

BGGN 213

Foundations of Bioinformatics

Barry Grant
UC San Diego

<http://thegrantlab.org/bggn213>

HELLO
my name is

BARRY

bjgrant@ucsd.edu

HELLO
HER — my name is

ILEENA

ileenamitra@eng.ucsd.edu

Introduce Yourself!

Your preferred name,
Place you identify with,
Major area of study/research,
Favorite joke (optional)!

Today's Menu

Course Logistics

Website, screencasts, survey, ethics, assessment and grading.

Learning Objectives

What you need to learn to succeed in this course.

Course Structure

Major lecture topics and specific learning goals.

Introduction to Bioinformatics

Introducing the *what, why and how* of bioinformatics?

Computer Setup

Ensuring your laptop is all set for future sections of this course.

<http://thegrantlab.org/bggn213/>

The screenshot shows a web browser window displaying the course website. The URL in the address bar is bioboot.github.io/bggn213_f17/. The page title is "Foundations of Bioinformatics (BGGN 213, Fall 2017)". On the left, there is a sidebar with the UC San Diego logo and links for Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, and Screen Cast Videos. Below these are social media icons for Twitter, GitHub, Email, and RSS. The main content area features the course title, course director information (Prof. Barry J. Grant), instructional assistant (Ileena Mitra), course syllabus (Fall 2017 PDF), and an overview section. The overview text discusses the driving force of bioinformatics in biosciences and its design for bioscience graduate students. Major topics include genomic and biomolecular bioinformatic resources.

UC San Diego

BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

Overview

Lectures

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

Screen Cast Videos

[Twitter](#) [GitHub](#) [Email](#) [RSS](#)

bioboot.github.io/bggn213_f17/

Home Gmail Gcal Bitbucket GitHub News Disqus

Foundations of Bioinformatics (BGGN 213, Fall 2017)



Course Director
Prof. Barry J. Grant (Email: bjgrant@ucsd.edu)

Instructional Assistant
Ileena Mitra (Email: ileenamitra@eng.ucsd.edu)

Course Syllabus
[Fall 2017 \(PDF\)](#)

Overview

Bioinformatics is driving the collection and analysis of big data in the biosciences. This course is designed for bioscience graduate students and provides a hands-on introduction to the computer-based analysis of genomic and biomolecular data.

Major topics include:

- Genomic and biomolecular bioinformatic resources,
- ...

<http://thegrantlab.org/bggn213/>

The screenshot shows a web browser window with the URL bioboot.github.io/bggn213_f17/ in the address bar. The page content is as follows:

UC San Diego
BGGN 213

A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

Overview
Lectures
Computer Setup
Learning Goals (highlighted with a red box)
Assignments & Grading
Ethics Code
Screen Cast Videos

Social media icons for Twitter, GitHub, Email, and RSS feed are at the bottom.

**Foundations of Bioinformatics
(BGGN 213, Fall 2017)**

Course Director
Prof. Barry J. Grant (Email: bjgrant@ucsd.edu)

Instructional Assistant
Ileena Mitra (Email: ileenamitra@eng.ucsd.edu)

Course Syllabus
[Fall 2017 \(PDF\)](#)

Overview

Bioinformatics is driving the collection and analysis of big data in the biosciences. This course is designed for bioscience graduate students and provides a hands-on introduction to the computer-based analysis of genomic and biomolecular data.

Major topics include:

- Genomic and biomolecular bioinformatic resources,
- Genome informatics

What essential concepts and skills should YOU attain from this course?

The screenshot shows a web browser window with the URL bioboot.github.io/bggm213_f17/goals/. The page is titled "Learning Goals". On the left, there is a sidebar for the course BGGN 213, which includes links for Overview, Lectures, Computer Setup, Learning Goals (which is highlighted with a red box), Assignments & Grading, Ethics Code, and Screen Cast Videos. Below the sidebar are social media sharing icons for Twitter, Facebook, Email, and RSS. The main content area contains the title "Learning Goals" and a list of learning objectives. At the bottom, a summary statement is provided.

Learning Goals

At the end of this course students will:

- Understand the increasing necessity for computation in modern life sciences research.
- Be able to use and evaluate online bioinformatics resources including major biomolecular and genomic databases, search and analysis tools, genome browsers, structure viewers, and select quality control and analysis tools to solve problems in the biological sciences.
- Be able to use the UNIX command line and the R environment to analyze bioinformatics data at scale.
- Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.
- Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genomics, Transcriptomics and Structural bioinformatics.

In short, students will develop a solid foundational knowledge of bioinformatics and be able to evaluate new biomolecular and genomic information using existing bioinformatic tools and resources.

At the end of this course students will:

- Understand the increasing necessity for computation in modern life sciences research.
- Be able to use and evaluate online bioinformatics resources and analysis tools to solve problems in the biological sciences.
- Be able to use the UNIX command line and the R environment to analyze bioinformatics data at scale.
- Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genome informatics, Transcriptomics and Structural informatics.

In short, you will develop a solid foundational knowledge of **bioinformatics** and be able to evaluate new biomolecular and genomic information using **existing bioinformatic tools and resources.**

Specific Learning Goals....

What I want you to know by course end!

The screenshot shows a web browser window with the URL bioboot.github.io/bggm213_f17/goals/. The page title is "Specific Learning Goals". The main text states: "Teaching toward the specific learning goals below is expected to occupy 60%-70% of class time. The remaining course content is at the discretion of the instructor with student body input. This includes student selected topics for peer presentation as well one student selected guest lecture from an industry based genomic scientist." Below this, a list of learning goals is provided, each mapped to specific lectures.

	Specific Learning Goals	Lecture(s):
1	Appreciate and describe in general terms the role of computation in hypothesis-driven discovery processes within the life sciences.	1, 2, 20
2	Be able to query, search, compare and contrast the data contained in major bioinformatics databases and describe how these databases intersect (GenBank, GENE, UniProt, PFAM, OMIM, PDB, UCSC, ENSEMBLE).	2, 12, 13
3	Describe how nucleotide and protein sequence and structure data are represented (FASTA, FASTQ, GenBank, UniProt, PDB).	3, 10
4	Be able to describe how dynamic programming works for pairwise sequence alignment and appreciate the differences	4, 5

Course Structure

Derived from specific learning goals

The screenshot shows a web browser window with the following details:

- Address Bar:** bioboot.github.io/bggn213_f17/lectures/
- Navigation:** Back, Forward, Stop, Refresh, Home, Gmail, Gcal, Bitbucket, GitHub, News, Disqus, and a Plus icon.
- Content Area:**
 - Section Header:** Lectures
 - Text:** All Lectures are Tu/Th 9:00-12:00 pm in Warren Lecture Hall 2015 (WLH 2015) ([Map](#)). Clicking on the class topics below will take you to corresponding lecture notes, homework assignments, pre-class video screen-casts and required reading material.
 - Table:** A table titled "Topics for Fall 2017" with columns for #, Date, and Topics. It lists two lectures:
 - Lecture 1:** Th, 09/28 - [Welcome to Foundations of Bioinformatics](#). Description: Course introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Student computer setup.
 - Lecture 2:** Tu, 10/03 - [Bioinformatics databases and key online resources](#). Description: NCBI & EBI resources for the molecular domain of bioinformatics, Focus on GenBank, UniProt, Entrez and Gene Ontology. Hands on with BLAST, GenBank, OMIM, GENE, UniProt, Muscle, PFAM and PDB bioinformatics tools and databases.

Course Structure

Derived from specific learning goals

The screenshot shows a web browser window with the URL bioboot.github.io/bggn213_f17/lectures/. The page title is "Lectures". The left sidebar of the website has a red box around the "Lectures" link. The main content area describes the lecture schedule and topics.

All Lectures are Tu/Th 9:00-12:00 pm in Warren Lecture Hall 2015 (WLH 2015) ([Map](#)). Clicking on the class topics below will take you to corresponding lecture notes, homework assignments, pre-class video screen-casts and required reading material.

#	Date	Topics for Fall 2017
1	Th, 09/28	Welcome to Foundations of Bioinformatics Course introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Student computer setup
2	Tu, 10/03	Bioinformatics databases and key online resources NCBI & EBI resources for the molecular domain of bioinformatics, Focus on GenBank, UniProt, Entrez and Gene Ontology. Hands on with BLAST, GenBank, OMIM, GENE, UniProt, Muscle, PFAM and PDB bioinformatics tools and databases

Class Details

Goals, Class material, Screencasts & **Homework**

The screenshot shows a web browser window with the URL bioboot.github.io/bggn213_f17/lectures/#1. The page content is as follows:

UC San Diego
BGGN 213
A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

Overview
Lectures
Computer Setup
Learning Goals
Assignments & Grading
Ethics Code
Screen Cast Videos

1: Welcome to Foundations of Bioinformatics

Topics:
Course introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Student 30-second introductions, Student computer setup.

Goals:

- Understand course scope, expectations, logistics and [ethics code](#).
- Understand the increasing necessity for computation in modern life sciences research.
- Get introduced to how bioinformatics is practiced.
- Complete the [pre-course questionnaire](#).
- Setup your [laptop computer](#) for this course.

Material:

- [Pre class screen cast](#),
- Lecture Slides: Large PDF, [Small PDF](#), (To be updated!)
- [Handout: Class Syllabus](#)
- [Computer Setup Instructions](#).

Homework

Goals, Class material, Screencasts & **Homework**

The screenshot shows a web browser window with the following details:

- URL:** bioboot.github.io/bggn213_f17/lectures/#1
- Page Content:**
 - UC San Diego Logo:** On the left side of the main content area.
 - BGGN 213 Section:** A large heading "BGGN 213" with a subtext: "A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD".
 - Navigation:** A sidebar on the left lists course sections: Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, and Screen Cast Videos.
 - Homework Section:** A heading "Homework:" followed by a bulleted list:
 - [Questions](#),
 - Readings:
 - PDF1: [What is bioinformatics? An introduction and overview](#),
 - PDF2: [Advancements and Challenges in Computational Biology](#),
 - Other: [For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights](#) New York Times, 2014.
 - Screen Casts Section:** A heading "Screen Casts:" followed by a video thumbnail:
 - **Welcome to "Foundations of Bioinformatics" (BGGN-213)**: A video thumbnail featuring Barry Grant, a man with short brown hair, wearing a dark shirt, standing in front of a background of colorful, glowing spheres. The video title includes "BGGN 213 Foundations of Bioinformatics" and "Barry Grant UC San Diego". Below the video is the URL <http://thegrantlab.org/bggn213>.

Homework

Goals, Class material, Screencasts & **Homework**

The screenshot shows a web browser window with the following content:

- UC San Diego BGGN 213 Course Page:** On the left, there's a sidebar with links: Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, and Screen Cast Videos.
- Homework Section:** The main content area has a heading "Homework:" followed by a bulleted list:
 - [Questions](#), (The "Questions" link is highlighted with a red box.)
 - Readings:
 - PDF1: [What is bioinformatics? An introduction and overview](#),
 - PDF2: [Advancements and Challenges in Computational Biology](#),
 - Other: [For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights](#) New York Times, 2014.
- Screen Casts:** Below the homework section, there's a video player showing a screen cast titled "Welcome to ‘Foundations of Bioinformatics’ (BGGN-21...)" featuring Barry Grant from UC San Diego. The video thumbnail includes the text "BGGN 213 Foundations of Bioinformatics" and "Barry Grant UC San Diego".

Homework

Goals, Class material, Screencasts & **Homework**

The screenshot shows a Google Forms survey window. At the top, the URL is `docs.google.com/forms/d/e/1FAIpQLSeN3pg-AaRg5la3PxZuqSj`. Below the URL, there are navigation icons and links to Home, Gmail, Gcal, Bitbucket, GitHub, News, and Disqus. The main title of the form is "BGGN213 Lecture 1 Homework (F17)". A sub-instruction says "Please answer the following questions". A red asterisk indicates a required field: "* Required". The first question asks for "Your UCSD username/email address *". It provides a hint: "The first part of your UCSD email address before the '@ucsd.edu' part". A text input field is labeled "Your answer". The second question is a multiple-choice question: "Which of the following operating systems is most frequently used for bioinformatics tool development?". The options are: Windows (radio button), iOS (radio button), Unix (radio button), and Perl (radio button). The bottom of the page shows a footer with "1 of 1 pages" and "1 question answered".

BGGN213 Lecture 1 Homework (F17)

Please answer the following questions

* Required

Your UCSD username/email address *

The first part of your UCSD email address before the '@[ucsd.edu](#)' part

Your answer

Which of the following operating systems is most frequently used for bioinformatics tool development?

- Windows
- iOS
- Unix
- Perl

1 of 1 pages

1 question answered

Homework

Goals, Class material, Screencasts & **Homework**

docs.google.com/forms/d/e/1FAIpQLSeN3pg-AaRg5la3PxZuqSj

Home Gmail Gcal Bitbucket GitHub News Disqus

BGGN213 Lecture 1 Homework

Please answer the following questions

* Required

Name/Email address *

Part of your UCSD email address before the '@ucsd.edu' part

Your answer

Which of the following operating systems is most frequently used for bioinformatics tool development

- Windows
- ios
- Unix
- Perl

Homework is due before the next weeks class!

Today's Menu

Course Logistics

Website, screencasts, survey, ethics, assessment and grading.

Learning Objectives

What you need to learn to succeed in this course.

Course Structure

Major lecture topics and specific learning goals.

Introduction to Bioinformatics

Introducing the *what, why* and *how* of bioinformatics?

Computer Setup

Ensuring your laptop is all set for future sections of this course.

OUTLINE

Overview of bioinformatics

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

Online 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

- How-to productively navigate major databases.

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

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!**

Q. What is Bioinformatics?

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

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

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

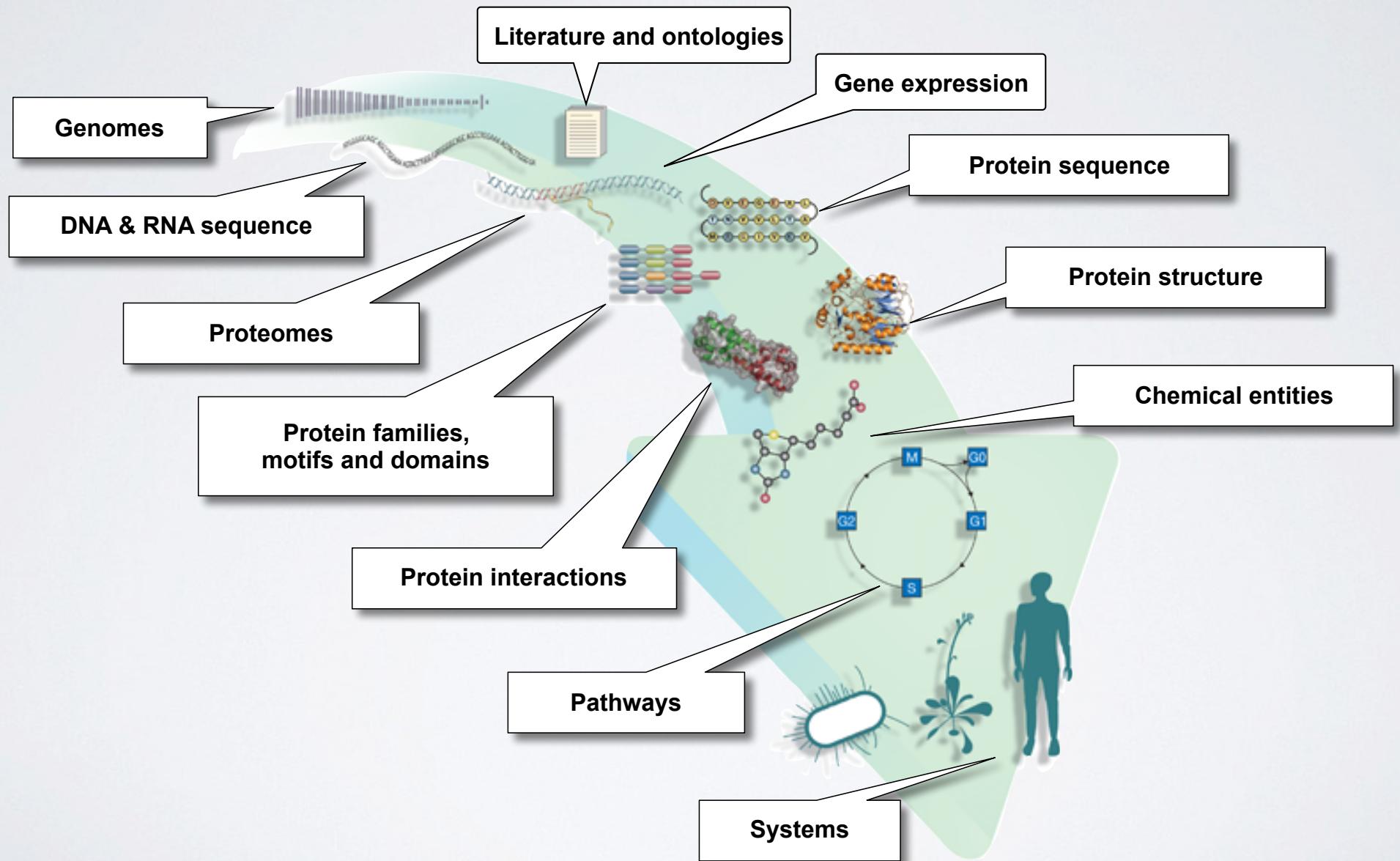
MORE DEFINITIONS

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

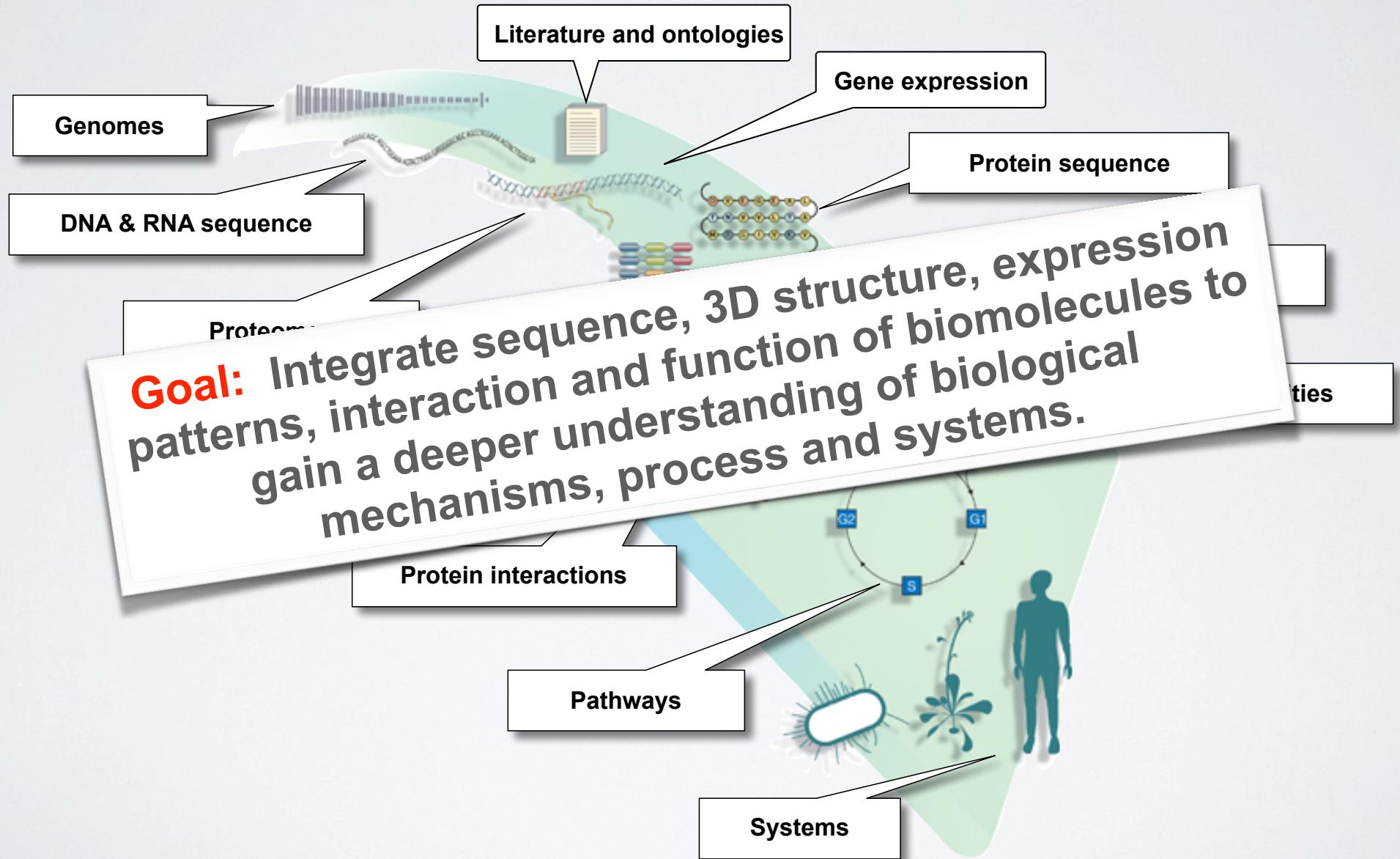
MORE DEFINITIONS

- ▶ “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 **analyze** the information associated with these molecules, on a **large-scale**.
Luscombe NM, et al. Methods 2001;40:346.
- ▶ “Bioinformatics is the search, development, or application of computing 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>)

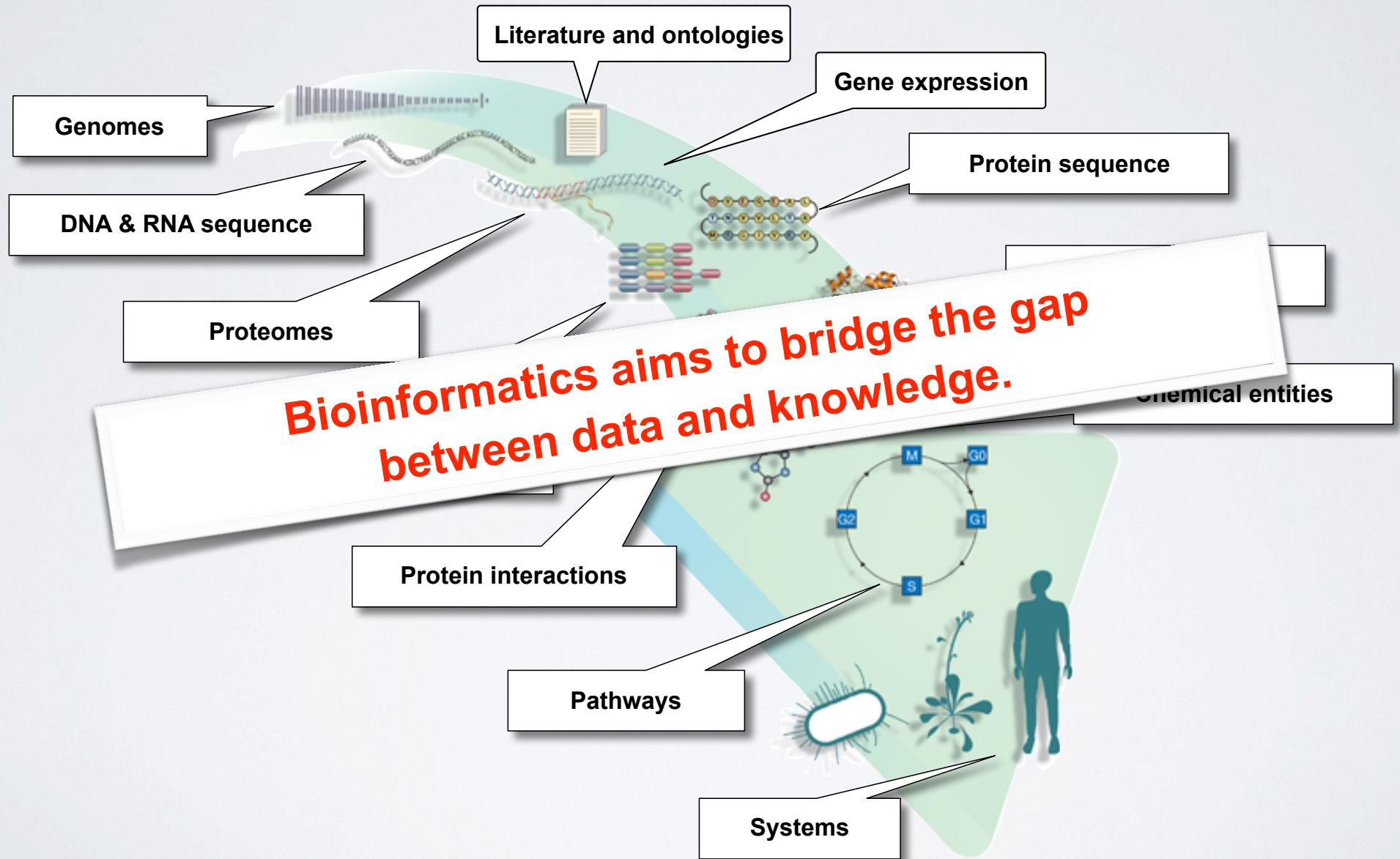
Major types of Bioinformatics Data



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

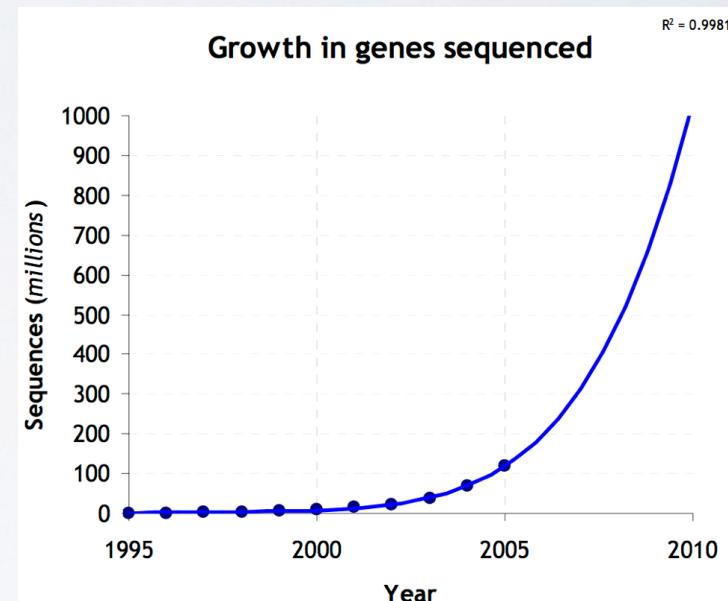
- DNA sequence determines *protein* sequence.
- *Protein* sequence determines *protein* structure.
- *Protein* structure determines *protein* function.
- *Regulatory mechanisms* (e.g. gene expression) determine the amount of a particular *function in space and time*.

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**

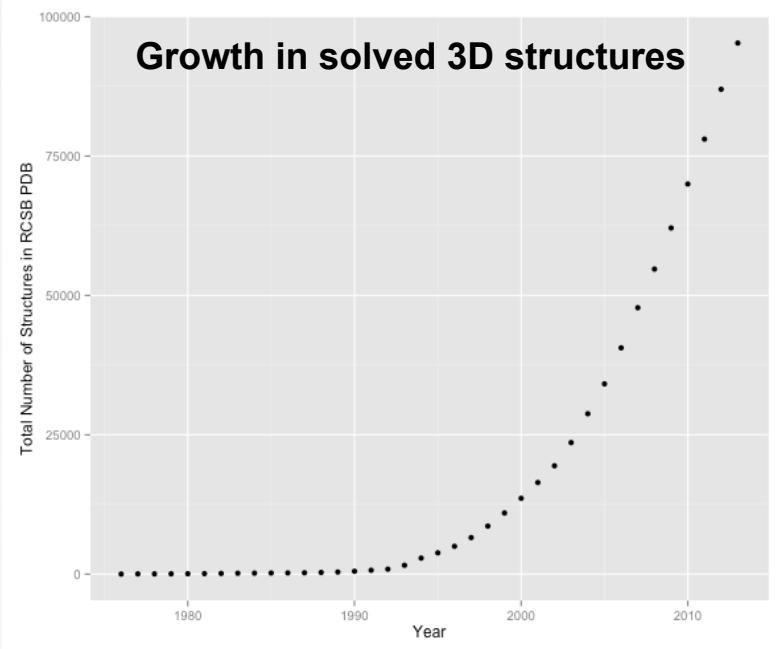


E.G. data from sequencing, structural genomics, proteomics, new high throughput assays, etc...

Why do we need Bioinformatics?

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

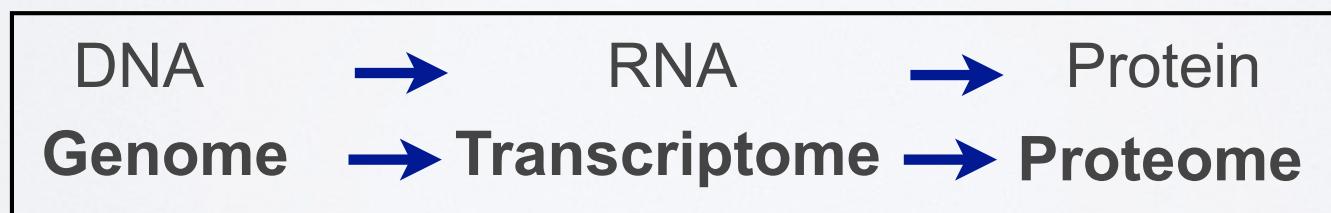
- Bioinformatics provides methods for the efficient:
 - ▶ **storage**
 - ▶ **annotation**
 - ▶ **search and retrieval**
 - ▶ **data integration**
 - ▶ **data mining and analysis**



E.G. data from sequencing, structural genomics, 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.



x 1,000

x 100,000

How do we actually do Bioinformatics?

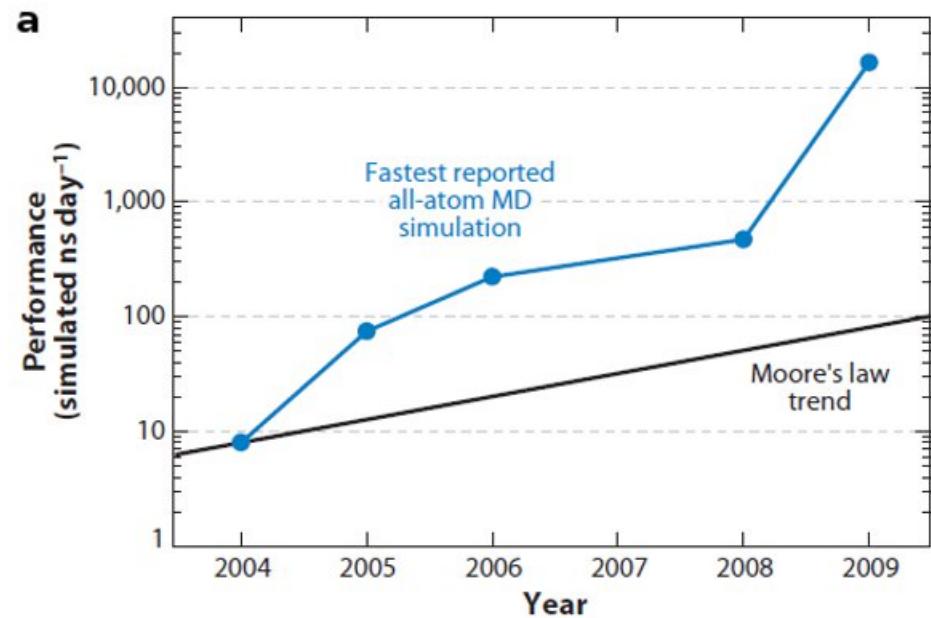
Pre-packaged tools and databases

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

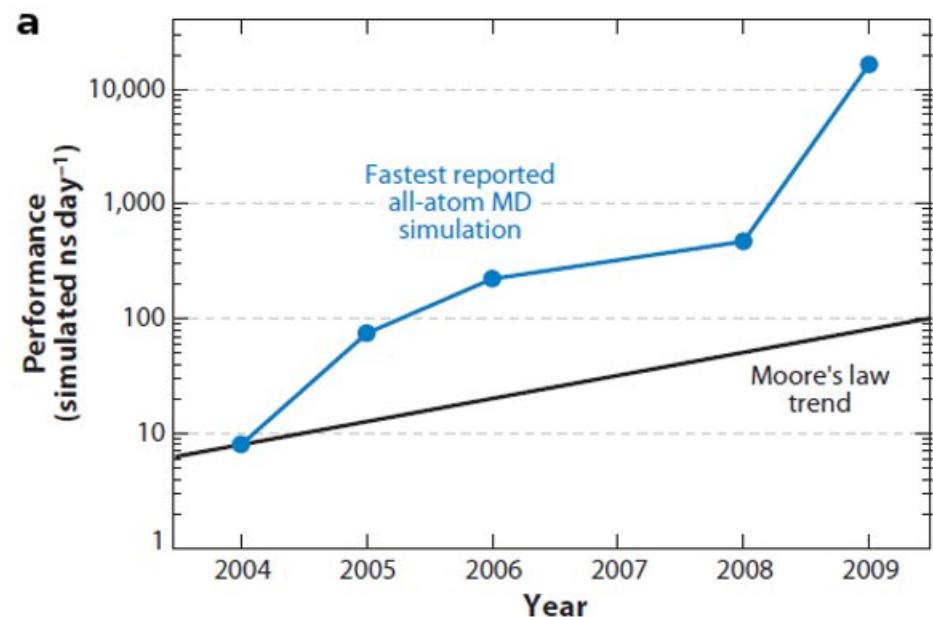
Tool development

- ▶ 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...

SIDE-NOTE: SUPERCOMPUTERS AND GPUS



SIDE-NOTE: SUPERCOMPUTERS AND GPUS



HOW COMPUTERS HAVE CHANGED

DATE	COST	SPEED	MEMORY	SIZE
1967	\$60M	0.1 MHz	1 MB	HALL
2013	\$6,000	1 GHz	10 GB	LAPTOP
CHANGE	10,000	10,000	10,000	10,000

If cars were like computers then a new Volvo would cost \$3, would have a top speed of 1,000,000 Km/hr, would carry 50,000 adults and would park in a shedbox



Skepticism & Bioinformatics

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

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

Common problems with Bioinformatics

Confusing multitude of tools available

- ▶ Each with many options and settable parameters

Most tools and databases are written by and for nerds

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

Most are developed independently

Notable exceptions are found at the:

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

General Parameters

Max target sequences

Select the maximum number of aligned sequences to display

Short queries Automatically adjust parameters for short input sequences

Expect threshold

Word size

Max matches in a query range

Scoring Parameters

Matrix

Gap Costs

Compositional adjustments

Filters and Masking

Filter Low complexity regions

Mask Mask for lookup table only
 Mask lower case letters

PSI/PHI/DELTA BLAST

Upload PSSM Optional no file selected

PSI-BLAST Threshold

Pseudocount

Even Blast has many settable parameters

Related tools with different terminology

STEP 3 - Set your
PROGRAM
FASTA

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

SCORE FORMAT

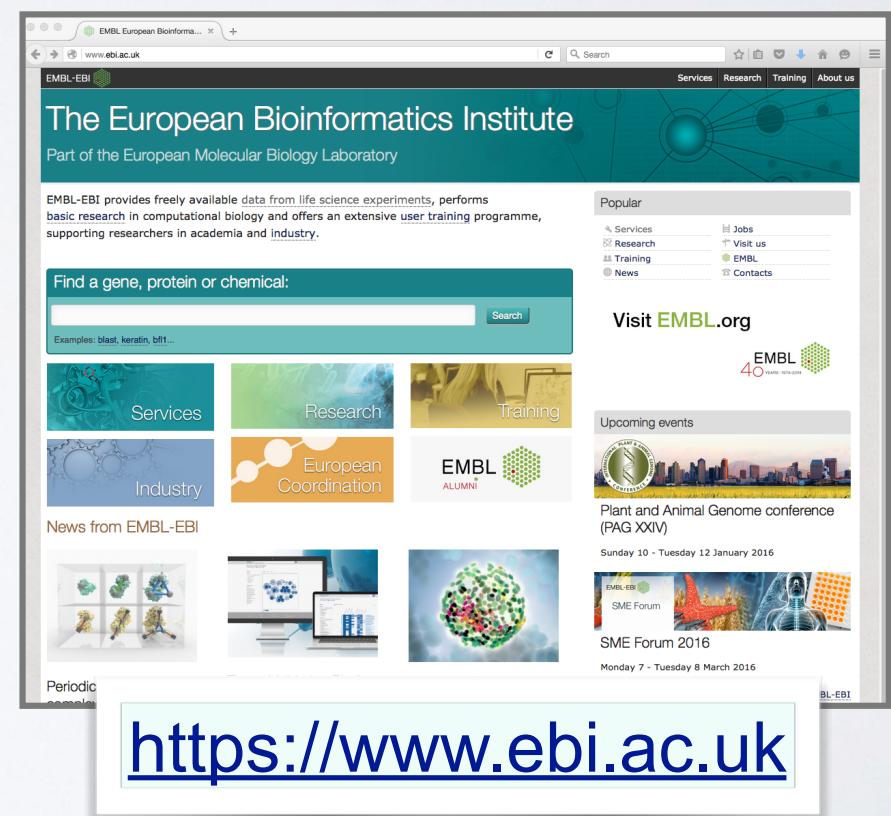
Key Online Bioinformatics Resources: NCBI & EBI

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



The screenshot shows the NCBI homepage with a blue header bar containing the NCBI logo, a search bar, and links for "Resources" and "How To". Below the header is a navigation menu 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 features a "Welcome to NCBI" section with a brief description of the center's mission, a "Get Started" section with links to "Tools", "Downloads", "How-Tos", and "Submissions", and sections for "3D Structures" (with a molecular model image) and "NCBI Announcements" (listing the new version of Genome Workbench, the July Newsletter, and the introduction to the 1000 Genomes project). A sidebar on the right lists "Popular Resources" such as PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem.

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



The screenshot shows the EMBL-European Bioinformatics Institute homepage. The header includes the EBI logo, a search bar, and links for "Services", "Research", "Training", and "About us". The main content area features a large banner for "The European Bioinformatics Institute, Part of the European Molecular Biology Laboratory". Below the banner, there is a "Find a gene, protein or chemical:" search bar with examples like "blast", "keratin", "bf1". The page is divided into several sections: "Services" (with a gear icon), "Research" (with a lab flask icon), "Training" (with a person icon), "Industry" (with a factory icon), "European Coordination" (with a network icon), "EMBL ALUMNI" (with a graduation cap icon), and "News from EMBL-EBI" (with a news icon). A sidebar on the right lists "Popular" links for "Services", "Research", "Training", "News", "Jobs", "Visit us", "EMBL", and "Contacts". At the bottom, there are sections for "Upcoming events" (listing the Plant and Animal Genome conference (PAG XXIV) and SME Forum 2016), and "Periodic" news items.

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

National Center for Biotechnology Information (NCBI)

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



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

National Center for Biotechnology Information

NCBI Resources How To Sign in to NCBI

All Databases Search

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Welcome to NCBI

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

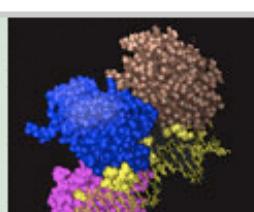
[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

Get Started

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

3D Structures

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



Popular Resources

PubMed

Bookshelf

PubMed Central

PubMed Health

BLAST

Nucleotide

Genome

SNP

Gene

Protein

PubChem

NCBI Announcements

New version of Genome Workbench available 06 Sep

An integrated, downloadable applicati

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

National Center for Biotechnology Information

NCBI Resources How To Sign in to NCBI

All Databases

Search

Popular Resources

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

Welcome to NCBI

The National Center for Biotechnology Information provides access to information and resources that support basic research and health by providing access to information.

About the NCBI | Mission | Our History

Get Started

- Tools: Analyze data using NCBI tools
- Downloads: Get NCBI data
- How-To's: Learn how to access and use NCBI resources
- Submissions: Submit data to NCBI databases

3D Structures

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

New version of Genome Workbench available

An integrated, downloadable application

06 Sep

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

National Center for Biotechnology Information

www.ncbi.nlm.nih.gov Reader Sign in to NCBI

NCBI Resources How To

All Databases Search

NCBI Home Resource List (A-Z)

Welcome to NCBI
The National Center for Biotechnology Information advances science

Popular Resources PubMed

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

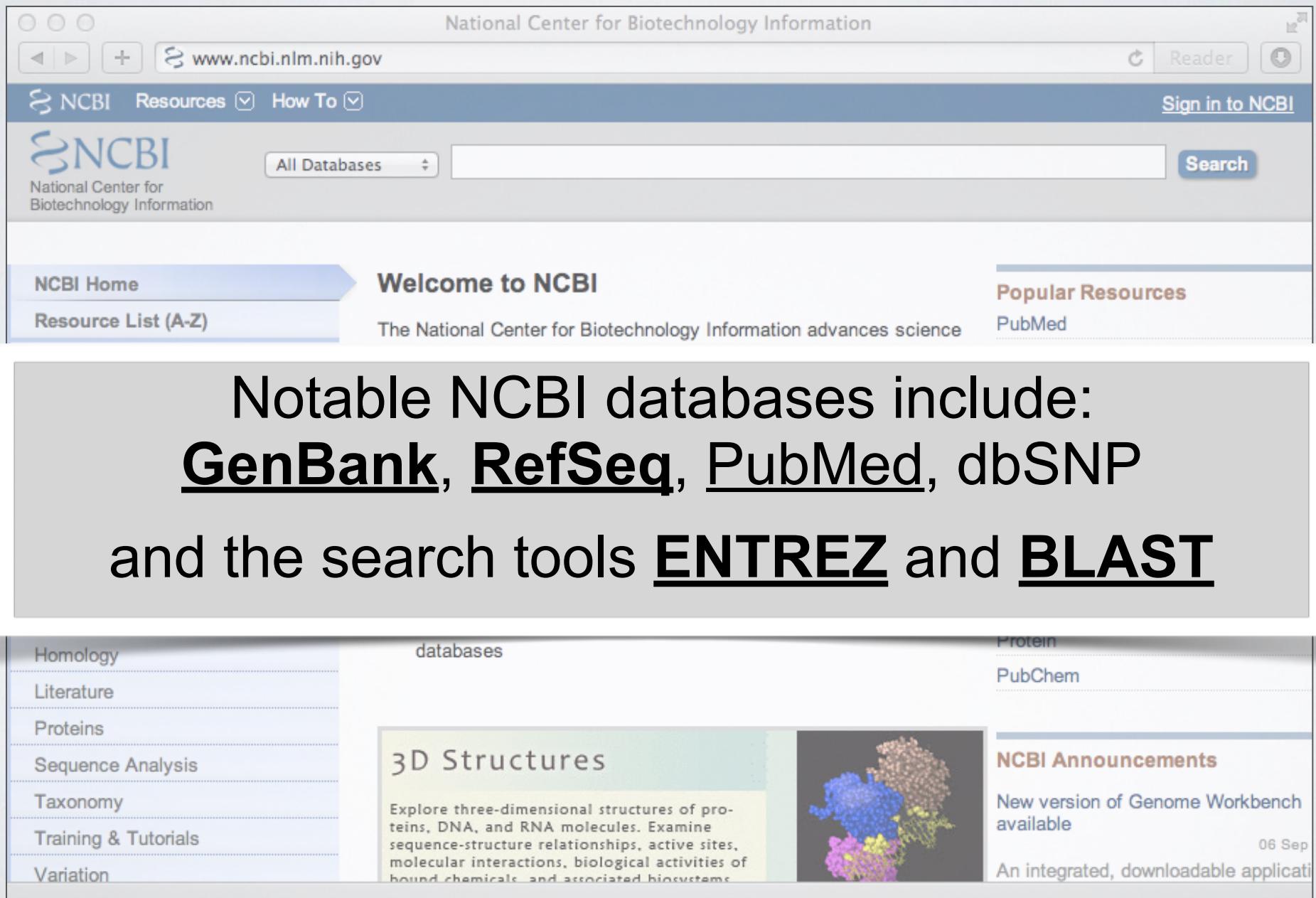
Homology Literature Proteins Sequence Analysis Taxonomy Training & Tutorials Variation

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.

Protein PubChem

NCBI Announcements
New version of Genome Workbench available 06 Sep
An integrated, downloadable applicati



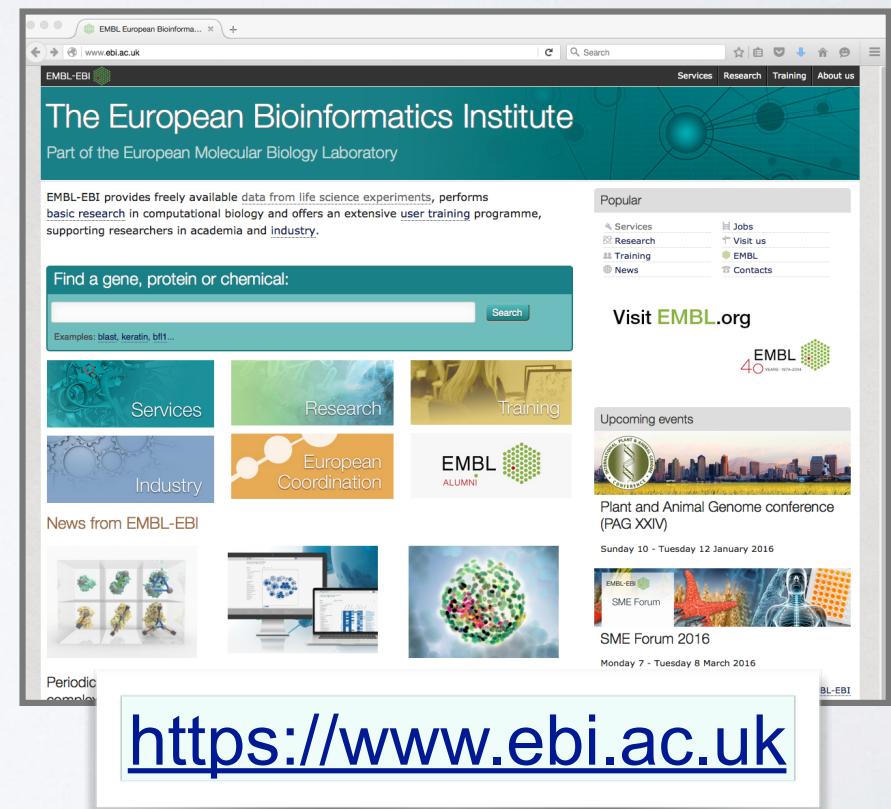
Key Online Bioinformatics Resources: NCBI & EBI

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



The screenshot shows the NCBI homepage with a blue header bar containing the NCBI logo, a search bar, and links for "Resources" and "How To". Below the header is a navigation menu 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 features a "Welcome to NCBI" section with a brief description of the center's mission, a "Get Started" section with links to "Tools", "Downloads", "How-Tos", and "Submissions", a "3D Structures" section showing a molecular model, and an "NCBI Announcements" section with news about the Genome Workbench and the July Newsletter. A sidebar on the right lists "Popular Resources" such as PubMed, Bookshelf, PubMed Central, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem.

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



The screenshot shows the EMBL-European Bioinformatics Institute homepage. The header includes the EBI logo, a search bar, and links for "Services", "Research", "Training", and "About us". The main content area features a banner for "The European Bioinformatics Institute, Part of the European Molecular Biology Laboratory". Below the banner is a "Find a gene, protein or chemical:" search bar. The page is divided into several sections: "Services" (with a gear icon), "Research" (with a scientist icon), "Training" (with a person icon), "Industry" (with a factory icon), "European Coordination" (with a network icon), and "EMBL ALUMNI" (with a green hexagonal icon). There are also sections for "News from EMBL-EBI" (with images of molecular models) and "Upcoming events" featuring the "Plant and Animal Genome conference (PAG XXIV)" (from Sunday 10 - Tuesday 12 January 2016) and the "SME Forum 2016" (from Monday 7 - Tuesday 8 March 2016). A sidebar on the right lists "Popular" links for "Services", "Research", "Training", "News", "Jobs", "Visit us", "EMBL", and "Contacts".

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

European Bioinformatics Institute (EBI)

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



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

The screenshot shows the homepage of the European Bioinformatics Institute (EBI) at www.ebi.ac.uk. The page features a dark teal header with the EMBL-EBI logo and navigation links for Services, Research, Training, and About us. Below the header, a large banner reads "The European Bioinformatics Institute" and "Part of the European Molecular Biology Laboratory". A central text block states: "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." To the right, a "Popular" sidebar lists links to Services, Research, Training, News, Jobs, Visit us, EMBL, and Contacts. A search bar at the top allows users to "Find a gene, protein or chemical:" with examples like blast, keratin, bfl1... Below the search bar are several promotional boxes: "Services" (highlighted with a red border), "Research", "Training", "Industry", "European Coordination", and "EMBL ALUMNI". At the bottom, there's a section for "News from EMBL-EBI" and a "Upcoming events" section featuring the "Plant and Animal Genome conference (PAG XXIV)".

EMBL European Bioinformatic...

www.ebi.ac.uk

Search

Services | Research | Training | About us

The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

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

Find a gene, protein or chemical:

Examples: blast, keratin, bfl1...

Search

Services

Research

Training

Industry

European Coordination

EMBL ALUMNI

Popular

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

Visit EMBL.org

40 YEARS | 1974-2016

Upcoming events

INTERNATIONAL PLANT & ANIMAL GENOME CONFERENCE

Sunday 10 - Tuesday 12 January 2016

EMBL-EBI

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

The screenshot shows the EBI Services website at www.ebi.ac.uk/services. The page features a dark header with the EMBL-EBI logo and navigation links for Services, Research, Training, and About us. Below the header is a large teal banner with the word "Services". The main content area is titled "Bioinformatics services". It includes a paragraph about maintaining molecular databases and links to various services like Ensembl, UniProt, PDB, and ChEMBL. There are also sections for "Service news" (with an image of a butterfly) and "Training". A footer at the bottom left mentions "Programmatic access".

Services < EMBL-EBI

www.ebi.ac.uk/services

EMBL-EBI

Services | Research | Training | About us

Services

Overview | A to Z | Data submission | Support

Bioinformatics services

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

DNA & RNA
genes, genomes & variation

Gene expression
RNA, protein & metabolite expression

Proteins
sequences, families & motifs

Structures
Molecular & cellular structures

Systems
reactions, interactions & pathways

Chemical biology
chemogenomics & metabolomics

Ontologies
taxonomies & controlled vocabularies

Literature
Scientific publications & patents

Cross domain
cross-domain tools & resources

Popular

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

Service news

Training

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

The screenshot shows the EBI Services website at www.ebi.ac.uk/services. The page features a navigation bar with links for Services, Research, Training, and About us. Below the navigation is a section titled "Popular" which lists several databases: Ensembl, UniProt, PDB, ArrayExpress, and ChEMBL. The "Proteins" service is highlighted with a red border. The main content area displays nine service categories in a grid:

- DNA & RNA: genes, genomes & variation
- Gene expression: RNA, protein & metabolite expression
- Proteins: sequences, families & motifs (highlighted)
- 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

A large banner on the right side of the page features a monarch butterfly and the word "Training".

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

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

Proteins

Popular services

 UniProt	UniProt: The Universal Protein Resource The gold-standard, comprehensive resource for protein sequence and functional annotation data.
 InterPro	InterPro A database for the classification of proteins into families, domains and conserved sites.
 PRIDE	PRIDE: The Proteomics Identifications Database An archive of protein expression data determined by mass spectrometry.
 Pfam	Pfam A database of hidden Markov models and alignments to describe conserved protein families and domains.
 Clustal Omega	Clustal Omega Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.
 HMMER	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 screenshot shows the homepage of the European Bioinformatics Institute (EBI) at www.ebi.ac.uk. The page features a dark teal header with the EMBL-EBI logo and navigation links for Services, Research, Training, and About us. Below the header, the main title "The European Bioinformatics Institute" and subtitle "Part of the European Molecular Biology Laboratory" are displayed. A large text block describes EMBL-EBI's mission to provide freely available data from life science experiments, perform basic research in computational biology, and offer user training. A search bar is present above a grid of links to various services. One link, "Training", is highlighted with a red box. To the right, a "Popular" sidebar lists links to Services, Research, Training, and News. A "Visit EMBL.org" section features the EMBL 40th anniversary logo. At the bottom, there's a "Upcoming events" section for the Plant and Animal Genome conference (PAG XXIV).

EMBL European Bioinformatic... [+ ↗](#)

www.ebi.ac.uk

Search

EMBL-EBI

The European Bioinformatics Institute
Part of the European Molecular Biology Laboratory

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

Find a gene, protein or chemical:

Search

Examples: blast, keratin, bfl1...

Services

Research

Training

Industry

European Coordination

EMBL ALUMNI

Popular

Jobs

Visit us

EMBL

News

Visit EMBL.org

40 YEARS | 1974-2016

Upcoming events

INTERNATIONAL PLANT & ANIMAL GENOME CONFERENCE

Sunday 10 - Tuesday 12 January 2016

EMBL-EBI

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

A screenshot of a web browser displaying an EBI online training course. The URL in the address bar is www.ebi.ac.uk/training/online/course/using-sequence-similarity-searching-tools-embl-ebi. The page title is "Using sequence similarity s...". The main navigation menu includes "Services", "Research", **Training**, "About us", and a search bar. A banner at the top says "Train online". Below it, a sub-navigation menu shows "Training", "Train online Home", "Course list", "Glossary", "Support & Feedback", and "Log in / Register". The breadcrumb trail indicates the current location: "training » online » course-list » using-sequence-similarity-searching-tools-embl-ebi".

The central content area is titled "Using sequence similarity searching tools at EMBL-EBI: webinar". It features a video player showing a slide with the title "Using sequence similarity search tools at EMBL-EBI" and the subtitle "Finding homologous sequences with BLAST, FASTA, PSI-Search etc.". The slide also lists "Andrew Cowley" with email addresses "andrew.cowley@ebi.ac.uk" and "support@ebi.ac.uk". A small portrait of Andrew Cowley is visible on the right. Below the video player, a yellow box contains the text: "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."

On the right side of the page, there are two sidebar boxes: "Popular" (with links to "Train online", "Find us", and "Funding") and "Find us at..." (with links to "Open days and career days", "Conference exhibitions", "EMBL courses and events", "Genome campus events", and "Science for schools").

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

A screenshot of a web browser displaying the EBI Train online website. The address bar shows 'www.ebi.ac.uk/training/online/'. The page features a green header with the EMBL-EBI logo and navigation links for Databases, Tools, Research, Training, Industry, About Us, Help, Site Index, and a 'Beta' badge. A 'Train online' section is highlighted in orange, containing a 'Train online Home' link.

Notable EBI databases include:

ENA, **UniProt**, **Ensembl**

and the tools **FASTA**, **BLAST**, **InterProScan**,
MUSCLE, **DALI**, **HMMER**

Find a course

Browse by subject



[Genes and Genomes](#)



[Gene Expression](#)

Interactions, Pathways and Networks

Next Class...

**MAJOR BIOINFORMATICS
DATABASES AND ASSOCIATED
ONLINE TOOLS**

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, Biolmage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty_cDB, DIP, DOGS, DOMO, DPD, 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, IMGT, Kabat, KDNA, KEGG, KloTho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, MycDB, NDB, NRSub, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-MODEL Repository, SWISS-PROT, TeIDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc ..!!!!

Bioinformatics Databases

There are lots of Bioinformatics Databases

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

< Handout_Major_Databases.pdf >

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGB, Beanref, CANSITE, CarbBank, CARBHYD, CATH, CAZy, ChickGBASE, Colibri, COPE, CottonDB, dbSTS, DDBJ, DGP, DictyDb, ECGC, EC02DBASE, FlyBase, GDB, HOMER, KEGG, MHCDB, MycoDB, PDB, PDBe, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc

Side-note: Databases come in all shapes and sizes



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

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.

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/>

NCBI Resources How To Sign in to NCBI

All Databases Search

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

Welcome to NCBI

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

About the NCBI | Mission | Organization | NCBI News

Get Started

- [Data](#): Search and analyze data using NCBI software
- [Tools](#): Get NCBI data or software
- [How Tos](#): Learn how to accomplish specific tasks at NCBI
 - [Submissions](#): Submit data to GenBank or other NCBI databases

Genotypes and Phenotypes

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

Resources

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

NCBI Announcements

RefSeq release 69 available on

The full RefSeq release 69 is now available on the FTP site with 74 records describing 50,370,460

Hands on demo (or see following slides)

ras - GQuery: Global Cross X

www.ncbi.nlm.nih.gov/gquery/?term=ras

NCBI Resources ▾ How To ▾ Sign in to NCBI

Search NCBI databases

Help

ras

Search

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

Literature			Genes		
Books	1,677	books and reports	EST	3,985	expressed sequence tag sequences
MeSH	402	ontology used for PubMed indexing	Gene	87,165	collected information about gene loci
NLM Catalog	223	books, journals and more in the NLM Collections	GEO DataSets	3,732	functional genomics studies
PubMed	54,672	scientific & medical abstracts/citations	GEO Profiles	1,622,789	gene expression and molecular abundance profiles
PubMed Central	96,114	full-text journal articles	HomoloGene	696	homologous gene sets for selected organisms
Health			PopSet	2,254	sequence sets from phylogenetic and population studies
ClinVar	759	human variations of clinical significance	UniGene	4,770	clusters of expressed transcripts
dbGaP	120	genotype/phenotype interaction studies	Proteins		
GTR	1,879	genetic testing registry			

62

ras - Gene - NCBI

www.ncbi.nlm.nih.gov/gene/?term=ras

NCBI Resources How To Sign in to NCBI

Gene Gene ras Search Save search Advanced Help

Show additional filters Hide sidebar >

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

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

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

Results: 1 to 20 of 85633

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

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

Filters: Manage Filters

Top Organisms [Tree]

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

[More...](#)

Find related data

Database: Select Find items

Search details

ras[All Fields] AND alive[property]

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

Results: 1 to 20 of 1126

Name/Gene ID	Description	Location	Aliases
<input checked="" type="checkbox"/> NRAS ID: 4893	neuroblastoma RAS viral (v-ras) oncogene homolog [<i>Homo sapiens</i> (human)]	Chromosome 1, NC_00001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
<input checked="" type="checkbox"/> KRAS ID: 3845	Kirsten rat sarcoma viral oncogene homolog [<i>Homo sapiens</i> (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, KI-RAS1, KRAS2, NS, NS2, RASK2

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

Filters: [Manage Filters](#)

Find related data

Database: Select

Search details

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

Recent activity

Turn Off Clear

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
- Pseudogene

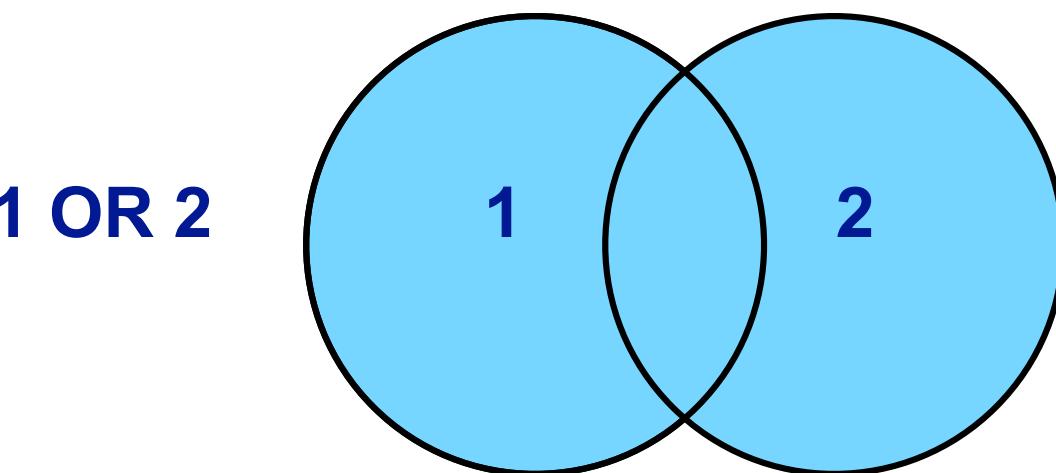
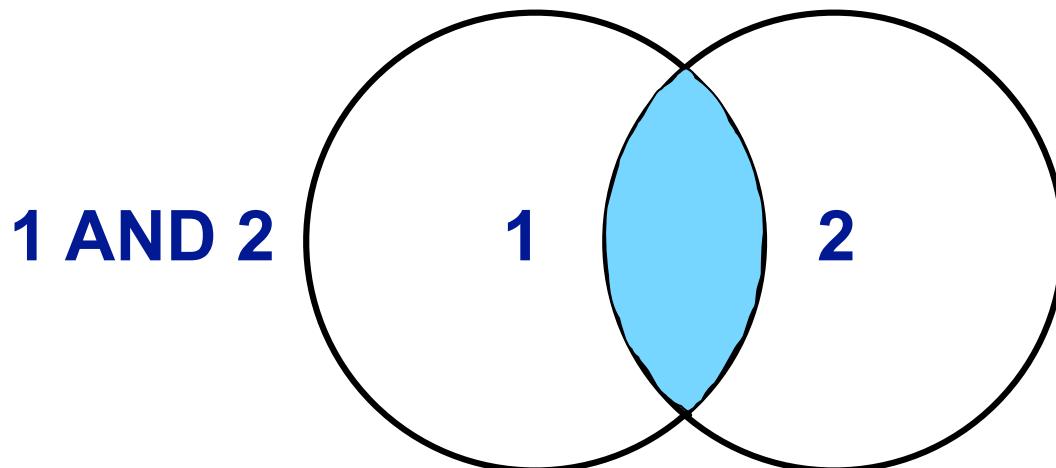
Sequence content

- CCDS
- Ensembl
- RefSeq

Status clear

Current only

Chromosome locations Select



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)**

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

Gene Gene (ras) AND "Homo sapiens"[porgn:txid9606] Search Save search Advanced Help

Show additional filters Hide sidebar >

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

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
<input type="checkbox"/> NRAS ID: 4893	neuroblastoma RAS viral (v-ras) oncogene homolog [<i>Homo sapiens</i> (human)]	Chromosome 1, NC_00001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
<input type="checkbox"/> KRAS ID: 3845	Kirsten rat sarcoma viral oncogene homolog [<i>Homo sapiens</i> (human)]	Chromosome 12, NC_000012.12 (25205246..25250923, complement)	C-K-RAS, CFC2, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, KI-RAS1, KRAS2, NS, NS2, RASK2

Find related data

Database: Select Find items

Search details

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

Search See more...

Recent activity

Turn Off Clear

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

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

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

Summary

Official Symbol KRAS provided by HGNC

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

Primary source HGNC:HGNC:6407

See related Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193

Gene type protein coding

RefSeq status REVIEWED

Organism *Homo sapiens*

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

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

Table of contents

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

67

KRAS Kirsten rat sarcoma

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

NCBI Resources How To Sign in to NCBI

Gene Display S Search Help

KRAS (human)

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

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

Summary

Official Symbol KRAS provided by HGNC

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

Primary source HGNC:HGNC:6407

See related Ensembl:ENSG00000133703; HPRD:01817; MIM:190070; Vega:OTTHUMG00000171193

Gene type protein coding

RefSeq status REVIEWED

Organism Homo sapiens

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

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

Table of contents

Summary

Genomic context

Genomic regions, transcripts, and products

Bibliography

Phenotypes

Variation

HIV-1 interactions

Pathways from BioSystems

Interactions

General gene information

Markers, Related pseudogene(s), Homology, Gene Ontology

General protein information

NCBI Reference Sequences (RefSeq)

Related documents

Hide sidebar >>

KRAS Kirsten rat sarcoma

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

Genomic context

Location: 12p12.1 See KRAS in [Epigenomics](#), [MapViewer](#)

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

Go to [reference sequence details](#)

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

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

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

KRAS Kirsten rat sarcoma

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

NCBI Resources How To Sign in to NCBI

Gene

Display Settings

KRAS Ki (human)] Gene ID: 3845

Summary

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:OTTHUMG00000171193

Gene type protein coding

RefSeq status REVIEWED

Organism Homo sapiens

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

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

Search Help Hide sidebar >

Table of contents

Summary

Genomic context

Genomic regions, transcripts, and products

Bibliography

Phenotypes

Variation

HIV-1 interactions

Pathways from BioSystems

Interactions

General gene information

Markers, Related pseudogene(s), Homology, Gene Ontology

General protein information

NCBI Reference Sequences (RefSeq)

Related documents

KRAS Kirsten rat sarcoma

Gene Ontology Provided by GOA

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

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

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



GO: Gene Ontology

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

The screenshot shows a web browser window with the UniProt-GOA database. The title bar includes tabs for "KRAS Kirsten rat sarcoma" and "UniProt-GOA < EMBL-EBI". The main content area features the "UniProt-GOA" logo and a search bar with examples like "GO:0006915, tropomyosin, P06727". A sidebar on the right is titled "Menu" and lists various resources.

UniProt-GOA

Examples: GO:0006915, tropomyosin, P06727

Search

Services | Research | Training | About us

Overview | New to UniProt-GOA | FAQ | Contact Us

Gene Ontology Annotation (UniProt-GOA) Database

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

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

Menu

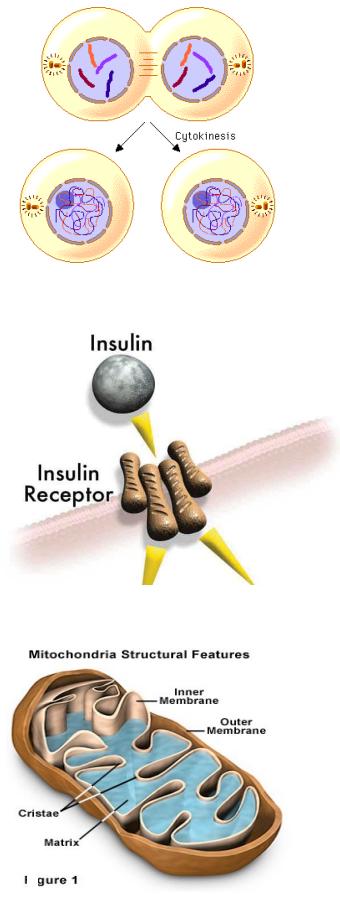
- Downloads
- Searching UniProt-GOA
- Annotation Methods
- Annotation Tutorial
- Manual Annotation Efforts
 - Reference Genome Annotation Initiative
 - Cardiovascular Gene Ontology Annotation Initiative
 - Renal Gene Ontology Annotation Initiative
 - Exosome Gene

Why do we need Ontologies?

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

GO Ontologies

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



KRAS Kirsten rat sarcoma

Gene Ontology Provided by GOA

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

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

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

Scroll down to
UniProt link

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

genomic X01669.1 CAA25828.1

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

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

Additional links

Scroll down to
Very bottom for
UniProt link

You are here: NCBI > Genes & Expression > Gene

Write to the Help Desk

GETTING STARTED

- [NCBI Education](#)
- [NCBI Help Manual](#)
- [NCBI Handbook](#)
- [Training & Tutorials](#)

RESOURCES

- [Chemicals & Bioassays](#)
- [Data & Software](#)
- [DNA & RNA](#)
- [Domains & Structures](#)
- [Genes & Expression](#)
- [Genetics & Medicine](#)
- [Genomes & Maps](#)
- [Homology](#)
- [Literature](#)
- [Proteins](#)
- [Sequence Analysis](#)
- [Taxonomy](#)

POPULAR

- [PubMed](#)
- [Bookshelf](#)
- [PubMed Central](#)
- [PubMed Health](#)
- [BLAST](#)
- [Nucleotide](#)
- [Genome](#)
- [SNP](#)
- [Gene](#)
- [Protein](#)
- [PubChem](#)

FEATURED

- [Genetic Testing Registry](#)
- [PubMed Health](#)
- [GenBank](#)
- [Reference Sequences](#)
- [Gene Expression Omnibus](#)
- [Map Viewer](#)
- [Human Genome](#)
- [Mouse Genome](#)
- [Influenza Virus](#)
- [Primer-BLAST](#)
- [Sequence Read Archive](#)

NCBI INFORMATION

- [About NCBI](#)
- [Research at NCBI](#)
- [NCBI News](#)
- [NCBI FTP Site](#)
- [NCBI on Facebook](#)
- [NCBI on Twitter](#)
- [NCBI on YouTube](#)

UniProt will detail much more information for protein coding genes

KRAS - GTPase KRas prec ×

www.uniprot.org/uniprot/P01116

UniProtKB Advanced

BLAST Align Retrieve/ID Mapping Help Contact Basket

P01116 - RASK_HUMAN

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

Display None

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

BLAST Align Format Add to basket History Feedback Help video

Functionⁱ

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

Enzyme regulationⁱ

Alternates between an inactive form bound to GDP and an active form bound to GTP. Activated by a guanine nucleotide-exchange factor (GEF) and inactivated by a GTPase-activating protein (GAP). Interaction with SOS1 promotes exchange of bound GDP by GTP. 3 Publications

Regions

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

KRAS - GTPase KRas precursor

www.uniprot.org/uniprot/P01116

UniProtKB Advanced

BLAST Align Retrieve/ID Mapping Help Contact Basket

P01116 - RASK_HUMAN

Protein GTPase KRas
Gene KRAS
Organism Homo sapiens (Human)
Status Reviewed - 5 Publications

Display None

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

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

Enzyme regulationⁱ
Alternates between an inactive form bound to GDP and an active form bound to GTP. Activated by a guanine nucleotide-exchange factor (GEF) and inactivated by a GTPase-activating protein (GAP). Interaction with SOS1 promotes exchange of bound GDP by GTP. 3 Publications

Regions

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

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?

KRAS - GTPase KRas prec x
www.uniprot.org/uniprot/P01116

Display None

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

Pathology & Biotech¹

Involvement in diseaseⁱ

LEUKEMIA, ACUTE MYELOGENOUS (AML)

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

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

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

LEUKEMIA, JUVENILE MYELOMONOCYTIC (JMML)

[MIM:607785]: An aggressive pediatric myelodysplastic syndrome/myeloproliferative disorder characterized by malignant transformation in the hematopoietic stem cell compartment with proliferation of differentiated progeny. Patients have splenomegaly, enlarged lymph nodes, rashes, and hemorrhages.

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

NOONAN SYNDROME 3 (NS3)

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



Example Questions:

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

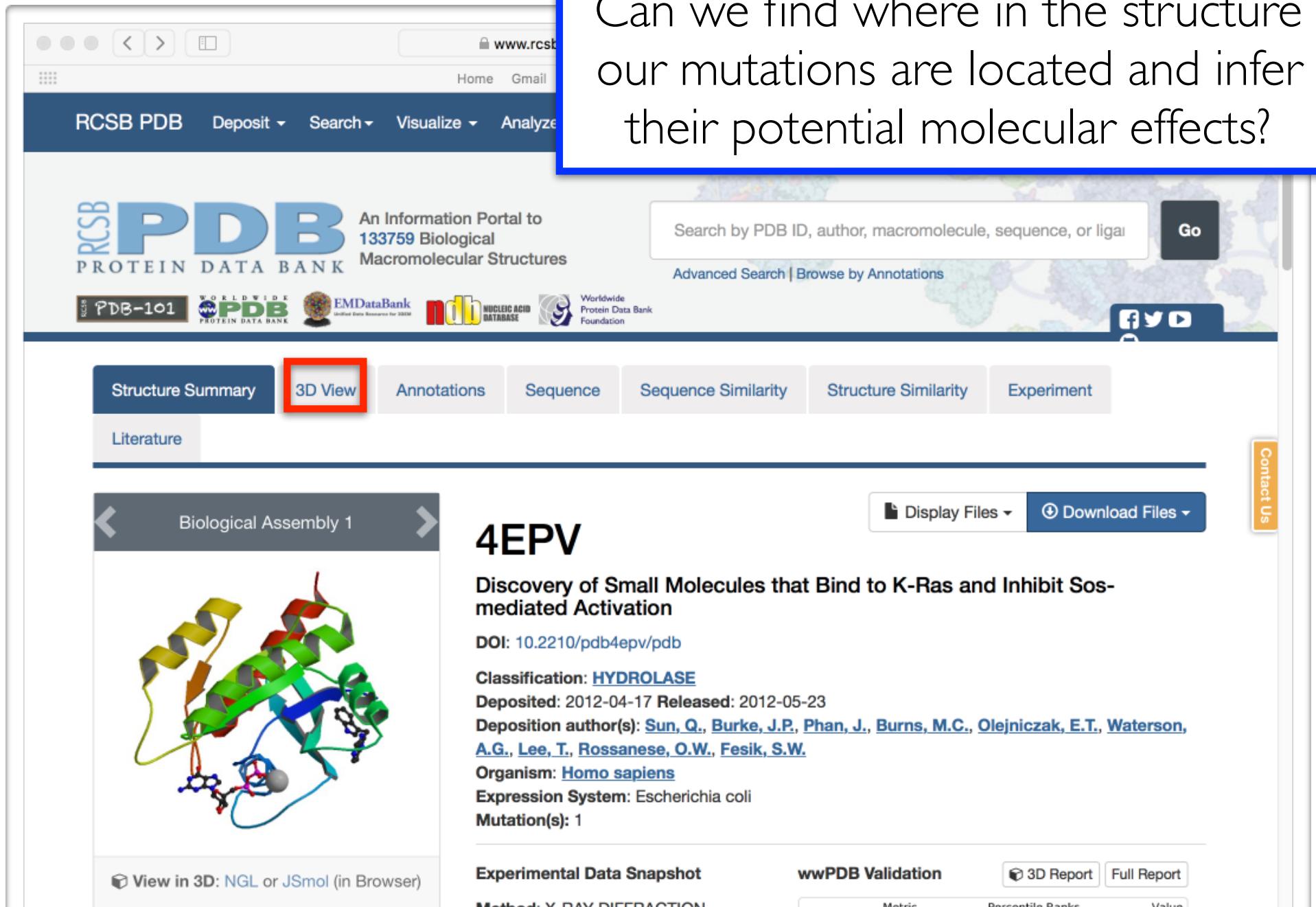
The screenshot shows the Uniprot protein details page for KRAS - GTPase KRas prec (P01116). The 'STRUCTURE' section is highlighted with a red box. In the '3D structure databases' table, the 'RCSB PDBⁱ' link is highlighted with a green box. A red arrow points from the '4EPV' entry in the table to a callout box containing the text 'Open link in a new tab!'. The table lists various PDB entries with their methods, resolutions, and chain information.

Entry	Method	Resolution (Å)	Chain	Positions	PDBsum
1D8D	X-ray	2.00	P	178-188	[»]
1D8E	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	[»]
4DSO	X-ray	1.85	A	2-164	[»]
4EPR	X-ray	2.00	A	1-164	[»]
4EPT	X-ray	2.00	A	1-164	[»]
4EPV	X-ray	1.35	A	1-164	[»]
4EPW	X-ray	1.70	A	1-1	
4EPX	X-ray	1.76	A	1-1	
4EPY	X-ray	1.80	A	1-1	
4L8G	X-ray	1.52	A	1-1	
4LDJ	X-ray	1.15	A	1-164	[»]
4LPK	X-ray	1.50	A/B	1-169	[»]

**Open link in
a new tab!**

Lets view the 3D structure:

Can we find where in the structure our mutations are located and infer their potential molecular effects?



The screenshot shows the RCSB PDB homepage. At the top, there is a navigation bar with links for Home, Gmail, and various menu options: Deposit, Search, Visualize, and Analyze. Below this is the main header featuring the RCSB PDB logo and the text "An Information Portal to 133759 Biological Macromolecular Structures". The header also includes links for PDB-101, Worldwide Protein Data Bank, EMDDataBank, Nucleic Acid Database, and the Worldwide Protein Data Bank Foundation. To the right of the header is a search bar with the placeholder "Search by PDB ID, author, macromolecule, sequence, or ligand" and a "Go" button. Below the header is a navigation menu with tabs: Structure Summary, 3D View (which is highlighted with a red box), Annotations, Sequence, Sequence Similarity, Structure Similarity, Experiment, and Literature. On the left side, there is a thumbnail image of a protein structure labeled "Biological Assembly 1" with a 3D ribbon model. To the right of the thumbnail, the PDB ID "4EPV" is displayed, along with the title "Discovery of Small Molecules that Bind to K-Ras and Inhibit Sos-mediated Activation". Below the title, the DOI is listed as "DOI: 10.2210/pdb4epv/pdb". Further down, the classification is listed as "HYDROLASE", and the deposition date is "2012-04-17" and the release date is "2012-05-23". The deposition authors are listed as "Sun, Q., Burke, J.P., Phan, J., Burns, M.C., Olejniczak, E.T., Waterson, A.G., Lee, T., Rossanese, O.W., Fesik, S.W.". The organism is "Homo sapiens", the expression system is "Escherichia coli", and there is 1 mutation. At the bottom of the page, there are sections for "Experimental Data Snapshot" and "wwPDB Validation", along with links for "3D Report" and "Full Report".

Lets view the 3D structure:

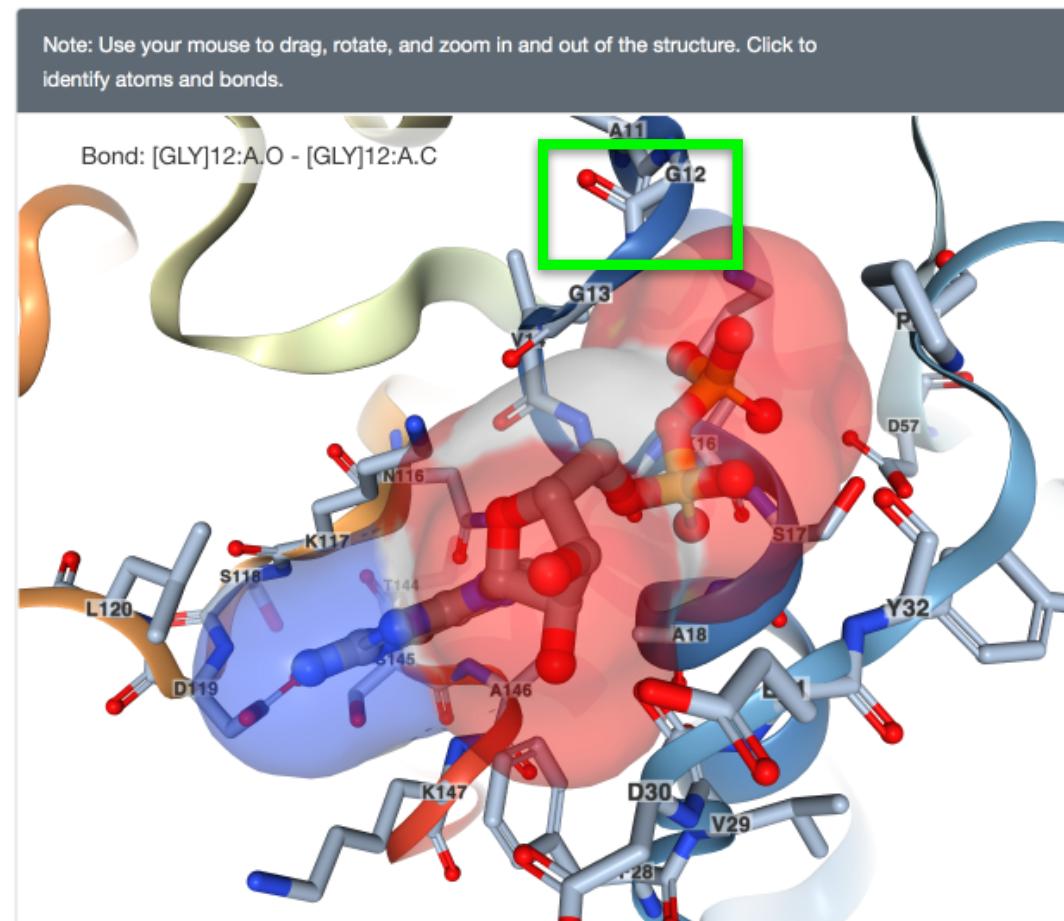
Can we find where in the structure our mutations are located and infer their potential molecular effects?

The screenshot shows the RCSB PDB interface with the identifier "4EPV". The main content area displays a 3D ribbon model of a protein structure. A green rectangular box highlights a specific region near the top center of the protein, specifically around residues A11, G12, and G13. A note at the top of the structure panel says: "Note: Use your mouse to drag, rotate, and zoom in and out of the structure. Click to identify atoms and bonds." Below the structure, a bond label "Bond: [GLY]12:A.O - [GLY]12:A.C" is shown. The top navigation bar includes links for Home, Gmail, Deposit, Search, Visualize, and Analyze.

Display Files ▾

4EPV

Discovery of Small Molecules that Bind to K-Ras and Inhibit Sos-mediated Activation



Display Options

Assembly ? Bioassembly 1

Model ? Model 1

Symmetry ? None

Interaction ? [GDP]201:A

Style ? Cartoon

Color ? Rainbow

Ligand ? None

Quality ? Automatic

Water ? Ions ?

Hydrogens ? Clashes ?

Contact Us

Viewer Options

Back to UniProt:

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

KRAS - GTPase KRas prec X
www.uniprot.org/uniprot/P01116

Display None

FUNCTION

NAMES & TAXONOMY

SUBCELL. LOCATION

PATHOL./BIOTECH

PTM / PROCESSING

EXPRESSION

INTERACTION

STRUCTURE

FAMILY & DOMAINS

SEQUENCES (2)

CROSS-REFERENCES

PUBLICATIONS

ENTRY INFORMATION

MISCELLANEOUS

SIMILAR PROTEINS

OrthoDB EVO
PhylomeDBⁱ P01
TreeFamⁱ TF3

Family and domain databases

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

PFAM is one of the best protein family databases

Sequences (2)ⁱ

Sequence statusⁱ: Complete.

Sequence processingⁱ: The displayed sequence is further processed into a mature form.

This entry describes 2 isoformsⁱ produced by alternative splicing. Align

Example Questions:

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

KRAS - GTPase KRas preo × Me Pfam: Family: Ras (PF0007) ×

EMBL-EBI pfam.xfam.org/family/PF00071

HOME

Family: Ras (PF00071)

Summary

Domain organisation

Clan

Alignments

HMM logo

Trees

Curation & model

Species (highlighted with a red box)

Interactions

Structures

Jump to... ⓘ

enter ID/acc Go

Summary: Ras family

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

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

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

Ras subfamily [Edit Wikipedia article](#)

This article is about p21/Ras protein. For the p21/waf1 protein, see [p21](#).

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

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

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

Because these signals result in cell growth and division, overactive Ras signaling can ultimately lead to [cancer](#).^[1] The 3 Ras genes in humans ([HRAS](#), [KRAS](#), and [NRAS](#)) are the most common [oncogenes](#) in human [cancer](#); mutations that permanently activate Ras are found in 20% to 25% of all human tumors and up to 90% in certain types of cancer (e.g., [pancreatic cancer](#)).^[2] For this reason, Ras inhibitors are being studied as a treatment for cancer, and other diseases with Ras overexpression.

Contents [hide]

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

Chemical structure

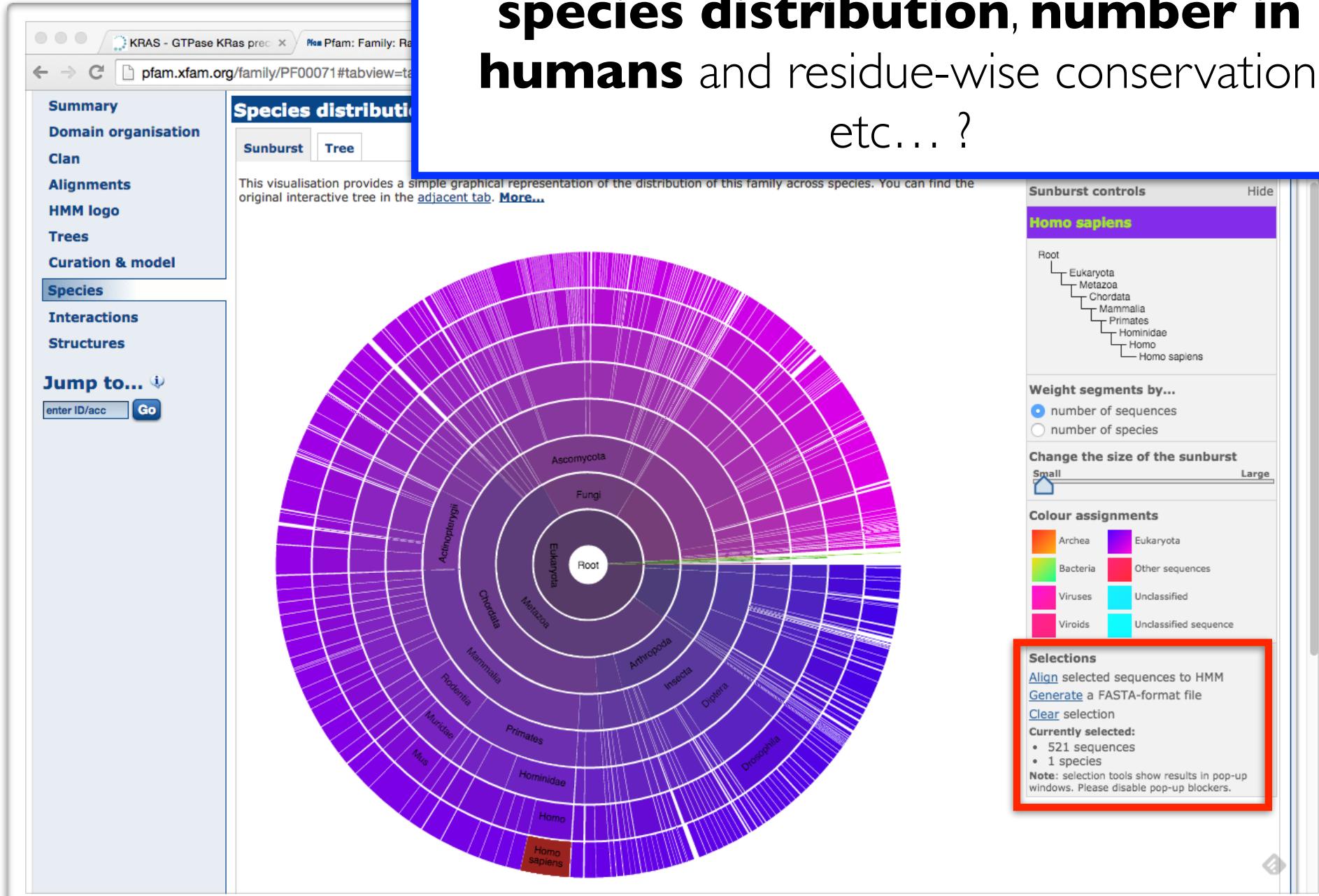
H-Ras structure PDB 121p, surface colored by conservation in Pfam seed alignment: gold, most conserved; dark cyan, least conserved.

Identifiers

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

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... ?

KRAS - GTPase KRas pre... [More Pfam: Family: Results](#)

Summary Domain organisation Species distribution

Species distribution

Pfam: Pfam alignment viewer

Alignment for selected sequences

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

Jump to enter ID/acc

Jump to enter ID/acc

1 2 3 4 5 6 7 8 9 10 11 ...

There are 18 pages in this alignment. Show page

Download this alignment.

Close window

can find the

Sunburst controls Hide

Homo sapiens

Root
└ Eukaryota
 └ Metazoa
 └ Chordata
 └ Mammalia
 └ Primates
 └ Hominidae
 └ Homo
 └ Homo sapiens

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

Change the size of the sunburst
Small Large

Colour assignments

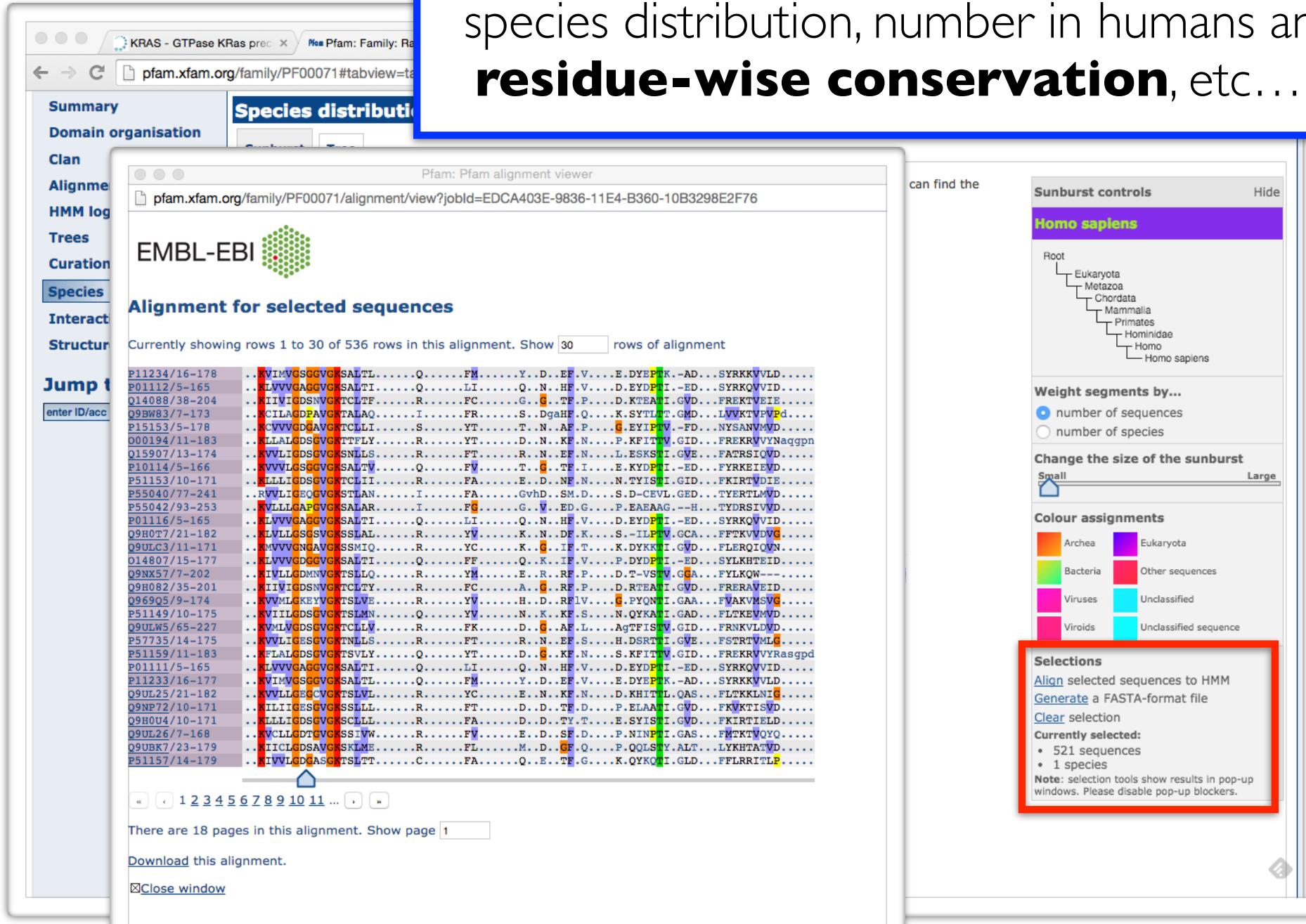
Archea	Eukaryota
Bacteria	Other sequences
Viruses	Unclassified
Viroids	Unclassified sequence

Selections

[Align](#) selected sequences to HMM
[Generate](#) a FASTA-format file
[Clear](#) selection

Currently selected:
• 521 sequences
• 1 species

Note: selection tools show results in pop-up windows. Please disable pop-up blockers.



Example Questions:

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

KRAS - GTPase KRas pre... More Pfam: Family: Ras

pfam.xfam.org/family/PF00071#tabview=tab4

EMBL-EBI 

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

Family: Ras (PF00071)

Summary Domain organisation Clan Alignments **HMM logo** HMM logo (highlighted with a red box)

Trees Curation & model Species Interactions Structures

Jump to... enter ID/acc Go

HMM logo

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

Contribution



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



Family: Kinesin (PF00225)

Loading page components (1 remaining)...

126 architectures 4150 sequences 6 Interactions 248 species 114 structures

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

Interactions

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

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



Family: Kinesin (PF00225)

126 architectures 4150 sequences 6 Interactions 248 species 114 structures

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

enter ID/acc

Go

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 [PDBe](#) group, to allow us to map Pfam domains onto UniProt sequences and three-dimensional protein structures. The table below shows the structures on which the **Kinesin** domain has been found.

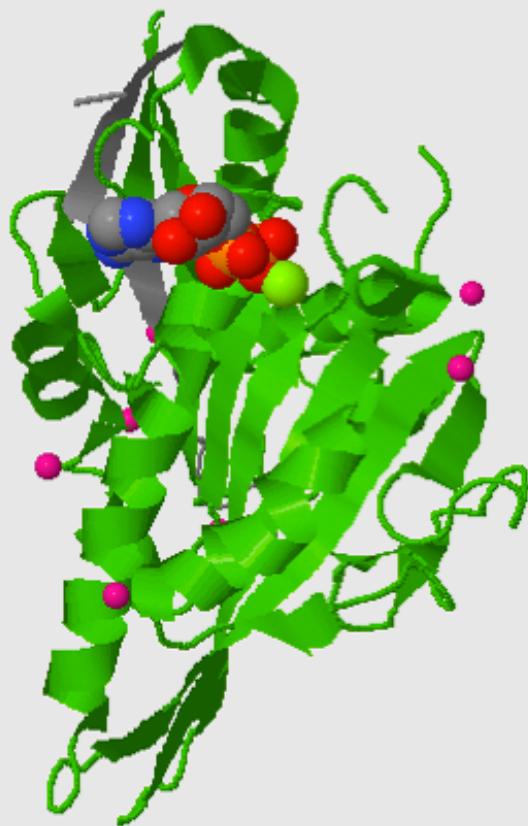
UniProt entry	UniProt residues	PDB ID	PDB chain ID	PDB residues	View
A8BKD1_GIALA	11 - 335	2vvg	A	11 - 335	Jmol AstexViewer SPICE
			B	11 - 335	Jmol AstexViewer SPICE
CENPE_HUMAN	12 - 329	1t5c	A	12 - 329	Jmol AstexViewer SPICE
			B	12 - 329	Jmol AstexViewer SPICE
KAR3_YEAST	392 - 723	1f9t	A	392 - 723	Jmol AstexViewer SPICE
		1f9u	A	392 - 723	Jmol AstexViewer SPICE
		1f9v	A	392 - 723	Jmol AstexViewer SPICE
		1f9w	A	392 - 723	Jmol AstexViewer SPICE
		3kar	B	392 - 723	Jmol AstexViewer SPICE
KI13B_HUMAN	11 - 352	3gbi	A	11 - 352	Jmol AstexViewer SPICE
			B	11 - 352	Jmol AstexViewer SPICE
			C	11 - 352	Jmol AstexViewer SPICE
		1ii6	A	24 - 359	Jmol AstexViewer SPICE
			B	24 - 359	Jmol AstexViewer SPICE
		1g0b	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: Family: Kinesin (PF00225)

Pfam: Jmol



PDB entry 3bfm



Jmol

Your turn:

What can you find out about “eg5”

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

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 Gene, UniProt, PDB databases as well as a number of ‘boutique’ databases including PFAM and OMIM.
- Introduced the notion of *controlled vocabularies* and *ontologies*.

HOMEWORK

https://bioboot.github.io/bggn213_f17/lectures/#1

- Complete the **initial course questionnaire**:
- Check out the “**Background Reading**” material online:
- Complete the **lecture 1 homework questions**:

THANK YOU