

INTRODUCTION TO BIOINFORMATICS

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

Barry Grant
University of Michigan
www.thegrantlab.org

BIOINF 525 http://bioboot.github.io/bioinf525_w16/ 12-Jan-2016



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



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



Hongyang Li (GSI)
hyangl@umich.edu

COURSE LOGISTICS

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

Labs: *Session I:* Thursdays 2:30 - 4:00 PM
Session II: Fridays 10:30 - 12:00 PM
Rm. 2036 Palmer Commons

Website: <http://tinyurl.com/bioinf525-w16>
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 on Ctools:
<http://tinyurl.com/bioinf525-w16>
- Complete the **lecture 1.1 homework questions**:
<http://tinyurl.com/bioinf525-quiz1>

Q. What is Bioinformatics?

Q. What is Bioinformatics?

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

[After Orengo, 2003]

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

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

MORE DEFINITIONS

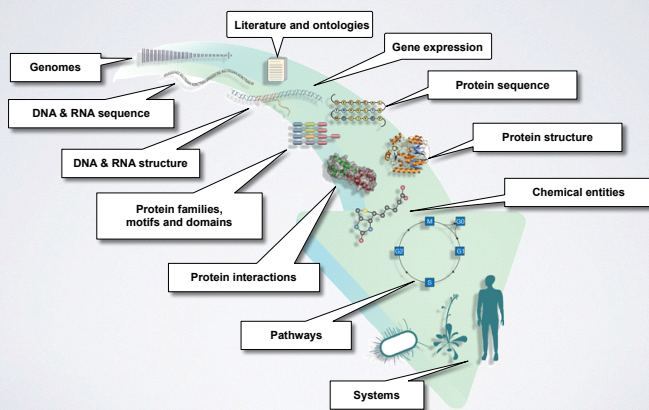
- ▶ "Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying "**informatics**" techniques (derived from disciplines such as applied maths, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a **large-scale**.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

MORE DEFINITIONS

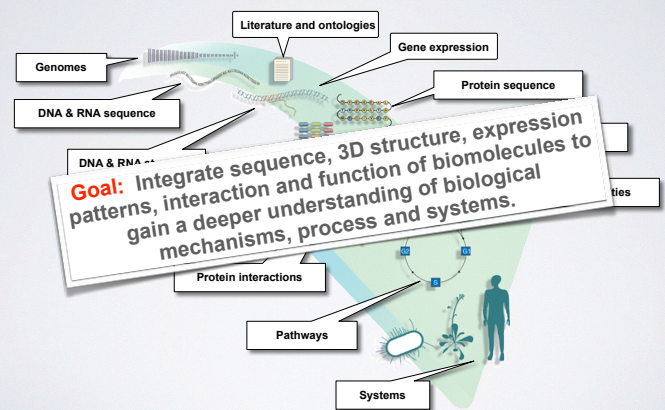
- ▶ "Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying "**informatics**" techniques (derived from disciplines such as applied mathematics, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on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

Key Point: Bioinformatics is Computer Aided Biology

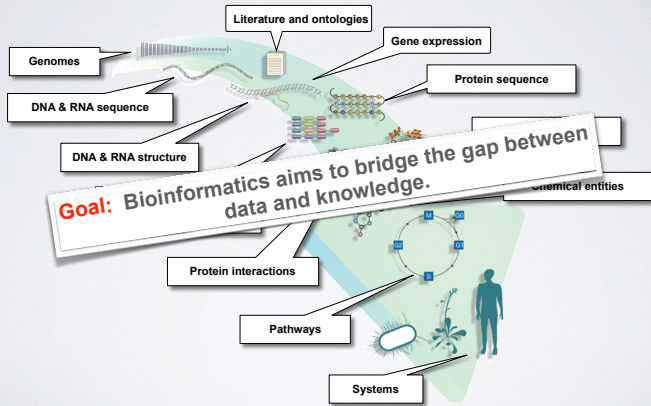
Major types of Bioinformatics Data



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

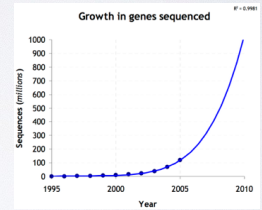
- DNA sequence determines protein sequence.
- Protein sequence determines protein structure.
- Protein structure determines protein function.
- Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.

Bioinformatics is now essential for the archiving, organization and analysis of data related to 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

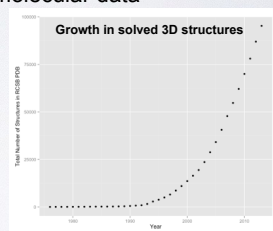


Bioinformatics is essential for the archiving, organization and analysis of data from sequencing, structural genomics, microarrays, proteomics and new high throughput assays.

Why do we need Bioinformatics?

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

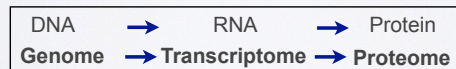
- Bioinformatics provides methods for the efficient:
 - ▶ storage
 - ▶ annotation
 - ▶ search and retrieval
 - ▶ data integration
 - ▶ data mining and analysis



Bioinformatics is essential for the archiving, organization and analysis of data from sequencing, structural genomics, microarrays, proteomics and new high throughput assays.

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?

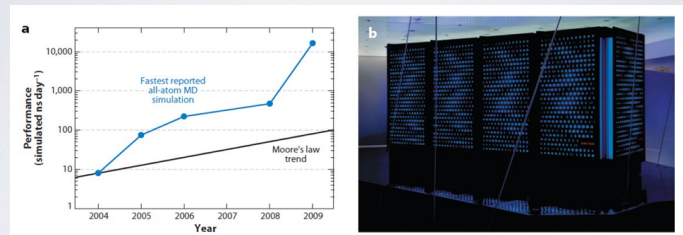
Pre-packaged tools and databases

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

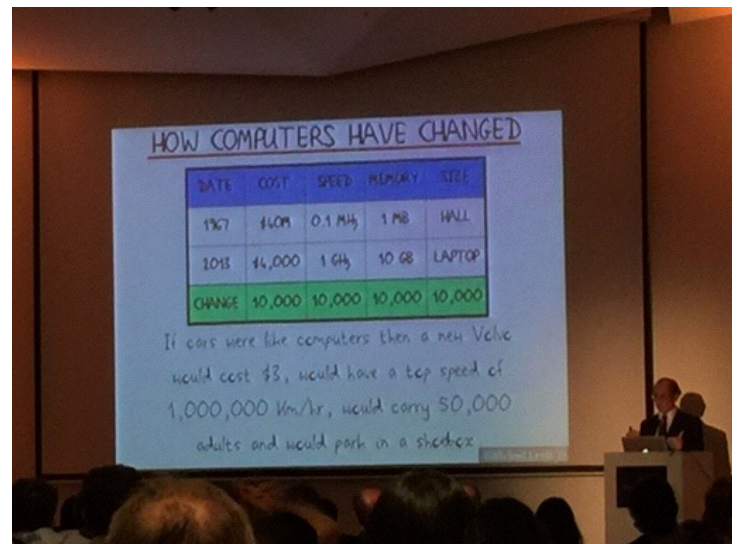
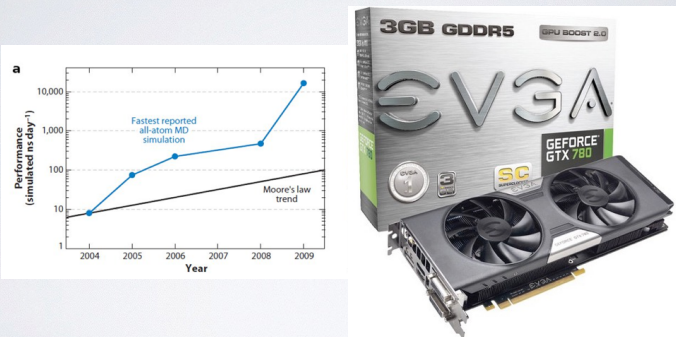
Tool development

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

SIDE-NOTE: SUPERCOMPUTERS AND GPU'S



SIDE-NOTE: SUPERCOMPUTERS AND GPU'S



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

Compositional adjustments: Conditional compositional scoring

Filters and Masking

Filter: Low complexity regions

Mask: Mask for lookup table only
 Mask lower case letters

PSI/PHI/DELTA BLAST

Optional: Choose File (no file selected)

PSI-BLAST Threshold: 0.005

Pseudocount: 0

Even Blast has many settable parameters

Related tools with different terminology

| MATRIX | GAP OPEN | GAP EXTEND | KTUP | EXPECTATION UPPER VALUE | EXPECTATION LOWER VALUE |
|----------|----------|------------|------|-------------------------|-------------------------|
| BLOSUM50 | -10 | 2 | 10 | 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

Key Online Bioinformatics Resources: NCBI & EBI

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

Please take the initial **BIOINF525** questionnaire:
tinyurl.com/bioinf525-questions

<http://www.ncbi.nlm.nih.gov> <https://www.ebi.ac.uk>

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>

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

Bethesda, MD

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

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

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

3D Structures

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

NCBI Announcements

New version of Genome Workbench available

An integrated, downloadable application

Popular Resources

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

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

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | RSS Feeds

Get Started

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

3D Structures

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

NCBI Announcements

New version of Genome Workbench available

An integrated, downloadable application

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

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

Services Research Training About us

Find a gene, protein or chemical:

Examples: Insd, keratin, MY1

Services Research Training Industry European Coordination EMBL ALUMNI

News from EMBL-EBI

Visit EMBL.org

Upcoming events

Plant and Animal Genome conference (PAG XXIV)

Sunday 10 - 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 Nucleic Acids Research*.

Popular

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

Service news

DNA & RNA genes, genomes & variation

Gene expression RNA, protein & metabolite expression

Proteins sequences, families & motifs

Structures Molecular & cellular structures

Systems reactions, interactions & pathways

Chemical biology chemogenomics & metabolomics

Ontologies taxonomies & controlled vocabularies

Literature Scientific publications & patents

Cross domain cross-domain tools & resources

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 Nucleic Acids Research*.

Popular

- Ensembl
- UniProt
- PDB
- ArrayExpress
- ChEMBL

Service news

DNA & RNA genes, genomes & variation

Gene expression RNA, protein & metabolite expression

Proteins sequences, families & motifs

Structures Molecular & cellular structures

Systems reactions, interactions & pathways

Chemical biology chemogenomics & metabolomics

Ontologies taxonomies & controlled vocabularies

Literature Scientific publications & patents

Cross domain cross-domain tools & resources

Training

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

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

Proteins

Popular services

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

Quick links

- Popular services in this category
- All services in this category
- Project websites in this category

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

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

Visit **EMBL.org**

Upcoming events: Plant and Animal Genome conference (PAG XXIV), Sunday 10 - Tuesday 12 January 2016

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

Train online

Using sequence similarity searching tools at EMBL-EBI: webinar

Andrew Cowley

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

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

Train online | EBI Train online

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

Find a course

Browse by subject

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

BIOINFORMATICS DATABASES AND ASSOCIATED TOOLS

What is a database?

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

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

Key database features allow for:

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

Desirable features include:

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

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, BiomagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAR, BSORF, BTKbase, CANSITE, CarbBank, CARB-HYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCF, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlilesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDb, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMG, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEPS, 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, SeqAnalRef, 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, Beanref, BiomagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAR, BSORF, BTKbase, CANSITE, CarbBank, CARB-HYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCF, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlilesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDb, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMG, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEPS, 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, SeqAnalRef, 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 Ctools handout
< [Handout Major Databases.pdf](#) >

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.

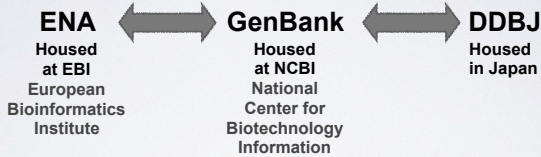
- **Primary databases** (or *archival databases*) consist of data derived experimentally.
 - ▶ **GenBank**: NCBI's primary nucleotide sequence database.
 - ▶ **PDB**: Protein X-ray crystal and NMR structures.
- **Secondary databases** (or *derived databases*) contain information derived from a primary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



- The underlying raw DNA sequences are identical
 - ▶ The different sites provide different views and ways to navigate through the data
- Access to GenBank (and other NCBI databases including RefSeq) is typically through **Entrez**, (the Google of NCBI) [more on this later](#)

GenBank sequence record

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

GenBank sequence record

GenBank sequence record

Can set different display formats here

FASTA sequence record

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.

GenBank 'graphics' sequence record

GenBank sequence record, cont.

GenBank sequence record, cont.

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

GenBank sequence record, cont.

The actual sequence entry starts after the word ORIGIN

RefSeq: NCBI's Derivative Sequence Database

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

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

UNIPROT: THE PREMIER PROTEIN SEQUENCE DATABASE

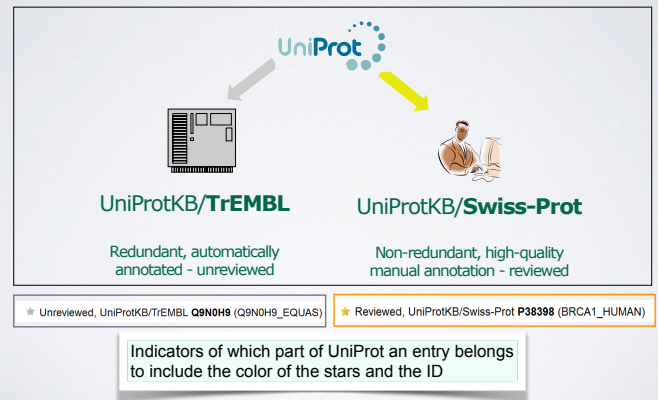
UniProt: Protein sequence database

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

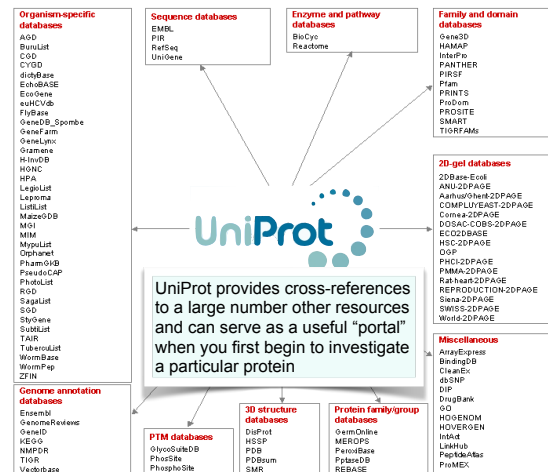
- UniProt comprises four databases:

- 1. UniProtKB** (Knowledgebase)
 - Containing **Swiss-Prot** and **TrEMBL** components (these correspond to hand curated and automatically annotated entries respectively)
- 2. UniRef** (Reference Clusters)
 - Filtered version of UniProtKB at various levels of sequence identity
 - e.g. **UniRef90** contains sequences with a maximum of 90% sequence identity to each other
- 3. UniParc** (Archive) with database cross-references to source.
- 4. UniMES** (Metagenomic and Environmental Sequences)

The two sides of UniProtKB



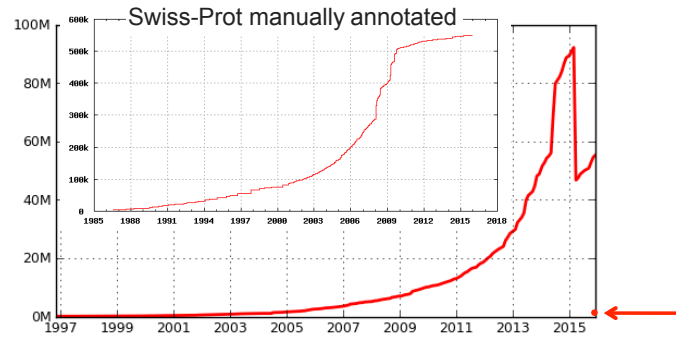
The main information added to a UniProt/Swiss-Prot entry



UniProt/Swiss-Prot vs UniProt/TrEMBL

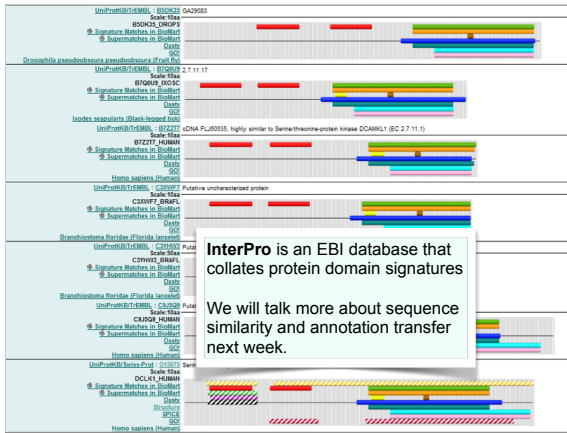
- *UniProtKB/Swiss-Prot* is a **non-redundant** database with one entry per protein
- *UniProtKB/TrEMBL* is a **redundant** database with one entry per translated ENA entry (ENA is the EBI's equivalent of GenBank)
 - Therefore TrEMBL can contain multiple entries for the same protein
 - Multiple UniProtKB/TrEMBL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



73

Same domain composition = same function = annotation transfer



74

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 |
|----------------|-----------|--|
| Literature | | |
| Books | 1,677 | books and reports |
| MeSH | 402 | ontology used for PubMed indexing |
| NLM Catalog | 223 | books, journals and more in the NLM Collections |
| PubMed | 54,672 | scientific & medical abstracts/citations |
| PubMed Central | 96,114 | full-text journal articles |
| Health | | |
| ClinVar | 759 | human variations of clinical significance |
| dbGaP | 120 | genotype/phenotype interaction studies |
| GTR | 1,879 | genetic testing registry |
| Genes | | |
| EST | 3,965 | expressed sequence tag sequences |
| Gene | 87,165 | collected information about gene loci |
| GEO DataSets | 3,732 | functional genomics studies |
| GEO Profiles | 1,622,789 | gene expression and molecular abundance profiles |
| HomoloGene | 696 | homologous gene sets for selected organisms |
| PopSet | 2,254 | sequence sets from phylogenetic and population studies |
| UniGene | 4,770 | clusters of expressed transcripts |
| Proteins | | |

77

NCBI Resources How To Sign in to NCBI

Gene

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

Clear all

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

Filters: Manage Filters

Top Organisms [Tree]

- Homo sapiens (1126)
- Mus musculus (823)
- Rattus norvegicus (525)
- Oreochromis niloticus (533)
- Neolamprologus brichardi (507)
- All other taxa (82019)
- More...

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

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

Find related data
Database:

Search details
[ras][All Fields] AND alive[property]

78

NCBI Resources How To Sign in to NCBI

Gene

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

Clear all

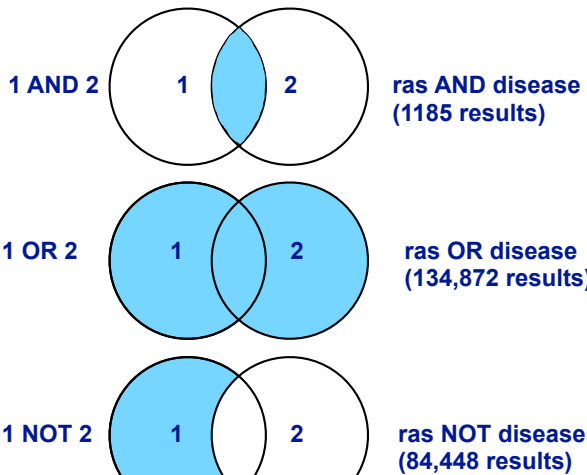
Results: 1 to 20 of 1126
Filters activated: Current only. Clear all to show 1499 items.

| Name/Gene ID | Description | Location | Aliases |
|------------------|---|---|---|
| NRAS ID: 4993 | neuroblastoma RAS viral (v-ras) oncogene homolog [Homo sapiens (human)] | Chromosome 1, NC_000011.11 (114704464..114716894, complement) | RPS, 1000E10.2, ALP54, CMNS, N-ras, NCMS1, NS6, NRAS |
| KRAS ID: 3845 | Kirsten rat sarcoma viral oncogene homolog [Homo sapiens (human)] | Chromosome 12, NC_000012.12 (25205246..25250923, complement) | C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS, N-ras |

Find related data
Database:

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

79



NCBI Resources How To Sign in to NCBI

Gene

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

Clear all

Results: 1 to 20 of 1126
Filters activated: Current only. Clear all to show 1499 items.

| Name/Gene ID | Description | Location | Aliases |
|------------------|---|---|---|
| NRAS ID: 4993 | neuroblastoma RAS viral (v-ras) oncogene homolog [Homo sapiens (human)] | Chromosome 1, NC_000011.11 (114704464..114716894, complement) | RPS, 1000E10.2, ALP54, CMNS, N-ras, NCMS1, NS6, NRAS |
| KRAS ID: 3845 | Kirsten rat sarcoma viral oncogene homolog [Homo sapiens (human)] | Chromosome 12, NC_000012.12 (25205246..25250923, complement) | C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, K-RAS1, KRAS2, NS, N-ras |

Find related data
Database:

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

81

NCBI Resources How To Sign in to NCBI

Gene

Display Settings: Full Report

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

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

Summary

Official Symbol KRAS provided by HSNCG
Official Full Name Kirsten rat sarcoma viral oncogene homolog provided by HSNCG
Primary source HGNC:HGNC:6407
See related Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193

Gene type protein coding
RefSeq status REVIEWED
Organism *Homo sapiens*
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Hominidae; Homo

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

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

82

NCBI Resources How To Sign in to NCBI

Gene

Display Settings: Full Report

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

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

Summary

Official Symbol KRAS provided by HSNCG
Official Full Name Kirsten rat sarcoma viral oncogene homolog provided by HSNCG
Primary source HGNC:HGNC:6407
See related Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193

Gene type protein coding
RefSeq status REVIEWED
Organism *Homo sapiens*
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Hominidae; Homo

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

Table of contents
Summary
Genomic context

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

83

Genomic context

Location: 12p12.1
Exon count: 6

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

Chromosome 12 - NC_000012.12

Genomic regions, transcripts, and products

84

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

Gene
KRAS (human)

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

85

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 |
|---------------------------------------|---------------|------|
| Fc-epsilon receptor signaling pathway | TAS | |
| GTP catabolic process | IEA | |
| MAPK cascade | TAS | |
| Ras protein signal transduction | TAS | |
| actin cytoskeleton organization | IEA | |
| activation of MAPKK activity | TAS | |
| axon guidance | TAS | |
| mRNA processing | TAS | |

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 biocuration. UniProt manual and electronic GO annotations are supplemented with manual annotations supplied by external collaborating GO Consortium groups, to ensure a comprehensive GO annotation dataset is supplied to users.

UniProt is a member of the GO Consortium.

87

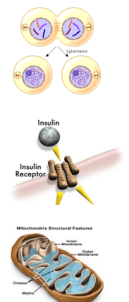
Why do we need Ontologies?

- 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

88

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



89

Gene Ontology Provided by GOA

Function

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

Process

- Fc-epsilon receptor signaling pathway
- GTP catabolic process
- MAPK cascade
- Ras protein signal transduction
- actin cytoskeleton organization
- activation of MAPKK activity
- axon guidance
- Nuclei organization

Evidence Code

- TAS
- IEA
- TAS
- TAS
- IEA
- 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

UniProtKB Link
UniProtKB/Swiss-Prot:P01116

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)

Function

Ras proteins bind GDP/GTP and possess intrinsic GTPase activity. Plays an important role in the regulation of cell proliferation (PubMed:23698361, PubMed:22711838). #2 Publications - @ Curated

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 ¹ | 10 - 18 | 9 | GTP @ 2 Publications | | | |
| Nucleotide binding ¹ | 29 - 35 | 7 | GTP @ 2 Publications | | | |
| Nucleotide binding ¹ | 59 - 60 | 2 | GTP @ 2 Publications | | | |

Display

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

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

Protein: GTPase KRas
Gene: KRAS
Organism: Homo sapiens (Human)

Function

Ras proteins bind GDP/GTP and possess intrinsic GTPase activity. Plays an important role in the regulation of cell proliferation (PubMed:23698361, PubMed:22711838). #2 Publications - @ Curated

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 ¹ | 10 - 18 | 9 | GTP @ 2 Publications | | | |
| Nucleotide binding ¹ | 29 - 35 | 7 | GTP @ 2 Publications | | | |
| Nucleotide binding ¹ | 59 - 60 | 2 | GTP @ 2 Publications | | | |

Display

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

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

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

Feature key

| Feature key | Position(s) | Length | Description | Graphical view | Feature identifier | Actions |
|------------------------------|-------------|--------|---|----------------|--------------------|---------|
| Natural variant ¹ | 10 - 10 | 1 | G - GG in one individual with AML; expression in 3T3 cell causes cellular transformation; expression in CCD5 cells activates the Ras-MAPK signaling pathway; lower GTPase activity; faster GDP dissociation rate. #1 Publications | | VAR_034601 | |

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

NOONAN SYNDROME 3 (NS3) [MIM:609942]: A form of Noonan syndrome, a disease characterized by short stature, facial dysmorphic features such as hypertelorism, a downward slanting and low-set posteriorly rotated ears, and a high incidence of congenital heart

Display

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

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

Structure

Secondary structure

Legend: Helix Turn Beta strand

3D structure databases

| Entry | Method | Resolution (Å) | Chain | Positions | PDBsum |
|-------|--------|----------------|-------------|-----------|--------|
| 10BD | X-ray | 2.00 | P | 178-188 | [+] |
| 10BE | X-ray | 3.00 | P | 178-188 | [+] |
| 1KZO | X-ray | 2.20 | C | 169-173 | [+] |
| 1KZP | X-ray | 2.10 | C | 169-173 | [+] |
| 3GFT | X-ray | 2.27 | A/B/C/D/E/F | 1-164 | [+] |
| 4DSN | X-ray | 2.03 | A | 2-164 | [+] |
| 4D50 | X-ray | 1.85 | A | 2-164 | [+] |
| 4EPR | X-ray | 2.00 | A | 1-164 | [+] |
| 4EPT | X-ray | 2.00 | A | 1-164 | [+] |
| 4EPV | X-ray | 1.35 | A | 1-164 | [+] |
| 4EPW | X-ray | 1.70 | A | 1-164 | [+] |
| 4EPX | X-ray | 1.76 | A | 1-164 | [+] |
| 4EPY | X-ray | 1.80 | A | 1-164 | [+] |
| 4L8G | X-ray | 1.52 | A | 1-169 | [+] |
| 4LDJ | X-ray | 1.15 | A | 1-164 | [+] |
| 4LPK | X-ray | 1.50 | A/B | 1-169 | [+] |

Display

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

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... ?

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... ?

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

Family: Kinesin (PF00225)

HHMI janelia farm research campus

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

keyword search

126 architectures 4150 sequences 6 interactions 248 species 114 structures

Family: Kinesin (PF00225)

Summary

Domain organisation

Alignments

HMM logo

Trees

Curations & models

Species

Interactions

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 |
|---------------|------------------|--------|--------------|--------------|--|
| ARBK01_GIALA | 11 - 335 | 2vva | A | 11 - 335 | Jmol AstexViewer SPICE |
| | | | B | 11 - 335 | Jmol AstexViewer SPICE |
| CENPL_HUMAN | 12 - 329 | 155c | A | 12 - 329 | Jmol AstexViewer SPICE |
| | | | B | 12 - 329 | Jmol AstexViewer SPICE |
| KAR3_YEAST | 392 - 723 | 191v | A | 392 - 723 | Jmol AstexViewer SPICE |
| | | | A | 392 - 723 | Jmol AstexViewer SPICE |
| | | | A | 392 - 723 | Jmol AstexViewer SPICE |
| | | | A | 392 - 723 | Jmol AstexViewer SPICE |
| K113B_HUMAN | 11 - 352 | 3qbi | A | 11 - 352 | Jmol AstexViewer SPICE |
| | | | B | 11 - 352 | Jmol AstexViewer SPICE |
| | | | C | 11 - 352 | Jmol AstexViewer SPICE |
| | | | A | 24 - 359 | Jmol AstexViewer SPICE |
| | | 1i6 | B | 24 - 359 | Jmol AstexViewer SPICE |
| | | | A | 24 - 359 | Jmol AstexViewer SPICE |
| | | | B | 24 - 359 | Jmol AstexViewer SPICE |
| | | | A | 24 - 359 | Jmol AstexViewer SPICE |
| | | 1a0b | B | 24 - 359 | Jmol AstexViewer SPICE |
| | | | A | 24 - 359 | Jmol AstexViewer SPICE |
| | | | B | 24 - 359 | Jmol AstexViewer SPICE |
| | | | A | 24 - 359 | Jmol AstexViewer SPICE |
| | | 1x88 | B | 24 - 359 | Jmol AstexViewer SPICE |
| | | | A | 24 - 359 | Jmol AstexViewer SPICE |
| | | | A | 24 - 359 | Jmol AstexViewer SPICE |
| | | | A | 24 - 359 | Jmol AstexViewer SPICE |

Family: Kinesin (PF00225)

HHMI janelia farm research campus

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

keyword search

126 architectures 4150 sequences 6 interactions 248 species 114 structures

Structure viewer: Jmol

PDB entry 3bfn

Jmol

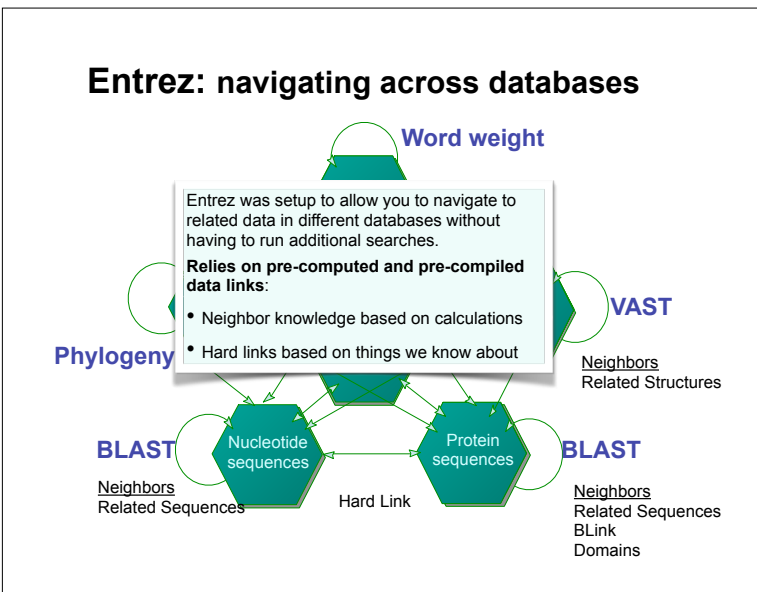
| Chain | PDB | Start | End | UniProt | Start | End | Pfam family | Colour |
|-------|-----|-------|-----|-------------|-------|-----|-------------------|--------|
| A | 49 | 368 | | KIF22_HUMAN | 49 | 368 | Kinesin (PF00225) | |

Close window

ENTREZ & BLAST:
TOOLS FOR SEARCHING AND ACCESSING MOLECULAR DATA AT NCBI

Entrez: Integrated search of NCBI databases

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



Global Entrez Query: All NCBI Databases

Search NCBI databases

ras

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

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

The Entrez system: 38 (and counting) integrated databases

| Database | Count | Description |
|----------------|--------|--|
| Literature | 3,985 | ontology used for PubMed indexing |
| MeSH | 402 | ontology used for PubMed indexing |
| NLM Catalog | 223 | ontology used for PubMed indexing |
| PubMed | 54,672 | full-text journal articles |
| 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 |
| EST | 3,985 | sequences |
| HomoloGene | 696 | homologous gene sets for selected organisms |
| PopSet | 2,254 | sequence sets from phylogenetic and population studies |
| UniGene | 4,770 | clusters of expressed transcripts |
| Proteins | | |

Search Results

Discovery Column (sort, filter, link)

1. **Danio rerio creatine kinase, muscle b (ckmb), mRNA**
 1,463 bp linear mRNA
 Accession: NM_001105683.1 GI: 157787190
 GenBank FASTA Graphics Related Sequences

Limits

Search Results

Discovery Column (sort, filter, link)

1. **Danio rerio creatine kinase, muscle b (ckmb), mRNA**
 1,463 bp linear mRNA
 Accession: NM_001105683.1 GI: 157787190
 GenBank FASTA Graphics Related Sequences

Advanced: Search Builder

Helps build complex fielded queries

Items from search history can be included / combined / modified

| Search | Add to builder | Query | Items found | Time |
|--------|----------------|---|-------------|----------|
| #7 | Add | Search zebrafish[organism] AND actin[title] | 71 | 12:41:16 |
| #8 | Add | Search zebrafish actin | 1288 | 12:40:07 |
| #1 | Add | Search zebrafish creatine kinase | 34 | 12:39:02 |

Complex Query Results

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

1. **Danio rerio creatine kinase, brain s (ckba), mRNA**
 1,481 bp linear mRNA
 Accession: NM_001077163.1 GI: 116004538
 GenBank FASTA Graphics Related Sequences

Controlled Vocabularies

– Taxonomy primary controlled vocabulary / classification system for molecular databases at NCBI

Search details
 "Porifera"[Organism]
 OR sponges"[All Fields]

► Medical Subject Headings (MeSH) primary controlled vocabulary / classification system (ontology) for molecular databases at NCBI

Search details
 "porifera"[MeSH Terms]
 OR "porifera"[All Fields]
 OR "sponges"[All Fields]

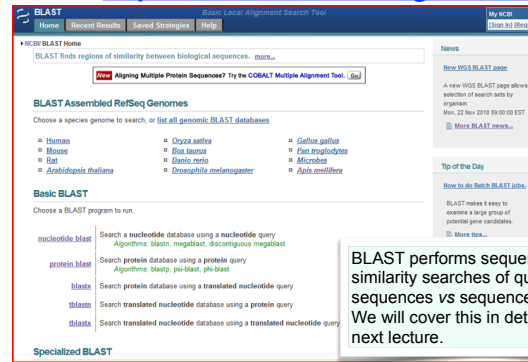
BLAST is a very important tool available from the NCBI Homepage

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



BLAST – Basic Local Alignment Search Tool

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



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 on Ctools:
<http://tinyurl.com/bioinf525-w16>
- ✓ Complete the **lecture 1.1 homework questions**:
<http://tinyurl.com/bioinf525-quiz1>

THANK YOU

ADDITIONAL DATABASES OF NOTE (SLIDES FOR YOUR REFERENCE)

NCBI Metadatabases

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

PubMed

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

120

PubMed results

Limits and Advanced search can be used to refine searches

Small molecule databases have been added at NCBI
<http://pubchem.ncbi.nlm.nih.gov/>

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

Online Mendelian Inheritance in Man – OMIM

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

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

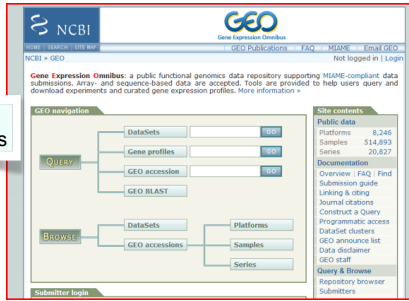
The NCBI Bookshelf includes many well known molecular biology texts.

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

GEO: Gene Expression Omnibus

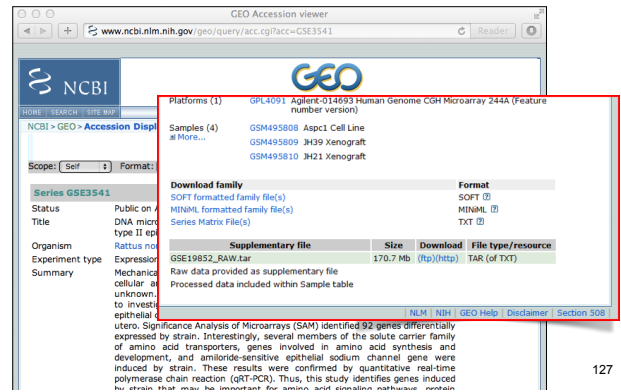
- Gene expression data (mostly from microarrays but also RNA-seq data, 2 methods for measuring RNA levels)

Query browse and download data sets



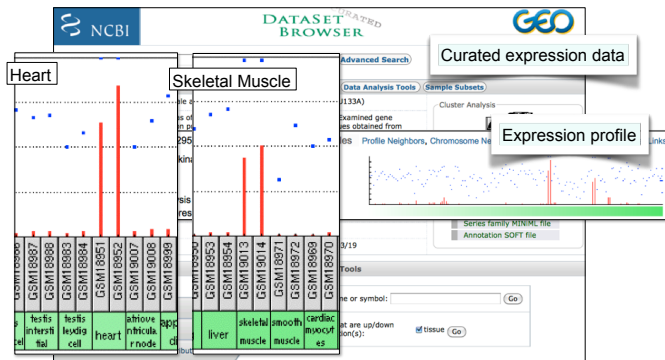
126

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



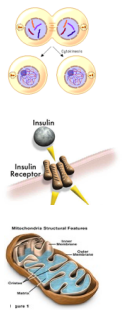
127

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



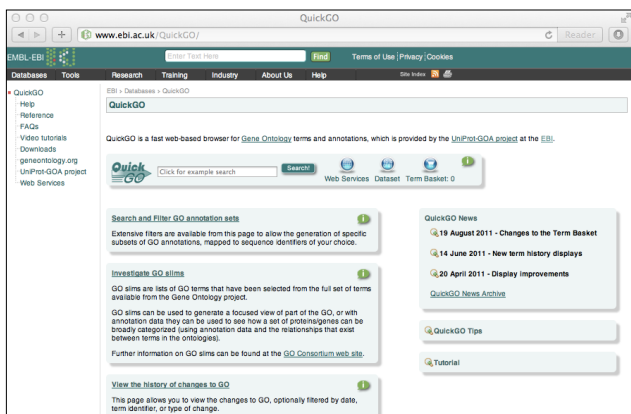
GO Ontologies

- 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



129

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



30

GO annotation in UniProt

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

