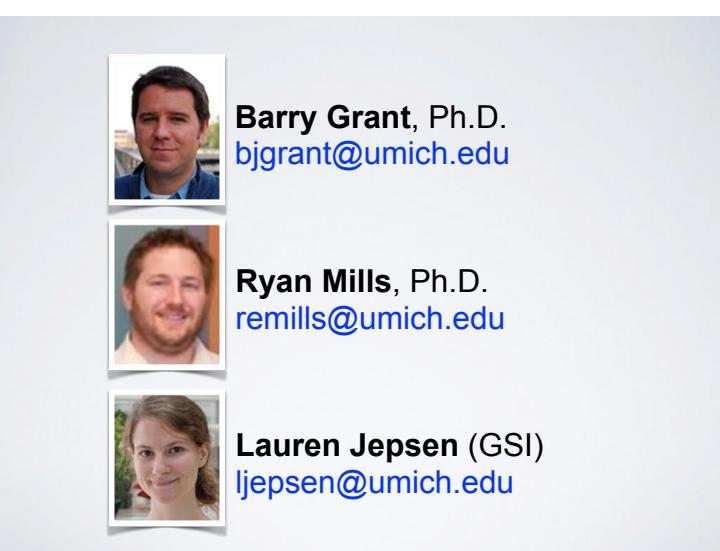


INTRODUCTION TO BIOINFORMATICS

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

Barry Grant
University of Michigan
www.thegrantlab.org

BIOINF 525 http://bioboot.github.io/bioinf525_w17/ 10-Jan-2017



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

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

Lauren Jepsen (GSI)
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:
[PDF 1 \(bioinformatics review\)](#),
[PDF 2 \(bioinformatics challenges\)](#).
- Complete the **lecture 1.1 homework questions**:
<http://tinyurl.com/bioinf525-quiz1>

MORE DEFINITIONS

Q. What is Bioinformatics?

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

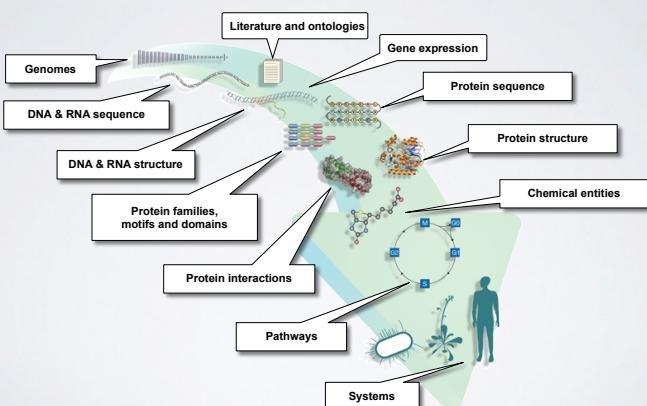
... Bioinformatics is a hybrid of biology and computer science
Bioinformatics is computer aided biology!

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

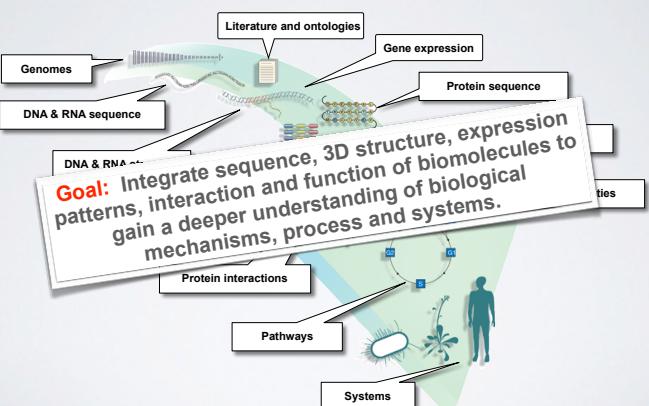
► "Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying "**informatics**" techniques (derived from disciplines such as applied maths, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a **large-scale**.
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>)

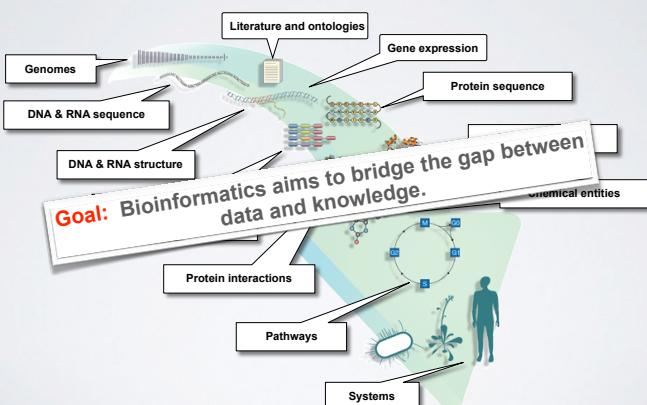
Major types of Bioinformatics Data



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

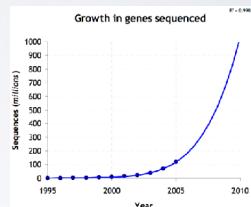
- DNA sequence determines protein sequence.
- Protein sequence determines protein structure.
- Protein structure determines protein function.
- Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.

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

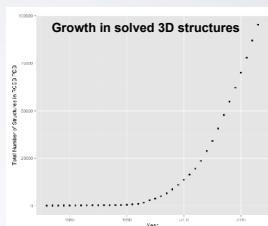


E.G. data from sequencing, structural genomics, microarrays, proteomics, new high throughput assays, etc...

Why do we need Bioinformatics?

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

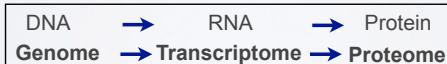
- Bioinformatics provides methods for the efficient:
 - storage
 - annotation
 - search and retrieval
 - data integration
 - data mining and analysis



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)

The screenshot shows the Protein BLAST search interface from NCBI. It includes sections for General Parameters, Scoring Parameters, Gap Costs, Compositional adjustments, Filters and Masking, and PSI/PHI/DELTA BLAST. A note on the right says "Even Blast has many settable parameters". Below the main interface is a box titled "Related tools with different terminology" containing links to programs like FASTA, HMMER, and BLAST.

Key Online Bioinformatics Resources: NCBI & EBI

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

The screenshot shows the NCBI homepage. It features a sidebar with links 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". The main content area includes sections for "Welcome to NCBI", "Popular Resources" (PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, PubChem), "Get Started" (with links to tools, downloads, how-tos, and submissions), "3D Structures" (with a molecular visualization image), and "NCBI Announcements" (mentioning a new version of Genome Workbench).

The screenshot shows the EBI homepage. It features a sidebar with links like "Get Started", "Tools", "Downloads", "How-Tos", and "Submissions". The main content area includes sections for "Welcome to EBI", "Popular Resources" (PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, PubChem), "Get Started" (with links to tools, downloads, how-tos, and submissions), "3D Structures" (with a molecular visualization image), and "NCBI Announcements" (mentioning a new version of Genome Workbench).

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

<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>

The screenshot shows the NCBI homepage. It features a sidebar with links 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". The main content area includes sections for "Welcome to NCBI", "Popular Resources" (PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, PubChem), "Get Started" (with links to tools, downloads, how-tos, and submissions), "3D Structures" (with a molecular visualization image), and "NCBI Announcements" (mentioning a new version of Genome Workbench).

The screenshot shows the NCBI homepage with several red arrows highlighting specific resources. One arrow points to "PubMed" in the "Popular Resources" section. Another arrow points to "BLAST". A third arrow points to "SNP". A fourth arrow points to "Gene". A fifth arrow points to "Protein". A sixth arrow points to "PubChem". The sidebar and main content area are identical to the previous screenshot.

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

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

The screenshot shows the NCBI homepage with a banner stating "Notable NCBI databases include: **GenBank**, **RefSeq**, **PubMed**, **dbSNP** and the search tools **ENTREZ** and **BLAST**". Below the banner are links for Homology, Literature, Proteins, Sequence Analysis, Taxonomy, Training & Tutorials, and Variation. A sidebar on the left lists "databases", "Protein", "PubChem", and "3D Structures". A news announcement about a new version of Genome Workbench is also present.

Key Online Bioinformatics Resources: NCBI & EBI

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

The image compares the NCBI and EBI websites. The NCBI page on the left shows its homepage with various databases and search tools. The EBI page on the right shows its homepage with sections for Services, Research, Training, Industry, European Coordination, and EMBL ALUMNI. Both pages have a "Popular" sidebar on the right.

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 EBI homepage with a search bar, a "Services" button highlighted in red, and sections for Research, Training, Industry, European Coordination, and EMBL ALUMNI. A news banner from EMBL-EBI is visible at the bottom.

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

The screenshot shows the EBI Services page with a "Bioinformatics services" section. It highlights "freely available" and up-to-date molecular databases. It lists services like DNA & RNA, Gene expression, Proteins, Structures, Systems, Chemical biology, Ontologies, Literature, and Cross domain. A "Popular" sidebar on the right includes Ensembl, UniProt, PDB, ArrayExpress, and ChEMBL.

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

The screenshot shows the EBI Services page with a "Bioinformatics services" section. It highlights "freely available" and up-to-date molecular databases. It lists services like DNA & RNA, Gene expression, Proteins, Structures, Systems, Chemical biology, Ontologies, Literature, and Cross domain. A "Popular" sidebar on the right includes Ensembl, UniProt, PDB, ArrayExpress, and ChEMBL. The "Proteins" link is also highlighted in red.

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

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

This screenshot shows the 'Proteins' section of the EBI website. It lists several popular services under the heading 'Popular services': UniProt (The Universal Protein Resource), InterPro (A database for the classification of proteins into families, domains and conserved sites), PRIDE (The Proteomics Identifications Database), Pfam (A database of hidden Markov models and alignments to describe eukaryotic protein families and domains), Clustal Omega (Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalN alignment tools), HMMER (protein homology search), and InterProScan 5 (Identifies protein sequences against InterPro's predictive protein signatures). A 'Quick links' sidebar on the right provides links to 'Regular services in this category', 'All services in this category', and 'Project websites in this category'.

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

This screenshot shows the homepage of the European Bioinformatics Institute (EMBL-EBI). It features a main banner with the text 'The European Bioinformatics Institute Part of the European Molecular Biology Laboratory'. Below the banner, there is a search bar with the placeholder 'Find a gene, protein or chemical:' and a 'Search' button. To the right of the search bar, there is a 'Popular' sidebar with links to 'Services', 'Research', 'Training' (which is highlighted with a red box), 'Industry', and 'European Coordination'. The 'Training' section includes a sub-section for 'EMBL ALUMNI'. On the right side of the page, there is a 'Upcoming events' section with a thumbnail for the 'Plant and Animal Genome conference (PAG XXV)'.

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

This screenshot shows a 'Train online' course titled 'Using sequence similarity searching tools at EMBL-EBI: webinar'. The course content is described as 'Using sequence similarity searching tools at EMBL-EBI: webinar'. The video player shows a thumbnail of a man speaking, with the caption 'Using sequence similarity search tools at EMBL-EBI: Finding homologous sequences with BLAST, FASTA, PSI-Blast etc.' and the name 'Andrew Cowley'. The course navigation bar includes links for 'Train online Home', 'Course list', 'Glossary', 'Support & Feedback', 'Log In / Register', and 'Navigation'.

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

This screenshot shows the 'Train online' section of the EBI website. It features a large callout box with the text 'Notable EBI databases include: ENA, UniProt, Ensembl' and 'and the tools FASTA, BLAST, InterProScan, MUSCLE, DALI, HMMER'. Below this, there is a 'Find a course' section with a 'Browse by subject' dropdown menu containing 'Genes and Genomes', 'Gene Expression', and '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, AM5db, 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, DIR, DOGS, DOMO, DPD, DPLinteract, ECDC, ECGC, EC02DATABASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlesne, 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, Klothe, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPep5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, MycDB, NDB, NRSub, 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, SeqAnaRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtList, 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, AM5db, 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, DIR, DOGS, DOMO, DPD, DPLinteract, ECDC, ECGC, EC02DATABASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlesne, 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, Klothe, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPep5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, MycDB, NDB, NRSub, 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, SeqAnaRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtList, 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 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 the Oxford Journals Nucleic Acids Research Database Summary Page. The URL in the address bar is <http://www.oxfordjournals.org/nar/database/c/>. The page features a navigation menu with links like 'ABOUT THIS JOURNAL', 'CONTACT THIS JOURNAL', 'SUBSCRIPTIONS', 'CUPERTINOBILE', 'ARCHIVE', and 'SEARCH'. Below the menu, there's a section titled '2014 NAR Database Summary Paper Category List' with a table of categories. A red box highlights the URL at the bottom of the page.

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 "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](#))

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)*. The page displays various fields such as LOCUS, DEFINITION, VERSION, KEYWORDS, SOURCE, ORGANISM, REFERENCES, AUTHORS, TITLE, JOURNAL, PAGES, REVIEWER, and COMMENTS. A callout box highlights the 'GenBank flat file format' which includes unique identifiers such as the **ACCESSION** number.

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

The screenshot shows the NCBI GenBank sequence record for *Homo sapiens kinesin family member 5A (KIF5A)*. The page displays detailed biological and experimental data for the gene, including its accession number NM_004984.2, its length (3897 bp), and its function as mRNA. A callout box highlights the 'GenBank flat file format' which includes unique identifiers such as the **ACCESSION** number.

GenBank sequence record

Home sapiens kinesin family member 5A (KIF5A). mRNA - Nucleotide - NCBI

www.ncbi.nlm.nih.gov/nucleotide/NM_004864.2

Home sapiens kinesin family member 5A (KIF5A). mRNA - Nucleotide - NCBI

NCBI Resources How To Sign in to NCBI

Nucleotides Nucleotide [KIF5A] AND "Homo sapiens" Search Help

Display Settings GenBank Send: Change region shown

NCBI Reference Sequence: NM_004864.2 FASTA Oracle Can set different display formats here

GenBank

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

NCBI Reference Sequence: NM_004864.2

FASTA Oracle

Can set different display formats here

Send: Change region shown

GenBank

LOCUS NM_004864 3887 bp m RNA linear PRX 10-JAN-2014

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

ACCESSION NM_004864

VERSION 4

KEYWORDS *h2SeqP*

SOURCE *Homo sapiens (Human)*

ORGANISM *Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Hominoidea; Hominidae; Homo*

REFERENCE *(Human 1C 3897)*

AUTHORS *Kaweguchi,K.*

TITLE *Loss-of-function of WIF10, a rare form of hereditary spastic paraparesis*

JOURNAL *Neuroscience* 19 (41), 336-344 (2013)

PONDER *Deacon, J.*

RENAME *Deacon, J. A review of the mechanism of pathogenicity involved in spastic paraparesis type 10 when KIF5A is inactivated by mutations. Recent articles*

REFERENCE *(Human 1C 3897)*

AUTHORS *Piotrowski,W., Breyer,S., Complioni,S., Budde,K., Bleck,R., Bobk,J. and Winkler,E.*

TITLE *Microtubule oligomers impair neuronal microtubule-kinesin interplay*

JOURNAL *J. Biol. Chem.* 288 (36), 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 - Nucleotide - NCBI

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

Nucleotide Resources How To

Nucleotide Nucleotide + [KIF5A] AND "Homo sapiens" **Search**

Units Advanced

Display Settings, **Geobank**

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

NCBI Reference Sequence: NM_004984.2

FASTA **Graphics**

Go!

LOCUS NM_004984
ORGANISM *Homo sapiens* kinesin family member 5A (KIF5A), mRNA.
ACCESSION NM_004984.2
VERSION NM_004984.2
KEYWORDS Kif5a; Kinesin-1; Human
SCHEME Human
ORIGIN Human
Bakayoko, Etienne; Chodat; Crneciate; Vertebrate; Euteleostomi; Mammal; Eutheria; Euarchontoglires; Primate; Haplorhini; Catarrhini; Hominoidea; Homo;

REFERENCE Bakayoko, E., et al. (2013) Human kinesin-1 is essential for GABA(A) receptor transport. *J Neurosci.* 33:34-344.

AUTHORS Bakayoko, E., et al.

TITLE Role of kinesin-1 in the pathogenesis of SCG10, a rare form of hereditary spastic paraparesis.

JOURNAL *Neurogenetics* 19 (4): 339-344 (2013)

PUBLISHED 23787106

REMARKS GeneID:11. A review of the mechanisms of pathogenesis involved in specific genetic type 10 when KIF5A is inactivated by mutations. Review article.

REFERENCE 2 (BABA, I. Tc 3897) BABA, I., KANG, S., CAMPIONI, S., DUDES, K., RICK, R., DABH, K.J., and WINTER, J.

AUTHORS BABA, I., KANG, S., CAMPIONI, S., DUDES, K., RICK, R., DABH, K.J., and WINTER, J.

TITLE alpha-Synuclein oligomers impair neuronal microtubule-kinesin interaction. *J Biol Chem.* 288 (30): 21742-21754 (2013)

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

ORIGIN 34477671

Send: **Change region shown**

Customize view

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find In This Sequence

Articles about the KIF5A gene

5'-Sugars oligomerize bacterial microcolicin *InjE* [J Neurosci. 2013]

Molecular motor KIF5A is essential for GABA(A) receptor transport. [J Neurosci. 2013]

Systematic analysis of ubiquitination dynamics reveals a key role [Nat Cell Biol. 2012]

See all...

Pathways for the KIF5A gene

Peptide hormone metabolism

MHC class I antigen presentation

GenBank sequence record, cont.

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

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. UniProkB (Knowledgebase)

Containing **Swiss-Prot** and **TrEMBL** components
(these correspond to hand curated and automatically annotated entries respectively)

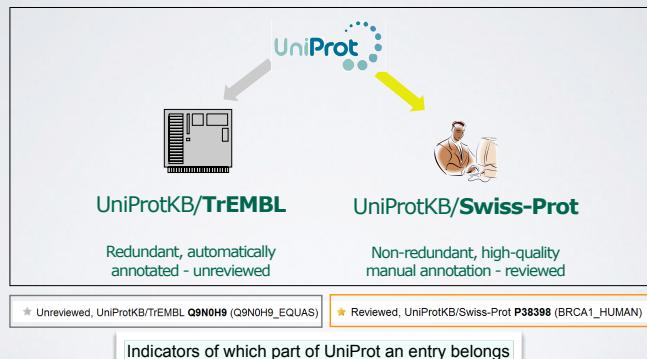
2. UniRef (Reference Clusters)

Filtered version of UniProkB at various levels of sequence identity
e.g. **UniRef90** contains sequences with a maximum of 90% sequence identity to each other

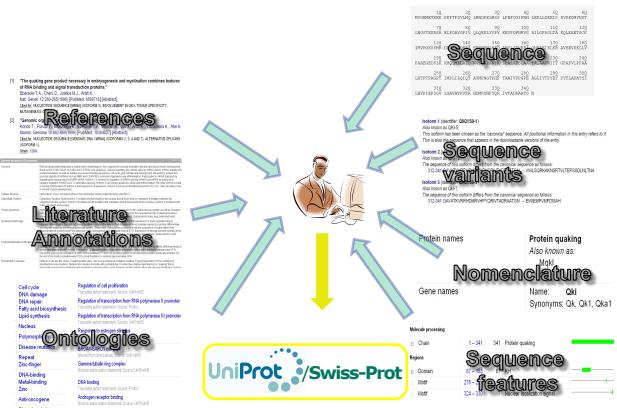
3. UniParc (Archive) with database cross-references to source.

4. UniMES (Metagenomic and Environmental Sequences)

The two sides of UniProtKB



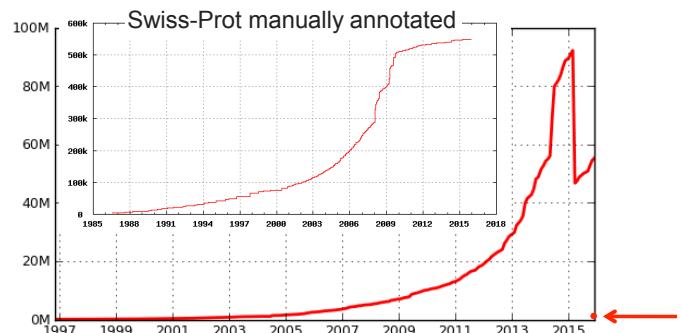
The main information added to a UniProt/Swiss-Prot entry



UniProt/Swiss-Prot vs UniProt/TrEMBL

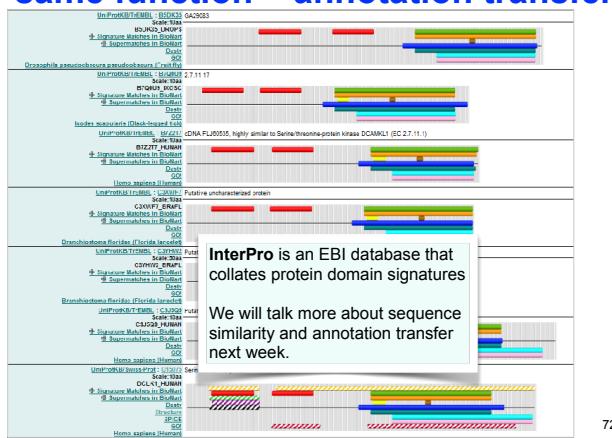
- UniProtKB/Swiss-Prot is a **non-redundant** database with one entry per protein
 - UniProtKB/TrEMBL is a **redundant** database with one entry per translated ENA entry (ENA is the EBI's equivalent of GenBank)
 - Therefore TrEMBL can contain multiple entries for the same protein
 - Multiple UniProtKB/TrEMBL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



74

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:

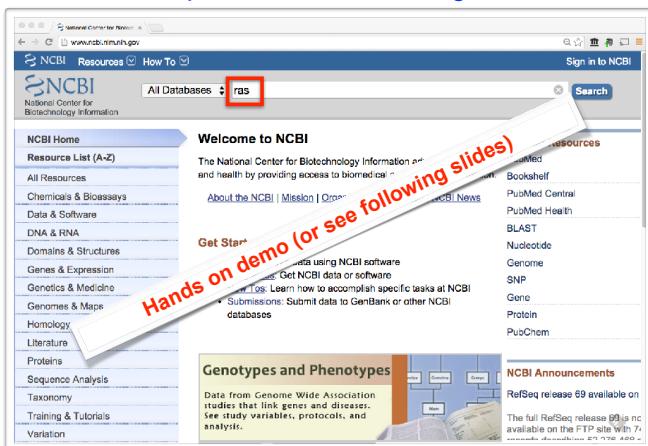
"What do you know about that 'Kras' gene the speaker kept taking about?"

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

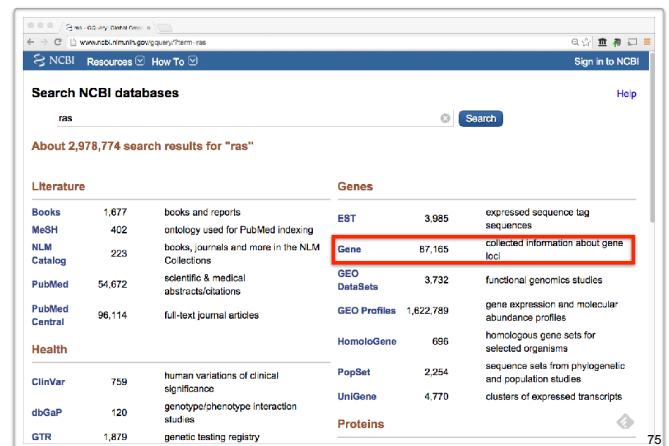
- Google?
 - Library?
 - **Bioinformatics databases at NCBI and EBI**

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

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



A screenshot of the NCBI homepage. At the top, there is a blue navigation bar with white text. Below it, the main content area features a large red banner with white text that reads "Hands on demo (or see following slides)". To the left of the banner, there is a sidebar with a blue header that says "Get Started". The main content area contains several sections: "About the NCBI | Mission | Core", "NCBI News", "Funding", "Programs", "Education", "NCBI Resources", "NCBI Databases", "NCBI Tools", "NCBI Software", and "NCBI Books".



75

NCBI Resources How To Sign in to NCBI

Gene Gene ras Search Help

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

Clear all Gene sources Genomic Metagenomic Organanelles Plasmids Plastids Categories Alternatively spliced Annotated genes Non-coding Protein-coding Pseudogenes Sequence content CCDS Ensembl RefSeq

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

Results: 1 to 20 of 8563 << First < Prev Page 1 of 4282 Next > Last >>

Filters activated: Current only. Clear all to show 87165 items.

Name/Gene ID	Description	Location	Aliases
ras	responsible for signaling	ras	
ID: 19412	solitaires (Mus musculus) [house mouse]	Drosophila melanogaster (fruit fly)	CG1789, NC_004354.4, CG1790, CG1791, DmelCG1790, EPX1063,
rasberry	Chromosome X, NC_004354.4	DmelCG1790, EPX1063,	
ID: 43873	[Drosophila melanogaster (fruit fly)]		

Find related data Database: Selected Find items Search details ras AND (All Fields) AND alive[property] 76

NCBI Resources How To Sign in to NCBI

Gene Gene (ras) AND "Homo sapiens" [prgn:txid9606] Search Help

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

Clear all Gene sources Genomic Categories Alternatively spliced Annotated genes Non-coding Protein-coding Pseudogenes Sequence content CCDS Ensembl RefSeq

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

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

Name/Gene ID	Description	Location	Aliases
NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog	Chromosome 1, NC_000011.11	RPS5, ALPS4, CMNS, N-ras, NCMS1, NS8, NRAS
ID: 4693	(114704464, 114716864)		
KRAS	Kirsten rat sarcoma viral oncogene homolog	Chromosome 12, NC_000012.12	C-K-RAS, CFC2, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS, NS2, RAS2
ID: 3645	(25205246, 25250923, complement)		
[Homo sapiens (human)]			

Find related data Database: Selected Find items Search details ras AND (All Fields) AND "Homo sapiens" [prgn] AND alive[property] 77

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)

78

NCBI Resources How To Sign in to NCBI

Gene Gene (ras) AND "Homo sapiens" [prgn:txid9606] Search Help

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

Clear all Gene sources Genomic Categories Alternatively spliced Annotated genes Non-coding Protein-coding Pseudogenes Sequence content CCDS Ensembl RefSeq

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

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

Name/Gene ID	Description	Location	Aliases
NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog	Chromosome 1, NC_000011.11	RPS5, ALPS4, CMNS, N-ras, NCMS1, NS8, NRAS
ID: 4693	(114704464, 114716864)		
KRAS	Kirsten rat sarcoma viral oncogene homolog	Chromosome 12, NC_000012.12	C-K-RAS, CFC2, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS, NS2, RAS2
ID: 3645	(25205246, 25250923, complement)		
[Homo sapiens (human)]			

Find related data Database: Selected Find items Search details ras AND (All Fields) AND "Homo sapiens" [prgn] AND alive[property] 79

NCBI Resources How To Sign in to NCBI

Gene Gene KRAS Search Help

Show additional filters Display Settings: Full Report Send to: Hide sidebar >>

Clear all Gene sources Genomic Categories Alternatively spliced Annotated genes Non-coding Protein-coding Pseudogenes Sequence content CCDS Ensembl RefSeq

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 (RefSeqs)

Official Symbol KRAS provided by HGNC
Official Full Name 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:OTTHUMG00000171193
Gene type protein coding
RefSeq status REVIEWED
Organism Homo sapiens
Lineage Eukaryota; Metazoa; Chordata; Craniota; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrini; Hominoidea; Homo
Also known as NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2; K-

30

NCBI Resources How To Sign in to NCBI

Gene Gene KRAS Search Help

Show additional filters Display Settings: Full Report Send to: Hide sidebar >>

Clear all Gene sources Genomic Categories Alternatively spliced Annotated genes Non-coding Protein-coding Pseudogenes Sequence content CCDS Ensembl RefSeq

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 (RefSeqs)

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

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:OTTHUMG00000171193
Gene type protein coding
RefSeq status REVIEWED
Organism Homo sapiens
Lineage Eukaryota; Metazoa; Chordata; Craniota; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrini; Hominoidea; Homo
Also known as NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2; K-

81

www.ncbi.nlm.nih.gov/gene/3845/genomic-context

Genomic context

Location: 12p12.1 **Exon count:** 6

See KRAS in Epigenomics MapViewer

Annotation release	Status	Assembly	Chr	Location
106	current	GRC:38 (GCF_000001405_25)	12	NC_000012.12 (25205246..25250923, complement)
105	previous assembly	GRC:37.p13 (GCF_000001405_25)	12	NC_000012.11 (25368100..25403670, complement)

Genomic regions, transcripts, and products

Chromosome 12 - NC_000012.12

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

Go to nucleotide: Graphics FASTA GenBank Nucleotide

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

Gene

KRAS KRAS (human)

Gene ID: 3845

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

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:OTTHUMG00001711193
Gene type protein coding
RefSeq status REVIEWED
Organism Homo sapiens
Lineage Eukaryota; Metazoa; Chordata; Craniota; Vertebrata; Euteleostomi; Mammalia; Eutheria; Eucarchontoglires; Primates; Haplorhini; Catarrhini; Hominoidea; Homo
Also known as NS; NS3; CFC2; KRAS1; KRAS2; RASK2; KI-RAS; C-K-RAS; K-RAS2A; K-RAS2B

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)

www.ncbi.nlm.nih.gov/gene/goterm-goa

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-gamma receptor signaling pathway	TAS	
GTP catabolic process	IEA	
MAPK cascade	TAS	
Ras protein signal transduction	IEA	
actin cytoskeleton organization	TAS	
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

www.ebi.ac.uk/GOA/

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.

Menu

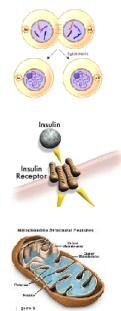
- Downloads
- Searching UniProt-GOA
- Annotation Methods
- Annotation Tutorial
- Manual Annotation Efforts
 - Reference Genome Annotation Initiative
 - Cardiovascular Gene Ontology Annotation Initiative
 - Renal Gene Ontology Annotation Initiative
 - Enzyme Gene

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



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

↓ Scroll down to
UniProt link

UniProt will detail much more information for protein coding genes

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?

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

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

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... ?

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... ?

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

Family: Kinesin (PF00225)
Loading page components (1 remaining)...

Interactions
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

Pfam: Family: Kinesin (PF00225) <http://pfam.janelia.org/family/kinesin#tabview=tab9>

HHMI
janelia farm research campus

Pfam
keyword search Go

Family: Kinesin (PF00225)

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 PDB[®] 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
ABBK01_GITALA	11 - 335	2vog	A	11 - 335	Jmol AstexViewer SPICE
			B	11 - 335	Jmol AstexViewer SPICE
CENPE_HUMAN	12 - 329	1tsc	A	12 - 329	Jmol AstexViewer SPICE
			B	12 - 329	Jmol AstexViewer SPICE
KAR83_YEAST	392 - 723	1pxt	A	392 - 723	Jmol AstexViewer SPICE
		1pxu	A	392 - 723	Jmol AstexViewer SPICE
		1pxv	A	392 - 723	Jmol AstexViewer SPICE
		1pxw	A	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
		186	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
		100b	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

Pfam: Jm1
<http://pfam.janelia.org/structure/viewer?viewer=jm1&id=3bfm>

Pfam: Family: Kinesin (PF00225) Pfam: Jm1

welcome trust sanger institute

PDB entry 3bfm

Jmol

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

[Close window](#)

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:
[PDF 1 \(bioinformatics review\)](#),
[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)