\_C Kinesin Tubulin Kinesin

Questions or comments: pfam@janelia.hhmi.org  
 Howard Hughes Medical Institute

Family: Kinesin (PF00225)

Summary

Domain organisation

Alignments

HMM logo

Trees

Curator & model

Species

Interactions

Structures

Jump to...

Interact

Enter ID/acc

126 architectures 4150 sequences 6 interactions 248 species 114 structures

There are 6 interactions for this family. More...

Tubulin Tubulin\_C Kinesin Tubulin Kinesin

Questions or comments: pfam@janelia.hhmi.org  
 Howard Hughes Medical Institute

HHMI janelia farm research campus

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam Family: **Kinesin (PF00225)**

126 architectures 4190 sequences 6 interactions 248 species 114 structures

**Summary**

**Domain organisation**

**Clares**

**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 [RSCB](#) 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
ARBKD1_GIALA	11 - 335	2vvo	A	11 - 335	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			B	11 - 335	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
CENPL_HUMAN	12 - 329	1t5c	A	12 - 329	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			B	12 - 329	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
KAR3_YEAST	392 - 723	1f9t	A	392 - 723	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			A	392 - 723	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			A	392 - 723	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			B	392 - 723	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
K113B_HUMAN	11 - 352	3gbl	A	11 - 352	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			B	11 - 352	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			C	11 - 352	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			A	24 - 359	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
		1i6	B	24 - 359	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			A	24 - 359	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			B	24 - 359	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			A	24 - 359	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
		1x88	B	24 - 359	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>
			A	24 - 359	<a href="#">Jmol</a> <a href="#">AstexViewer</a> <a href="#">SPICE</a>

Pfam: Jmol

structure/viewer/viewer=jmol&id=3bfn

welcome to sanger institute

PDB entry 3bfn

Jmol

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

## 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

## ADDITIONAL DATABASES OF NOTE (SLIDES FOR YOUR REFERENCE)