



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 <i>what, why</i> and <i>how</i> 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 with the address bar displaying `bioboot.github.io/bggn213_f17/`. The browser's navigation bar includes links for Home, Gmail, Gcal, Bitbucket, GitHub, News, and Disqus. The page content is divided into a left sidebar and a main content area. The sidebar features the UC San Diego logo, the course title 'BGGN 213', a descriptive paragraph, and a list of navigation links: Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, and Screen Cast Videos. At the bottom of the sidebar are icons for Twitter, GitHub, Email, and RSS. The main content area has a large heading 'Foundations of Bioinformatics (BGGN 213, Fall 2017)' with a DNA double helix icon containing the numbers '101' and '110'. Below the heading is a light gray box containing the 'Course Director' (Prof. Barry J. Grant), 'Instructional Assistant' (Ileena Mitra), and 'Course Syllabus' (Fall 2017 PDF) sections. The 'Overview' section below this box describes the course's focus on big data in biosciences and lists major topics, including 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

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/

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

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,
- Genome informatics

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

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

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.

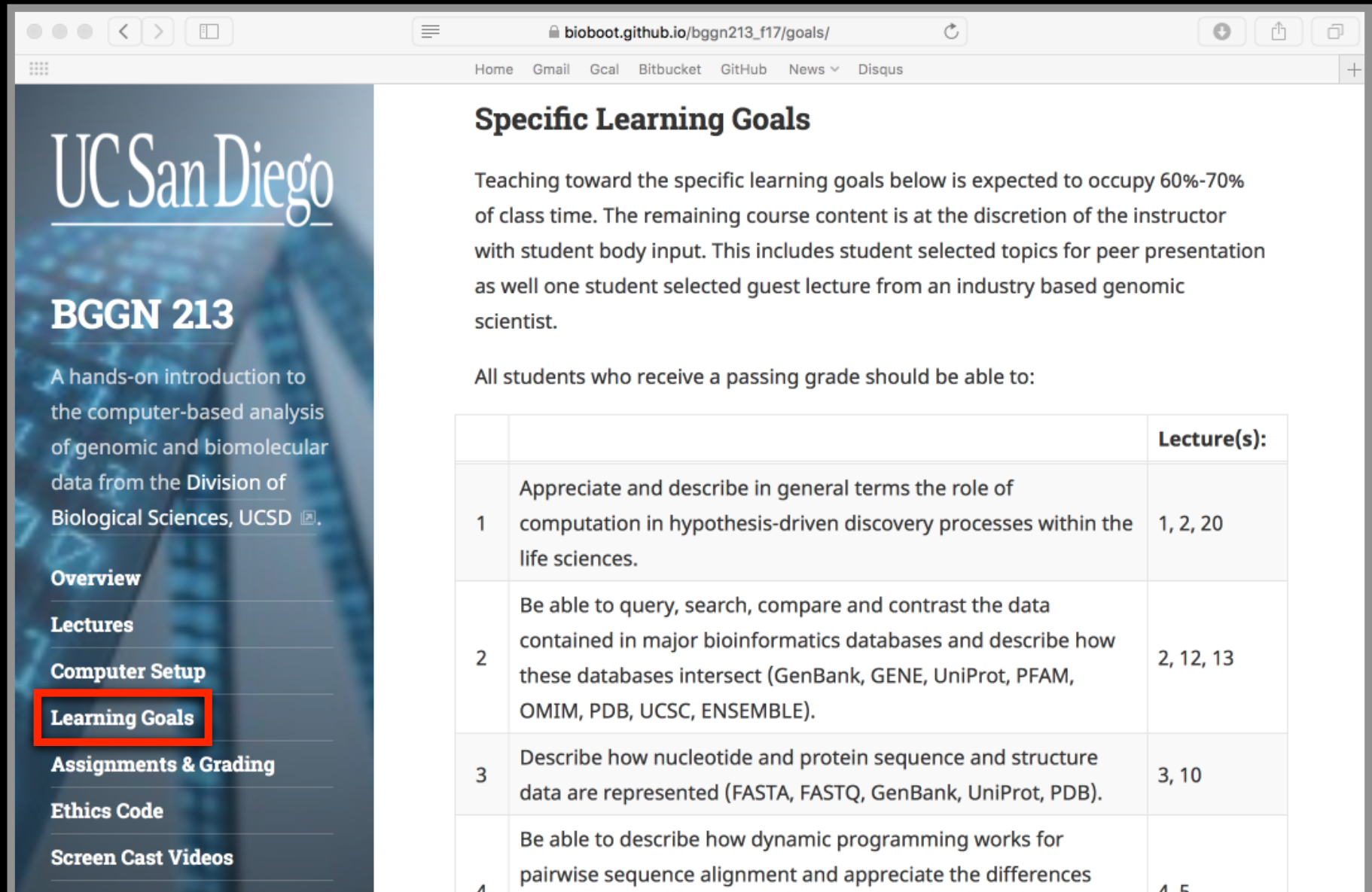
At the end of this course students will:

- 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!



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

Specific Learning Goals

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.

All students who receive a passing grade should be able to:

		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

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 [\[Map\]](#).

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

Lectures

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

Course Structure

Derived from specific learning goals

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 [\[Map\]](#).

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

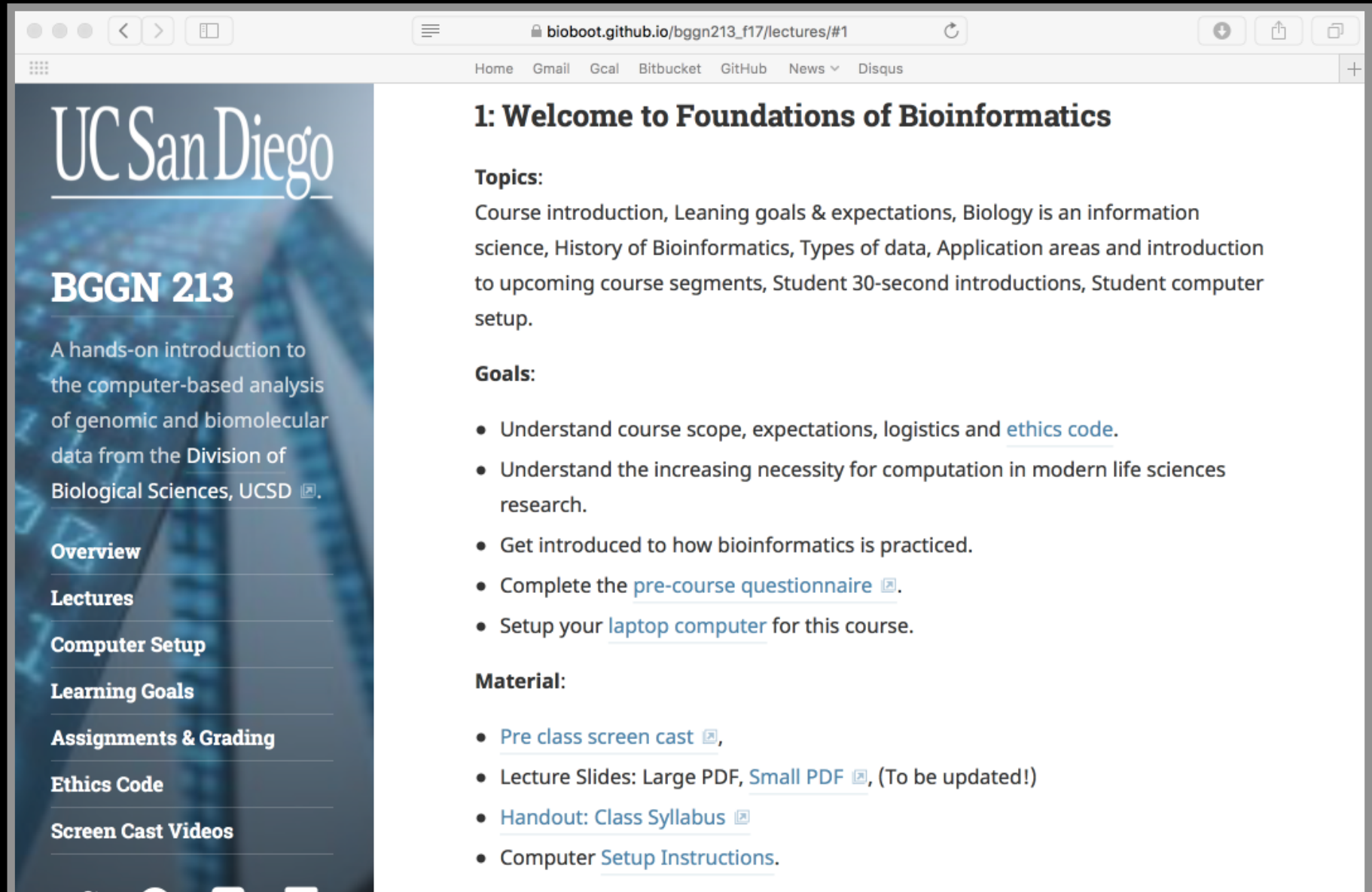
Lectures

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



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 [\[x\]](#).

- 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:

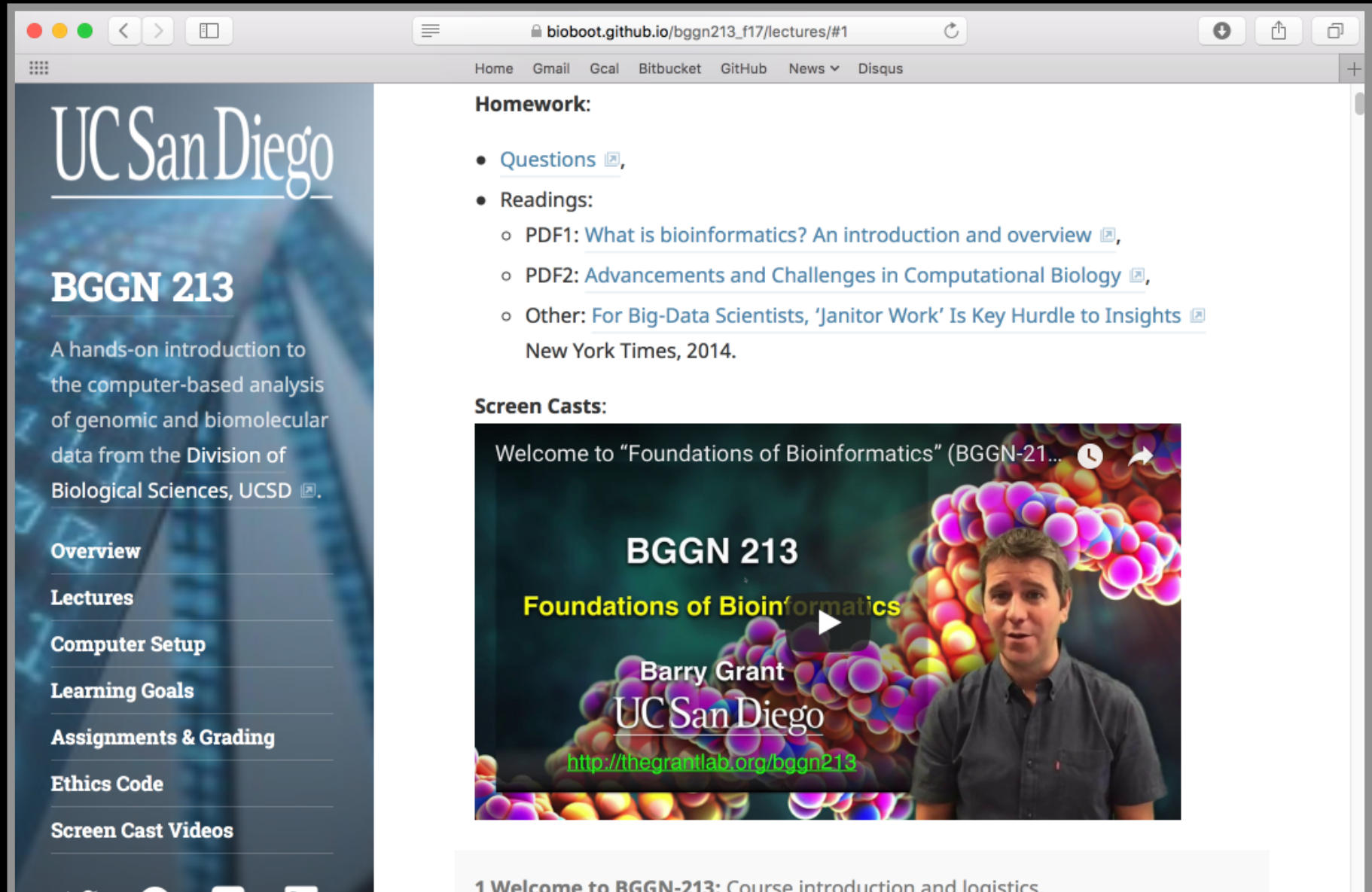
- Understand course scope, expectations, logistics and [ethics code](#).
- Understand the increasing necessity for computation in modern life sciences research.
- Get introduced to how bioinformatics is practiced.
- Complete the [pre-course questionnaire](#) [\[x\]](#).
- Setup your [laptop computer](#)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

Material:

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

Homework

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



The screenshot shows a web browser window with the address bar displaying `bioboot.github.io/bgggn213_f17/lectures/#1`. The browser's address bar also shows navigation icons and a search icon. 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 [\[external link\]](#).

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

Homework:

- [Questions](#) [\[external link\]](#),
- Readings:
 - PDF1: [What is bioinformatics? An introduction and overview](#) [\[external link\]](#),
 - PDF2: [Advancements and Challenges in Computational Biology](#) [\[external link\]](#),
 - Other: [For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights](#) [\[external link\]](#) New York Times, 2014.

Screen Casts:



1 Welcome to BGGN-213: Course introduction and logistics.

Homework

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

The screenshot shows a web browser window with the address bar displaying `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 [\[link\]](#).

Overview

Lectures

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

Screen Cast Videos

Homework:

- Questions** [\[link\]](#)
- Readings:**
 - PDF1: [What is bioinformatics? An introduction and overview](#) [\[link\]](#)
 - PDF2: [Advancements and Challenges in Computational Biology](#) [\[link\]](#)
 - Other: [For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights](#) [\[link\]](#) New York Times, 2014.

Screen Casts:

Welcome to "Foundations of Bioinformatics" (BGGN-21... [\[link\]](#) [\[share\]](#)

BGGN 213

Foundations of Bioinformatics

Barry Grant
UC San Diego

<http://thegrantlab.org/bggn213>

1 Welcome to BGGN-213: Course introduction and logistics.

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 (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

Which of the following databases contains information about the

Homework

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

The image shows a browser window displaying a Google Form titled "BGGN213 Lecture 1 Homework". The form includes a navigation bar with links for Home, Gmail, Gcal, Bitbucket, GitHub, News, and Disqus. The main content area has a purple header and a white form body. A prominent red banner with white text is overlaid diagonally across the form, stating "Homework is due before the next weeks class!". Below the banner, the form asks for the user's name/email address, with a note that it should be the part of the UCSD email address before the '@ucsd.edu' part. A "Your answer" field is provided. The first question is a multiple-choice question: "Which of the following operating systems is most frequently used for bioinformatics tool development". The options are Windows, iOS, Unix, and Perl, each with an unselected radio button.

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

Which of the following databases contains information about the

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 <i>what, why</i> and <i>how</i> 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

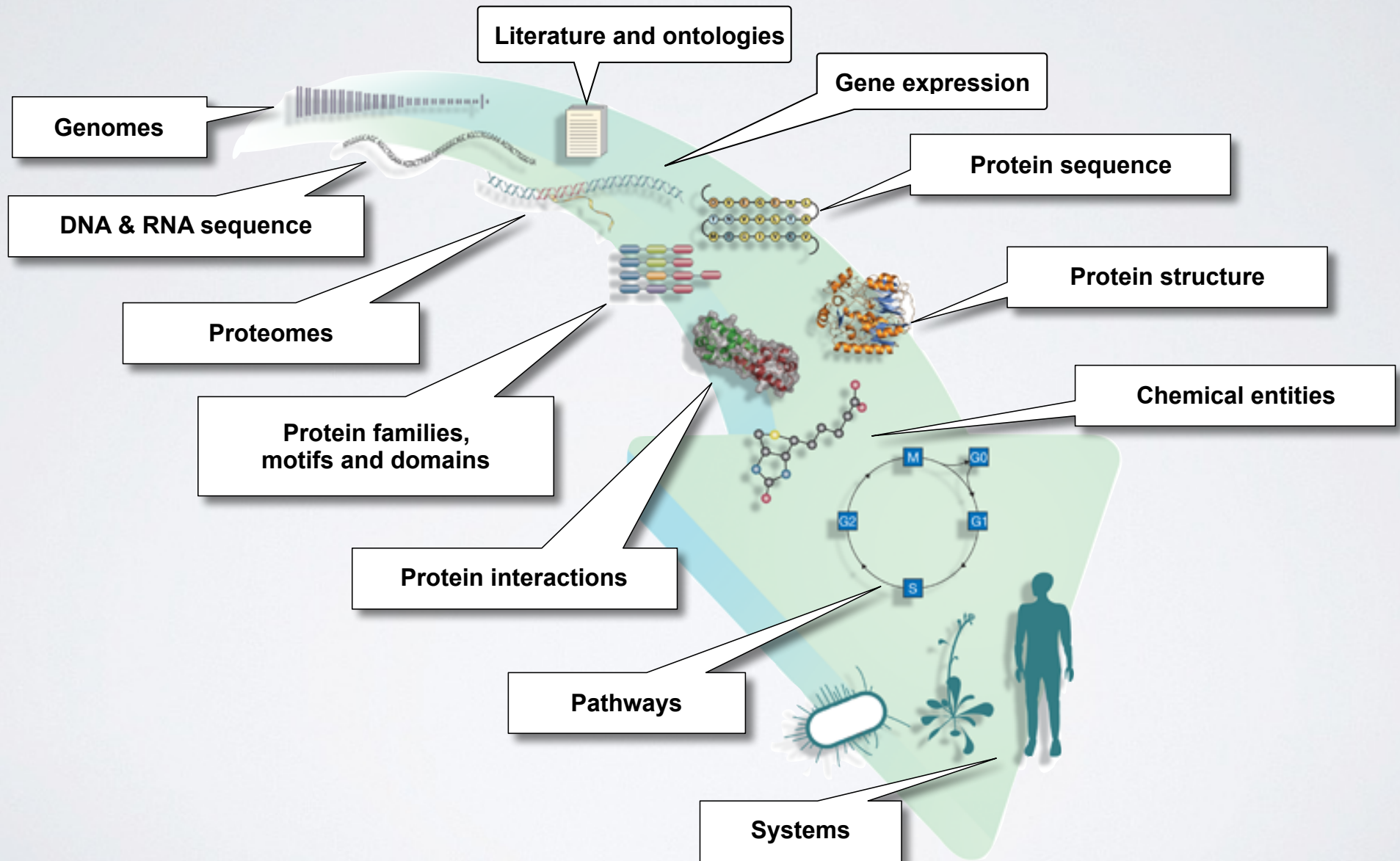
- ▶ “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “**informatics**” **techniques** (derived from disciplines such as applied maths, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a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

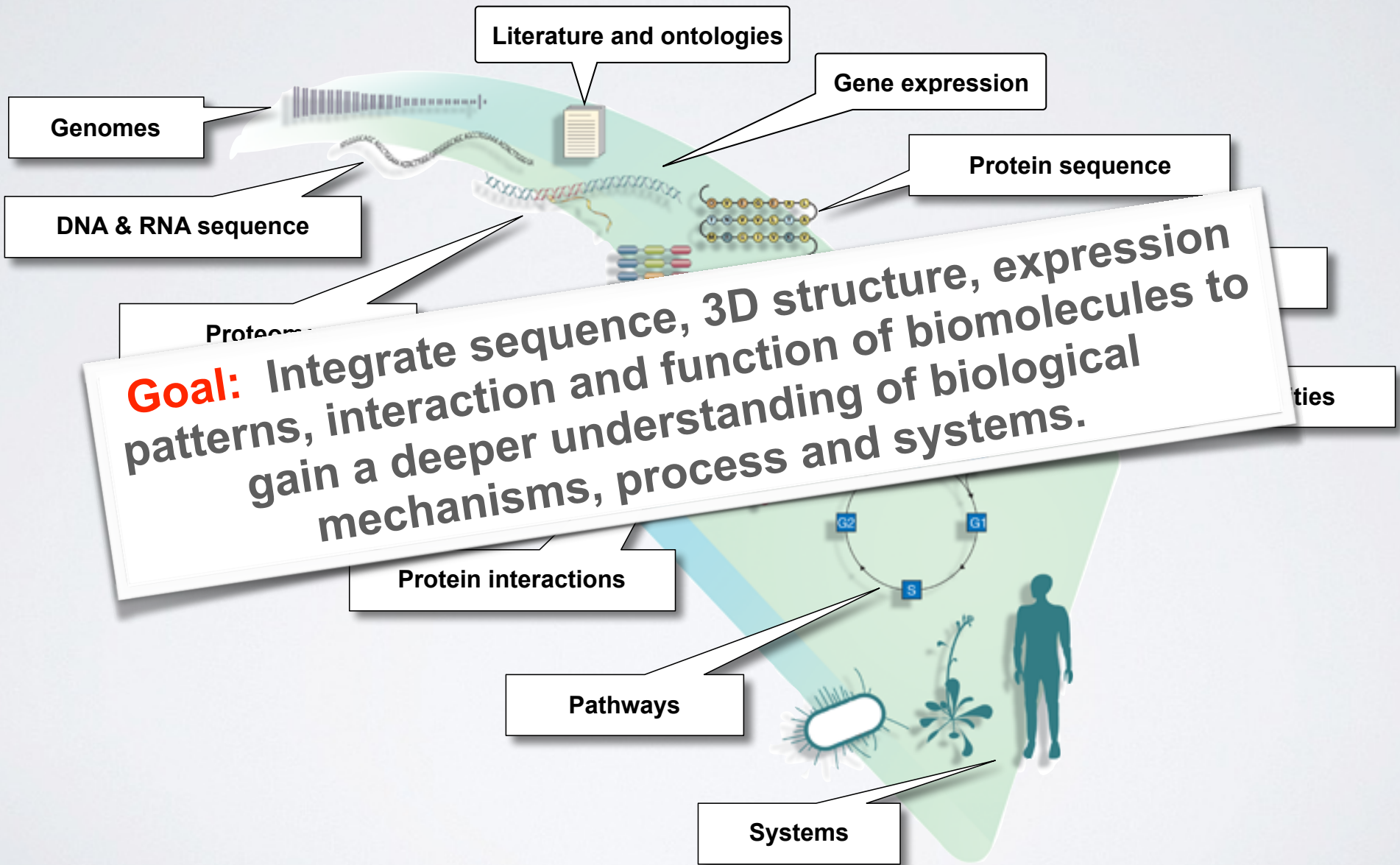
- ▶ “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “informatics” techniques (derived from disciplines such as applied mathematics, computer science, and statistics) to **understand and analyze** the information associated with these molecules, on a **large-scale**.
Luscombe NM, et al. Methods Mol Biol 2001;40:346.
- ▶ “Bioinformatics is the 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

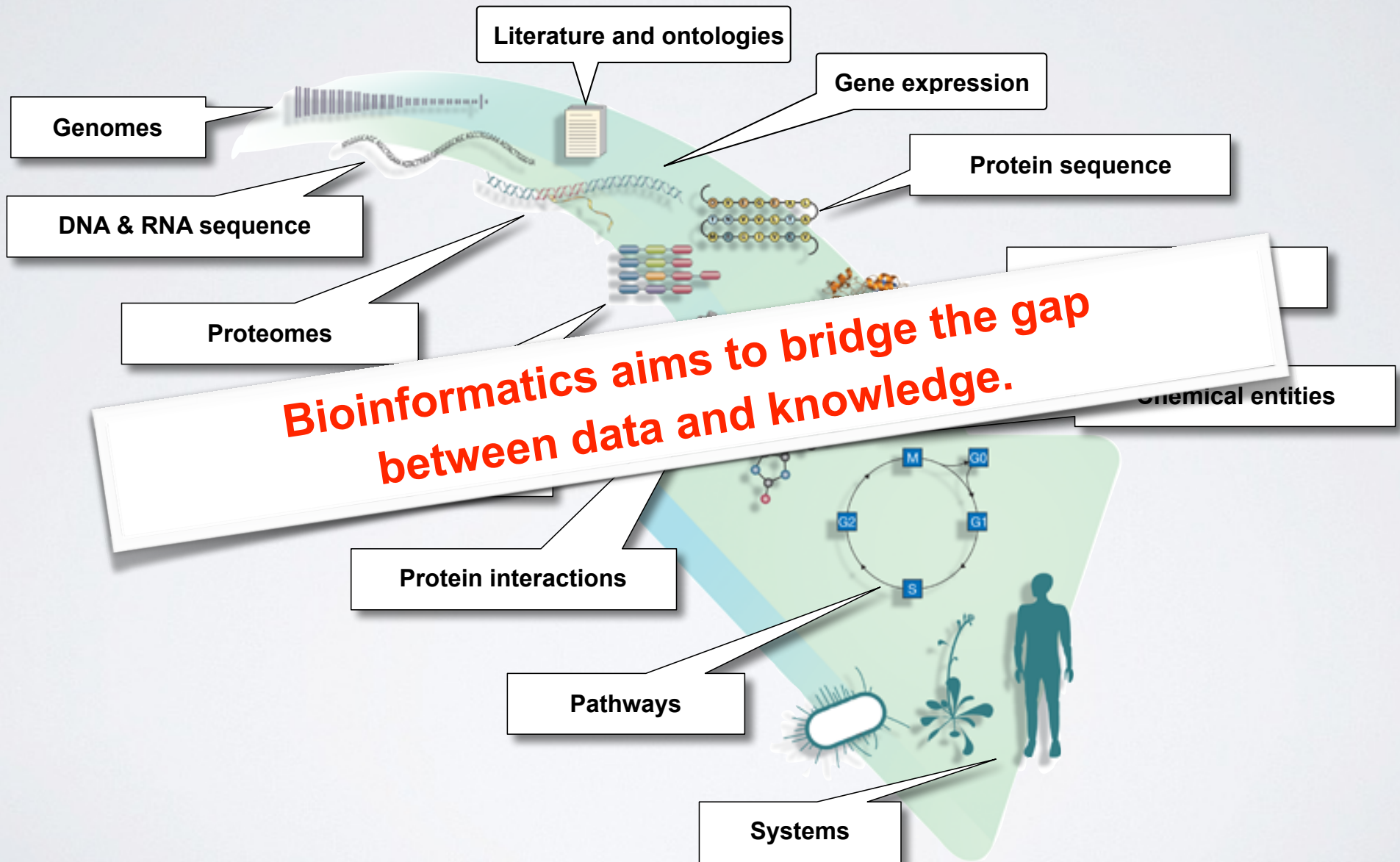
Major types of Bioinformatics Data



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

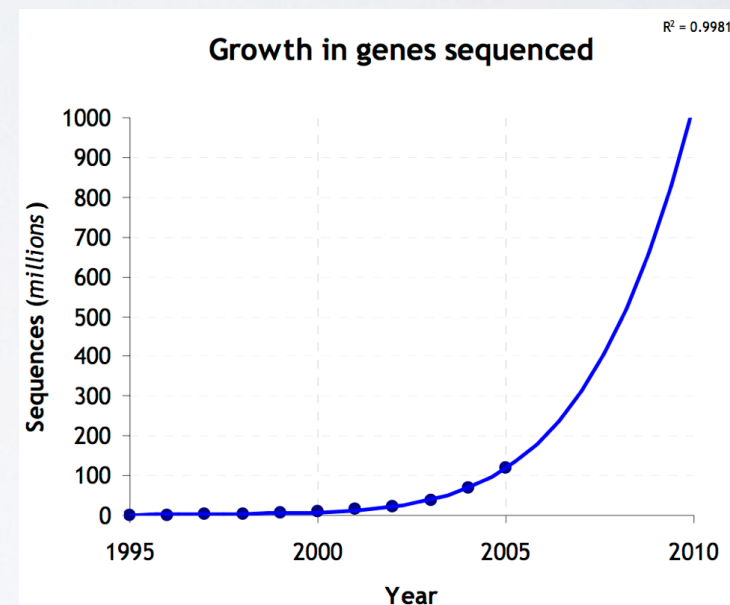
- *DNA sequence determines protein sequence.*
- *Protein sequence determines protein structure.*
- *Protein structure determines protein function.*
- *Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.*

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

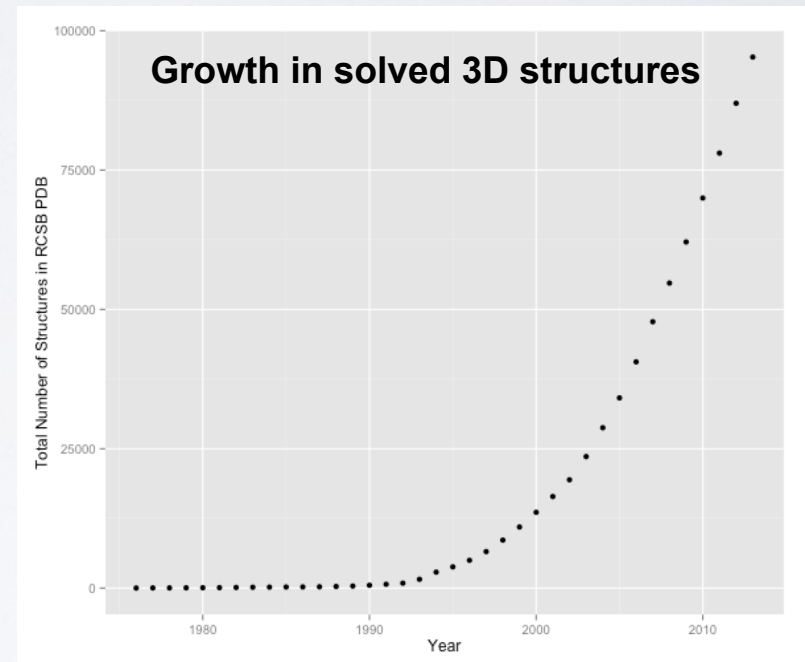


E.G. data from sequencing, structural genomics, proteomics, new high throughput assays, etc...

Why do we need Bioinformatics?

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

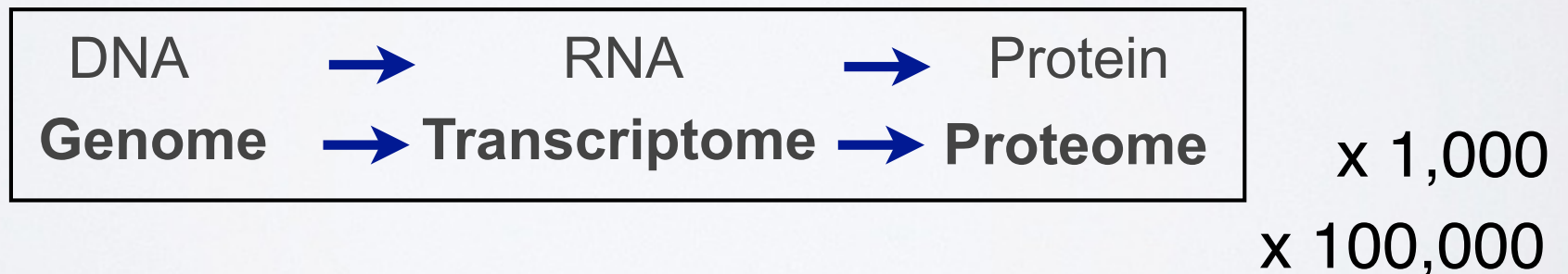
- Bioinformatics provides methods for the efficient:
 - ▶ **storage**
 - ▶ **annotation**
 - ▶ **search and retrieval**
 - ▶ **data integration**
 - ▶ **data mining and analysis**



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.



How do we actually do Bioinformatics?

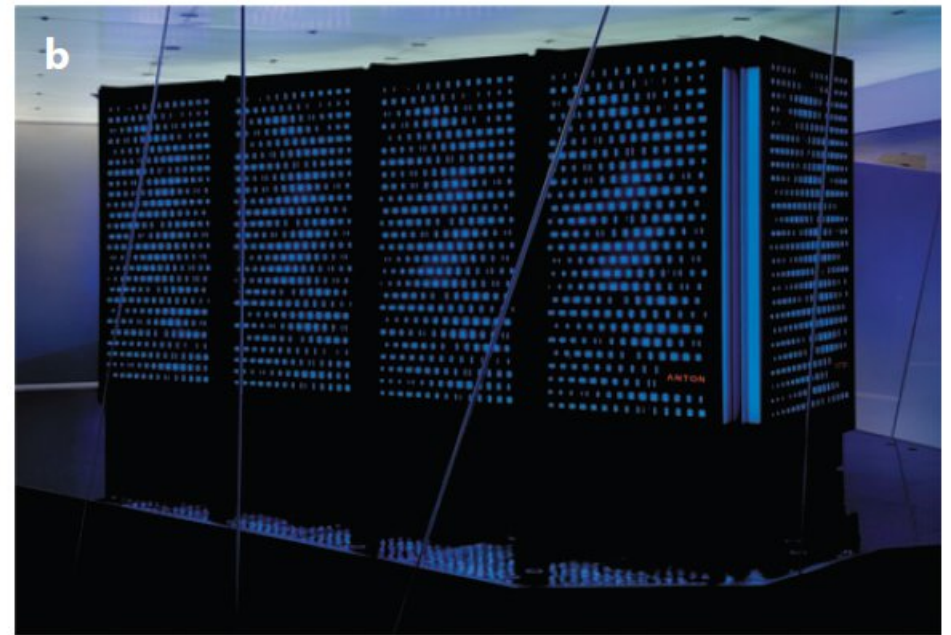
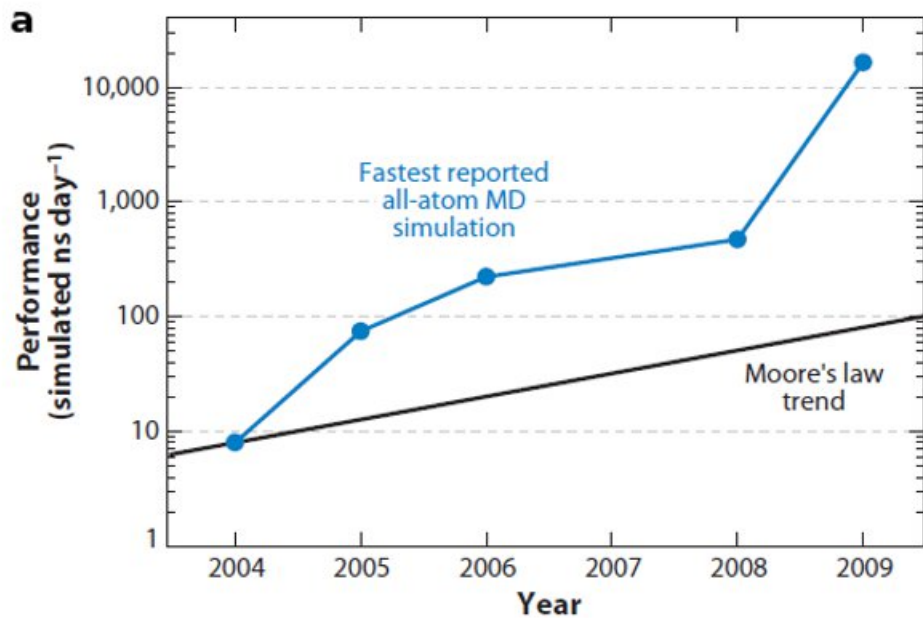
Pre-packaged tools and databases

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

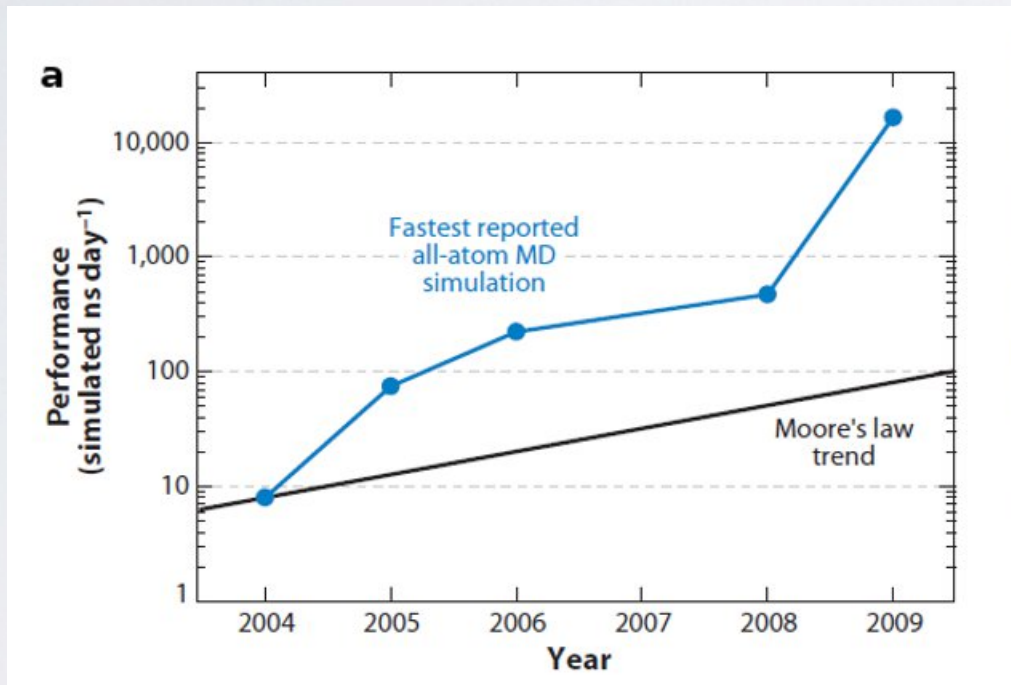
Tool development

- ▶ 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	\$40M	0.1 MHz	1 MB	HALL
2013	\$4,000	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 shoebox.



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 Existence: 11 Extension: 1

Compositional adjustments

Filters and Masking

Filter Low complexity regions

Mask Mask for lookup table only
 Mask lower case letters

PSI/PHI/DELTA BLAST

Upload PSSM no file selected

PSI-BLAST Threshold

Pseudocount

Even Blast has many settable parameters

STEP 3 - Set your

PROGRAM

Related tools with different terminology

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

Key Online Bioinformatics Resources: NCBI & EBI

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

The screenshot shows the NCBI website homepage. The browser address bar displays 'www.ncbi.nlm.nih.gov'. The page features a navigation menu on the left with categories like 'All Resources', 'Data & Software', 'Genes & Expression', and 'Proteins'. The main content area includes a 'Welcome to NCBI' message, a 'Get Started' section with links to 'Tools', 'Downloads', 'How-To's', and 'Submissions', and a '3D Structures' section. A 'Popular Resources' sidebar lists 'PubMed', 'Bookshelf', 'PubMed Central', 'PubMed Health', 'BLAST', 'Nucleotide', 'Genome', 'SNP', 'Gene', 'Protein', and 'PubChem'. The page also includes 'NCBI Announcements' and a search bar at the top.

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

The screenshot shows the EBI website homepage. The browser address bar displays 'www.ebi.ac.uk'. The page features a navigation menu at the top with 'Services', 'Research', 'Training', and 'About us'. The main content area includes a 'Find a gene, protein or chemical:' search bar with a search button and examples like 'blast', 'keratin', and 'bft1...'. Below the search bar are sections for 'Services', 'Research', 'Training', 'Industry', 'European Coordination', and 'EMBL ALUMNI'. The page also includes 'News from EMBL-EBI' and 'Upcoming events' such as 'Plant and Animal Genome conference (PAG XXIV)' and 'SME Forum 2016'. The page also features a 'Visit EMBL.org' section and a 'Popular' sidebar with links to 'Services', 'Research', 'Training', 'News', 'Jobs', 'Visit us', 'EMBL', and 'Contacts'.

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

National Center for Biotechnology Information (NCBI)

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



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

National Center for Biotechnology Information

www.ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

All Databases Search

NCBI Home

Resource List (A-Z)

- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

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

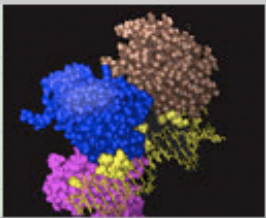
[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

Get Started

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

3D Structures

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



Popular Resources

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

NCBI Announcements

New version of Genome Workbench available

06 Sep

An integrated, downloadable applicati

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

The image shows a screenshot of the National Center for Biotechnology Information (NCBI) website. The browser address bar displays www.ncbi.nlm.nih.gov. The page header includes the NCBI logo and navigation links for 'Resources' and 'How To'. A search bar is visible on the right side of the header.

The main content area features a 'Welcome to NCBI' message and a 'Get Started' section with links to 'Tools', 'Downloads', 'How-To's', and 'Submissions'. A '3D Structures' section is also visible at the bottom.

A white overlay box titled 'Popular Resources' is positioned in the center-right of the page. It lists the following resources: PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. Red arrows point to 'PubMed', 'BLAST', and 'SNP'. A red bracket groups 'Nucleotide', 'Genome', 'SNP', 'Gene', and 'Protein', with an arrow pointing to the 'SNP' link.

On the left side of the page, there is a 'Resource List (A-Z)' menu with the following items: 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.

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

The image shows a screenshot of the National Center for Biotechnology Information (NCBI) website. The browser address bar displays "www.ncbi.nlm.nih.gov". The page header includes the NCBI logo, navigation links for "Resources" and "How To", and a "Sign in to NCBI" link. A search bar is present with a dropdown menu set to "All Databases" and a "Search" button. Below the header, there are navigation links for "NCBI Home", "Resource List (A-Z)", "Welcome to NCBI", "Popular Resources", and "PubMed". A large text overlay in the center of the page reads: "Notable NCBI databases include: **GenBank**, **RefSeq**, **PubMed**, dbSNP and the search tools **ENTREZ** and **BLAST**". The bottom of the page features a sidebar with categories like "Homology", "Literature", "Proteins", "Sequence Analysis", "Taxonomy", "Training & Tutorials", and "Variation". The main content area includes a section for "3D Structures" with a molecular model image and a "Protein" section with a "PubChem" link. An "NCBI Announcements" section is also visible, mentioning a new version of Genome Workbench.

Key Online Bioinformatics Resources: NCBI & EBI

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

The screenshot shows the NCBI website homepage. The browser address bar displays 'www.ncbi.nlm.nih.gov'. The page features a navigation menu on the left with categories like 'NCBI Home', 'Resource List (A-Z)', and 'All Resources'. The main content area includes a 'Welcome to NCBI' message, a 'Get Started' section with links to Tools, Downloads, How-Tos, and Submissions, and a '3D Structures' section. A 'Popular Resources' sidebar lists various databases like PubMed, Bookshelf, and BLAST. The footer contains 'NCBI Announcements' and a '3D Structures' section with a molecular model.

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

The screenshot shows the EBI website homepage. The browser address bar displays 'www.ebi.ac.uk'. The page features a navigation menu on the left with categories like 'Services', 'Research', 'Training', and 'About us'. The main content area includes a 'The European Bioinformatics Institute' header, a search bar, and a 'Find a gene, protein or chemical:' section. The page also features a 'Popular' section, 'Upcoming events' section, and a 'SME Forum 2016' section. The footer contains 'EMBL ALUMNI' and 'Periodic complete'.

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

European Bioinformatics Institute (EBI)

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



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

The screenshot shows the EMBL-EBI website homepage. At the top, the browser address bar shows 'www.ebi.ac.uk'. The main header features the EMBL-EBI logo and navigation links for 'Services', 'Research', 'Training', and 'About us'. The main title is 'The European Bioinformatics Institute', with the subtitle 'Part of the European Molecular Biology Laboratory'. Below this, a paragraph 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.' A search bar is present with the text 'Find a gene, protein or chemical:' and a 'Search' button. Below the search bar are examples: 'blast, keratin, bf1...'. A grid of six service tiles is shown: 'Services' (highlighted with a red border), 'Research', 'Training', 'Industry', 'European Coordination', and 'EMBL ALUMNI'. To the right, a 'Popular' section lists links for Services, Research, Training, News, Jobs, Visit us, EMBL, and Contacts. Below this is a 'Visit EMBL.org' section with the EMBL 40th anniversary logo (1974-2014). The 'Upcoming events' section features a banner for the 'Plant and Animal Genome conference (PAG XXIV)' on Sunday 10 - Tuesday 12 January 2016. The bottom of the page shows a row of three small image thumbnails.

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

The screenshot shows the EBI Services website. The browser address bar displays 'www.ebi.ac.uk/services'. The page features a teal header with the 'Services' title and navigation tabs for 'Services', 'Research', 'Training', and 'About us'. Below the header, there are navigation links for 'Overview', 'A to Z', 'Data submission', and 'Support'. The main content area is titled 'Bioinformatics services' and includes a paragraph describing the availability of molecular databases. To the right, a 'Popular' section lists tools like Ensembl, UniProt, PDBe, ArrayExpress, ChEMBL, BLAST, Europe PMC, Reactome, Train online, and Support. Below this is a 'Service news' section with a butterfly image. At the bottom, a 'Training' section is visible with a laptop and whiteboard image.

Services < EMBL-EBI

www.ebi.ac.uk/services

Services Research Training About us

Services

Overview | A to Z | Data submission | Support

Bioinformatics services

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

DNA & RNA
genes, genomes & variation

Gene expression
RNA, protein & metabolite expression

Proteins
sequences, families & motifs

Structures
Molecular & cellular structures

Systems
reactions, interactions & pathways

Chemical biology
chemogenomics & metabolomics

Ontologies
taxonomies & controlled vocabularies

Literature
Scientific publications & patents

Cross domain
cross-domain tools & resources

Popular

- Ensembl
- UniProt
- PDBe
- 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. The main navigation bar includes 'Services', 'Research', 'Training', and 'About us'. The 'Services' section is highlighted, and the 'Bioinformatics services' category is selected. A grid of service tiles is displayed, with 'Proteins' highlighted by a red box. A 'Popular' sidebar lists several databases, with 'Ensembl' and 'UniProt' highlighted by a red box. A 'Training' banner is visible at the bottom right.

Services < EMBL-EBI

www.ebi.ac.uk/services

Services Research Training About us

Services

Overview A to Z Data submission Support

Bioinformatics services

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

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

Programmatic access

Popular

- Ensembl**
- UniProt**
- PDB**
- ArrayExpress**
- ChEMBL**








Training

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

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

Proteins

Popular services

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

Quick links

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

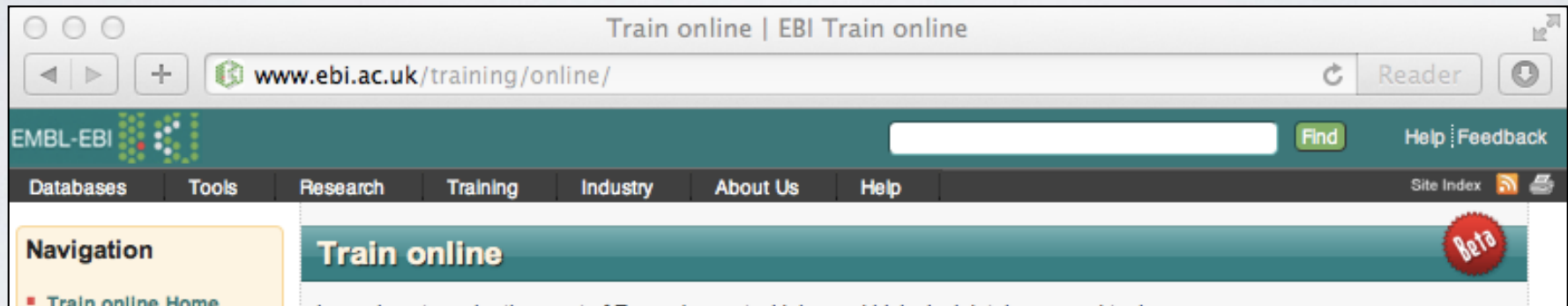
The EBI also provides a growing selection of **online tutorials** on EBI databases and tools

The screenshot shows the EMBL-EBI website homepage. At the top, there is a navigation bar with the EMBL-EBI logo and links for Services, Research, Training, and About us. The main header features the text "The European Bioinformatics Institute" and "Part of the European Molecular Biology Laboratory". Below this, a paragraph 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." A search bar is prominently displayed with the text "Find a gene, protein or chemical:" and a "Search" button. Below the search bar, there are several navigation tiles: Services, Research, Training (highlighted with a red border), Industry, European Coordination, and EMBL ALUMNI. To the right, a "Popular" section lists links for Services, Research, Training, News, Jobs, Visit us, EMBL, and Contacts. Below this is a "Visit EMBL.org" section with the EMBL 40th anniversary logo (1974-2014). The "Upcoming events" section features a banner for the "Plant and Animal Genome conference (PAG XXIV)" held from Sunday 10 to Tuesday 12 January 2016. The bottom of the page shows a row of three small image thumbnails.

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

The screenshot shows a web browser window with the URL www.ebi.ac.uk/training/online/course/using-sequence-similarity-searching-tools-emb-ebl. The page features a navigation menu with 'Services', 'Research', 'Training', and 'About us'. The main heading is 'Train online'. Below this, there are links for 'Training', 'Train online Home', 'Course list', 'Glossary', 'Support & Feedback', and 'Log in / Register'. The breadcrumb trail is 'training » online » course-list » using-sequence-similarity-searching-tools-emb-ebl'. The page title is 'Using sequence similarity searching tools at EMBL-EBI: webinar'. On the left, there is a 'Course content' section with a highlighted item 'Using sequence similarity searching tools at EMBL-EBI: webinar' and a 'Contributors' section. Below that is a 'Print Course' link. The main content area contains a video player with a thumbnail showing the webinar title, subtitle 'Finding homologous sequences with BLAST, FASTA, PSI-Search etc.', and the presenter's name 'Andrew Cowley' with his email and support address. The video player shows a progress bar at 0:00 / 37:42. On the right, there are 'Popular' and 'Find us at...' sections with various links like 'Train online', 'Find us', 'Funding', 'Open days and career days', 'Conference exhibitions', 'EMBL courses and events', 'Genome campus events', and 'Science for schools'. At the bottom, a text box states: 'This webinar focuses on how to use tools like **BLAST** and PSI-Search to find homologous sequences in EMBL-EBI databases, including tips on which tool and database to use, input formats, how to change parameters and how to interpret the results pages.'

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



Notable EBI databases include:

[ENA](#), [UniProt](#), [Ensembl](#)

and the tools [FASTA](#), [BLAST](#), [InterProScan](#),

[MUSCLE](#), [DALI](#), [HMMER](#)

Find a course

Browse by subject



[Genes and Genomes](#)



[Gene Expression](#)



[Interactions, Pathways and Networks](#)

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

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCCP, Beanref, BiImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVM, TKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CAP, ChickGBASE, Colibri, COPE, CottonDB, dbEST, dbSTS, DDBJ, DGP, DictyDb, ECGC, EC02DBASE, FlyBase, G, H, K, MHC, Myc, PD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-MODEL Repository, SWISS-PROT, TeIDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc !!!!

There are lots of Bioinformatics Databases

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

< [Handout Major Databases.pdf](#) >

Side-note: Databases come in all shapes and sizes



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

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/>

The image shows a screenshot of the National Center for Biotechnology Information (NCBI) website. The browser's address bar displays www.ncbi.nlm.nih.gov/. The NCBI logo and navigation links are visible at the top. A search bar is present, with the text "All Databases" and a dropdown menu showing "ras" highlighted by a red box. A blue "Search" button is to the right of the search bar. A diagonal banner with red text reads "Hands on demo (or see following slides)". The main content area features a "Welcome to NCBI" message and a "Get Started" section with links for "About the NCBI", "Mission", "Organizations", and "NCBI News". A "Resources" list is on the right, including "PubMed", "Bookshelf", "PubMed Central", "PubMed Health", "BLAST", "Nucleotide", "Genome", "SNP", "Gene", "Protein", and "PubChem". A "Genotypes and Phenotypes" section is visible at the bottom, with a diagram showing a family tree and a newspaper clipping.

Search NCBI databases

[Help](#)

ras

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

Literature

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

Health

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

Genes

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

Proteins

Show additional filters

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

- Filters: Manage Filters
- Top Organisms [Tree]
 - Homo sapiens (1126)**
 - Mus musculus (823)
 - Rattus norvegicus (625)
 - Oreochromis niloticus (533)
 - Neolamprologus brichardi (507)
 - All other taxa (82019)
- More...

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

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

Find related data

Database: Select

Find items

Search details

ras[All Fields] AND alive[property]

- Clear all
- Gene sources
- Genomic
 - Mitochondria
 - Organelles
 - Plasmids
 - Plastids
- Categories
- Alternatively spliced
 - Annotated genes
 - Non-coding
 - Protein-coding
 - Pseudogene
- Sequence content
- CCDS
 - Ensembl
 - RefSeq

Gene [Help](#)

[Show additional filters](#)

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.

Filters: [Manage Filters](#)

Find related data

Database:

Search details

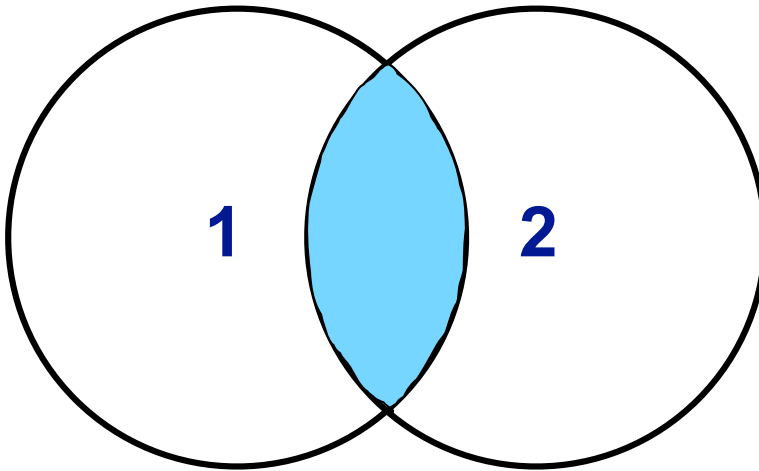
ras[All Fields] AND "Homo sapiens"[porgn] AND alive[property] [See more...](#)

Recent activity

- [Clear all](#)
- Gene sources**
 - Genomic
- Categories**
 - Alternatively spliced
 - Annotated genes
 - Non-coding
 - Protein-coding
 - Pseudogene
- Sequence content**
 - CCDS
 - Ensembl
 - RefSeq
- Status**
- Current only**
- Chromosome locations**
 - Select

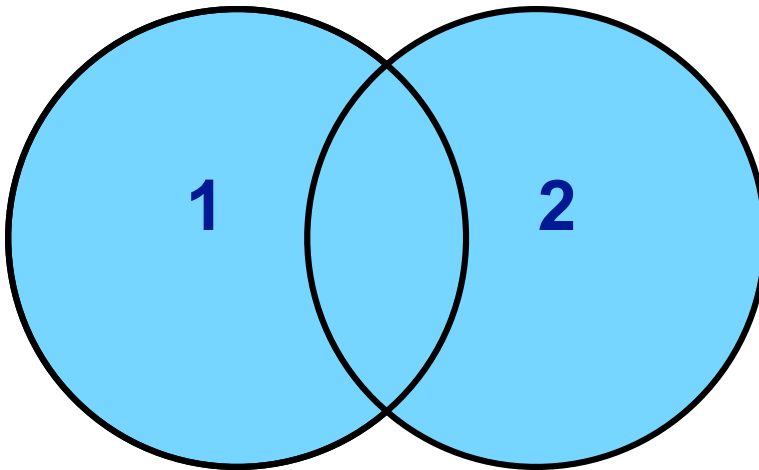
Name/Gene ID	Description	Location	Aliases
<input type="checkbox"/> NRAS ID: 4893	neuroblastoma RAS viral (v-ras) oncogene homolog [Homo sapiens (human)]	Chromosome 1, NC_000001.11 (114704464..114716894, complement)	RP5-1000E10.2, ALPS4, CMNS, N-ras, NCMS1, NS6, NRAS
<input type="checkbox"/> KRAS ID: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

1 AND 2



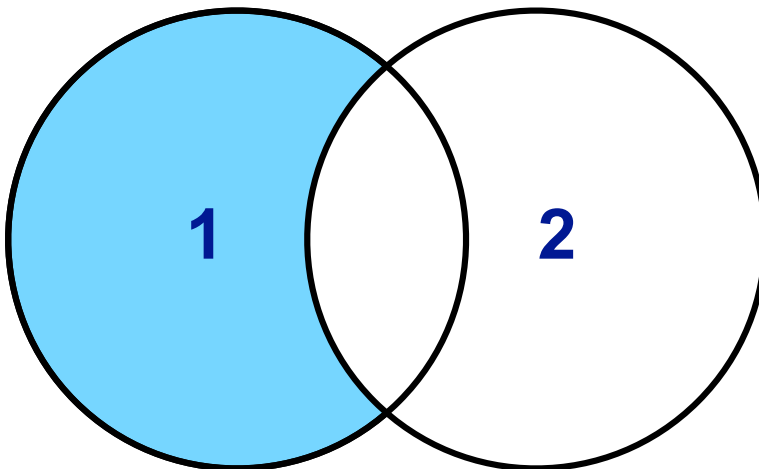
**ras AND disease
(1185 results)**

1 OR 2



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

1 NOT 2



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

Show additional filters

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

Hide sidebar >>

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

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

Filters: Manage Filters

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

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

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

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

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

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

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

Summary

Official Symbol KRAS provided by [HGNC](#)
Official Full Name Kirsten rat sarcoma viral oncogene homolog provided by [HGNC](#)
Primary source [HGNC:HGNC:6407](#)
See related [Ensembl:ENSG00000133703](#); [HPRD:01817](#); [MIM:190070](#);
[Vega:OTTHUMG00000171193](#)
Gene type protein coding
RefSeq status REVIEWED
Organism [Homo sapiens](#)
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; 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)



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

NCBI Resources How To Sign in to NCBI

Gene Search Help

Display S Hide sidebar >>

KRAS
(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)
- Related sequences

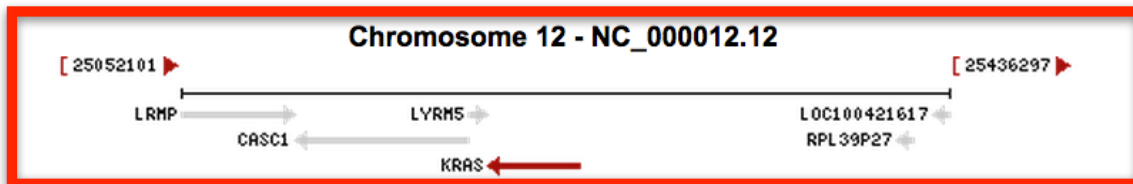
Genomic context

Location: 12p12.1

Exon count: 6

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

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



Genomic regions, transcripts, and products

Go to [reference sequence details](#)

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

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

- BioAssay by Target (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

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

NCBI Resources How To Sign in to NCBI

Gene

Search Help

Display Settings

Hide sidebar >>

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

Table of contents

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

KRAS Kirsten rat sarcoma (human)]
Gene ID: 3845

Summary

Official Symbol KRAS provided by HGNC

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

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

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

Gene type protein coding

RefSeq status REVIEWED

Organism [Homo sapiens](#)

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

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

Gene Ontology [Provided by GOA](#)

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

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


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



GO: Gene Ontology

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

The screenshot shows the UniProt-GOA website interface. At the top, there is a navigation bar with links for Services, Research, Training, and About us. Below this is a search bar with a 'Search' button and examples of search terms: GO:0006915, tropomyosin, P06727. The main heading is 'Gene Ontology Annotation (UniProt-GOA) Database'. The introductory text explains that the UniProt GO annotation program aims to provide high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB). It mentions that the assignment of GO terms to UniProt records is an integral part of UniProt biocuration, and that UniProt manual and electronic GO annotations are supplemented with manual annotations supplied by external collaborating GO Consortium groups. A 'Menu' section on the right lists various resources: Downloads, Searching UniProt-GOA, Annotation Methods, Annotation Tutorial, Manual Annotation Efforts, Reference Genome Annotation Initiative, Cardiovascular Gene Ontology Annotation Initiative, Renal Gene Ontology Annotation Initiative, and Exosome Gene.

EMBL-EBI  Services Research Training About us

UniProt-GOA

Examples: [GO:0006915](#), [tropomyosin](#), [P06727](#)

[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

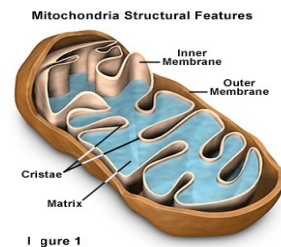
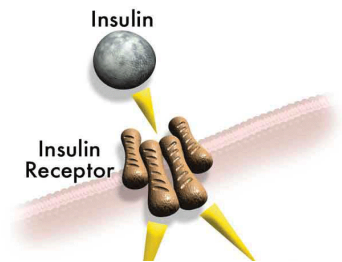
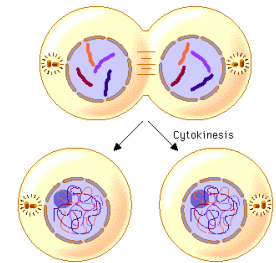
- [Downloads](#)
- [Searching UniProt-GOA](#)
- [Annotation Methods](#)
- [Annotation Tutorial](#)
- [Manual Annotation Efforts](#)
- [Reference Genome Annotation Initiative](#)
- [Cardiovascular Gene Ontology Annotation Initiative](#)
- [Renal Gene Ontology Annotation Initiative](#)
- [Exosome Gene](#)

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



Gene Ontology [Provided by GOA](#)

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

Process

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

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

⋮ Scroll down to
▼ **UniProt** link

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

The screenshot shows the NCBI Gene page for KRAS (3845). The browser address bar shows the URL: www.ncbi.nlm.nih.gov/gene/3845#gene-ontology. The page displays genomic information for X01669.1 and CAA25828.1. A table lists protein accession numbers and their corresponding UniProtKB links. The UniProtKB link for P01116.1 is highlighted with a red box. A red arrow points to the UniProt link with the text: "Scroll down to Very bottom for UniProt link".

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

Additional links

You are here: [NCBI](#) > [Genes & Expression](#) > [Gene](#) [Write to the Help Desk](#)

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

UniProt will detail much more information for protein coding genes

UniProtKB Advanced

BLAST Align Retrieve/ID Mapping Help Contact

P01116 - RASK_HUMAN

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

BLAST Align Format Add to basket History Feedback Help video

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).

Enzyme regulationⁱ

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

Regions

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Nucleotide binding ⁱ	10 - 18	9	GTP <input type="button" value="2 Publications"/>			
Nucleotide binding ⁱ	29 - 35	7	GTP <input type="button" value="2 Publications"/>			
Nucleotide binding ⁱ	59 - 60	2	GTP <input type="button" value="2 Publications"/>			

P01116 - RASK_HUMAN

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

Display None

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

Functionⁱ

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

Enzyme regulationⁱ

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

Regions

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

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?

Display None **Pathology & Biotech¹**

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

[▲ Top](#)

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?

Display None **Structure¹**

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

[Show more details](#)

3D structure databases

Select the link destinations:
 PDBe¹
 RCSB PDB¹
 PDBj¹

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	1-164	[>>]
4EPX	X-ray	1.76	A	1-164	[>>]
4EPY	X-ray	1.80	A	1-164	[>>]
4L8G	X-ray	1.52	A	1-164	[>>]
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 website interface. At the top, there is a navigation bar with 'RCSB PDB', 'Deposit', 'Search', 'Visualize', and 'Analyze'. Below this is the PDB logo and the text 'An Information Portal to 133759 Biological Macromolecular Structures'. A search bar is present with the text 'Search by PDB ID, author, macromolecule, sequence, or ligand' and a 'Go' button. Below the search bar are links for 'Advanced Search' and 'Browse by Annotations'. The main navigation bar includes 'Structure Summary', '3D View' (highlighted with a red box), 'Annotations', 'Sequence', 'Sequence Similarity', 'Structure Similarity', and 'Experiment'. Below this is a 'Literature' tab. The main content area shows the protein ID '4EPV' and the title 'Discovery of Small Molecules that Bind to K-Ras and Inhibit Sos-mediated Activation'. To the left of the text is a 3D ribbon diagram of the protein structure. Below the title are fields for 'Display Files' and 'Download Files'. The text includes 'DOI: 10.2210/pdb4epv/pdb', 'Classification: HYDROLASE', 'Deposited: 2012-04-17 Released: 2012-05-23', 'Deposition author(s): Sun, Q., Burke, J.P., Phan, J., Burns, M.C., Olejniczak, E.T., Waterson, A.G., Lee, T., Rossanese, O.W., Fesik, S.W.', 'Organism: Homo sapiens', 'Expression System: Escherichia coli', and 'Mutation(s): 1'. At the bottom, there are sections for 'Experimental Data Snapshot' (Method: X-RAY DIFFRACTION), 'wwPDB Validation' (with a table for Metric, Percentile Ranks, and Value), and '3D Report' and 'Full Report' buttons.

Structure Summary **3D View** Annotations Sequence Sequence Similarity Structure Similarity Experiment

Literature

Biological Assembly 1

4EPV

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

DOI: 10.2210/pdb4epv/pdb

Classification: [HYDROLASE](#)

Deposited: 2012-04-17 Released: 2012-05-23

Deposition author(s): [Sun, Q.](#), [Burke, J.P.](#), [Phan, J.](#), [Burns, M.C.](#), [Olejniczak, E.T.](#), [Waterson, A.G.](#), [Lee, T.](#), [Rossanese, O.W.](#), [Fesik, S.W.](#)

Organism: [Homo sapiens](#)

Expression System: Escherichia coli

Mutation(s): 1

View in 3D: [NGL](#) or [JSmol](#) (in Browser)

Experimental Data Snapshot

Method: X-RAY DIFFRACTION

wwPDB Validation

Metric	Percentile Ranks	Value
--------	------------------	-------

3D Report Full Report

Contact Us

Lets view the 3D structure:

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

www.rcsb.org

Home Gmail

RCSB PDB Deposit Search Visualize Analyze

4EPV

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

Note: Use your mouse to drag, rotate, and zoom in and out of the structure. Click to identify atoms and bonds.

Bond: [GLY]12:A:O - [GLY]12:A:C

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

Viewer Options

Contact Us

Back to UniProt:

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

Display None

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

[Top](#)

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]

Sequences (2)ⁱ

Sequence status¹: Complete.
Sequence processing¹: The displayed sequence is further processed into a mature form.
This entry describes **2** isoforms¹ produced by **alternative splicing**. [Align](#)

Example Questions:

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

EMBL-EBI  HOME

Family: Ras (PF00071)

332 architectures 21243 sequences 30 interactions 1006 species 663 structures

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

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



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... ?

Species distribution

This visualisation provides a simple graphical representation of the distribution of this family across species. You can find the original interactive tree in the [adjacent tab](#). [More...](#)

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... ?

Summary
Domain organisation
Clan
Alignment
HMM log
Trees
Curation
Species
Interact
Structure
Jump to
enter ID/acc

Species distribution

Pfam: Pfam alignment viewer
pfam.xfam.org/family/PF00071/alignment/view?jobId=EDCA403E-9836-11E4-B360-10B3298E2F76

EMBL-EBI

Alignment for selected sequences

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

```
P11234/16-178 ..KIVMVGSGGVCSSALTI...Q...FM...Y.D..EF.V...E.DYEPFK.-AD...SYRKKVLD...
P01112/5-165 ..KLVVVGAGGVCSSALTI...Q...LI...Q.N..HF.V...D.EYDPTI.-ED...SYRKQVVID...
Q14088/38-204 ..KIIIVIGDSNVGTCCLTF...R...FC...G..G..TF.P...D.KTEATI.GVD...FREKTVEIE...
Q9BW83/7-173 ..KCIILAGDPVAGCTALAQ...I...FR...S..DgaHF.Q...K.SYTLTT.GMD...LVVKTVPVEd...
P15153/5-178 ..KCVVVGDCGAVGTCLLI...S...YT...T.N..AF.P...G.EYIPTV.-FD...NYSANVMVD...
O00194/11-183 ..KLLALGDSGVCSTFFLY...R...YT...D.N..KF.N...P.KFITV.GID...FREKRVVYNagqpn...
Q15907/13-174 ..KVVLLGDSGVCSSNLLS...R...FT...R.N..EF.N...L.ESKSTI.GVE...FATRSIQVD...
P10114/5-166 ..KVVVLGSGGVCSSALTV...Q...Q...FV...T..G..TF.I...E.KYDPTI.-ED...FYRKEIEVD...
P51153/10-171 ..KLLILGDSGVCSTCLII...R...FA...E.D..NF.N...N.TYISII.GID...FKIRTVDIE...
P55040/77-241 ..KVVLLGEGGVCSTLAN...I...FA...GvhD..SM.D...S.D-CEVL.GED...TYERTLMVD...
P55042/93-253 ..KVVLLGAPGVCSSALAR...I...FG...G..V..ED.G...P.EAEAG.--H...TYDRSVVD...
P01116/5-165 ..KLVVVGAGGVCSSALTI...Q...LI...Q.N..HF.V...D.EYDPTI.-ED...SYRKQVVID...
Q9H0T7/21-182 ..KLVLLGSGGVCSSSLAL...R...YV...K.N..DF.K..S.-ILPTV.GCA...FFTRKVDVG...
Q9ULC3/11-171 ..KVVVVGNGAVGSSMIQ...R...YC...K..G..IF.T...K.DYKPTI.GVD...FLERQIQVN...
Q14807/15-177 ..KLVVVGSGGVCSSALTI...Q...FF...Q.K..IF.V...P.DYDPTI.-ED...SYLKHTEID...
Q9NX57/7-202 ..KIVLLGDMNVGTSLLQ...R...YM...E.R..RF.P...D.T-VSTV.GAA...FYLKQW---...
Q9H082/35-201 ..KIIIVIGDSNVGTCCLTY...R...FC...A..G..RF.P...D.RTEATI.GVD...FRERAVEID...
Q969Q5/9-174 ..KVVVLGKEYVGTSLVE...R...YV...H.D..RFLV...G.PYQNTI.GAA...FVAKVMSV...
P51149/10-175 ..KVIILGDSGVCSTSLMN...Q...YV...N.K..KF.S...N.QYKATI.GAD...FLTKRVVVD...
Q9ULW5/65-227 ..KVVMLGDSGVCSTCLL...R...FK...D..G..AF.L..AgTFISTV.GID...FRNKVLVD...
P57735/14-175 ..KVVLLGESGVCSTNLLS...R...FT...R.N..EF.S...H.DSRITI.GVE...FSTRIVML...
P51159/11-183 ..KFLALGDSGVCSTSVLY...Q...YT...D..G..KF.N...S.KFITV.GID...FREKRVVYRasgpd...
P01111/5-165 ..KLVVVGAGGVCSSALTI...Q...LI...Q.N..HF.V...D.EYDPTI.-ED...SYRKQVVID...
P11233/16-177 ..KIVMVGSGGVCSSALTI...Q...FM...Y.D..EF.V...E.DYEPFK.-AD...SYRKKVLD...
Q9UL25/21-182 ..KVVLLGEGCVGTSVLV...R...YC...E.N..KF.N...D.KHITL.QAS...FLTKKLNIG...
Q9NPF2/10-171 ..KILLIGESGVCSSLL...R...FT...D.D..TF.D...P.ELAATI.GVD...FKVKTISVD...
Q9H0U4/10-171 ..KILLIGDSGVCSTL...R...FA...D.D..TY.T...E.SYISII.GVD...FKIRTIELD...
Q9UL26/7-168 ..KVCLLGDTGVCSSIVW...R...FV...E.D..SF.D...P.NINPTI.GAS...FMTKTVOYQ...
Q9UBK7/23-179 ..KIIICLGSVAGSKLME...R...FL...M.D..EF.Q...P.QQLSTY.ALT...LYKHTATVD...
P51157/14-179 ..KVVLLGDSGVCSTLTT...C...FA...Q.Q.E..TF.G...K.QYKATI.GLD...FLRLRITLE...
```

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.

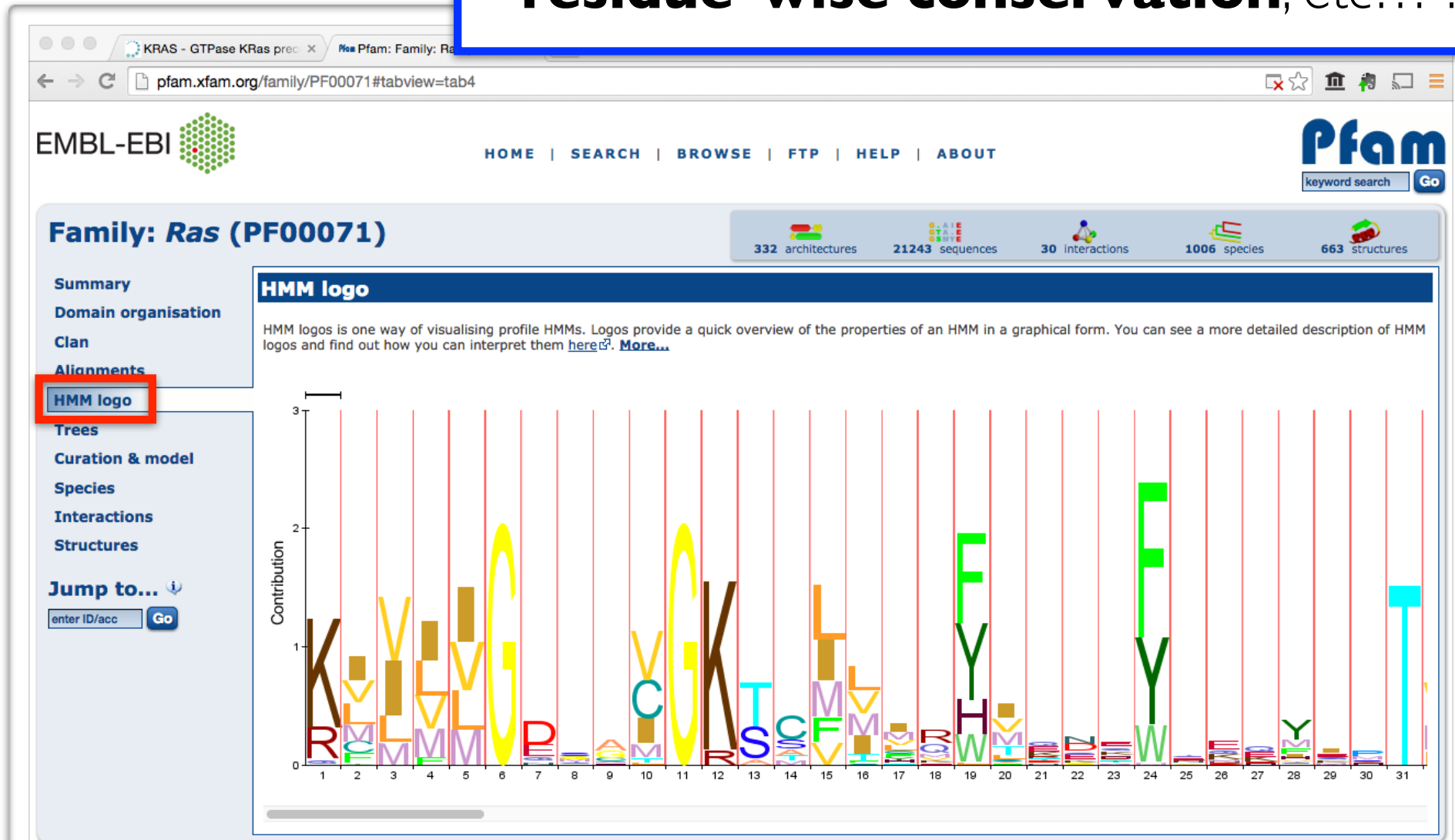
There are 18 pages in this alignment. Show page

[Download](#) this alignment.

Close window


Example Questions:

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



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

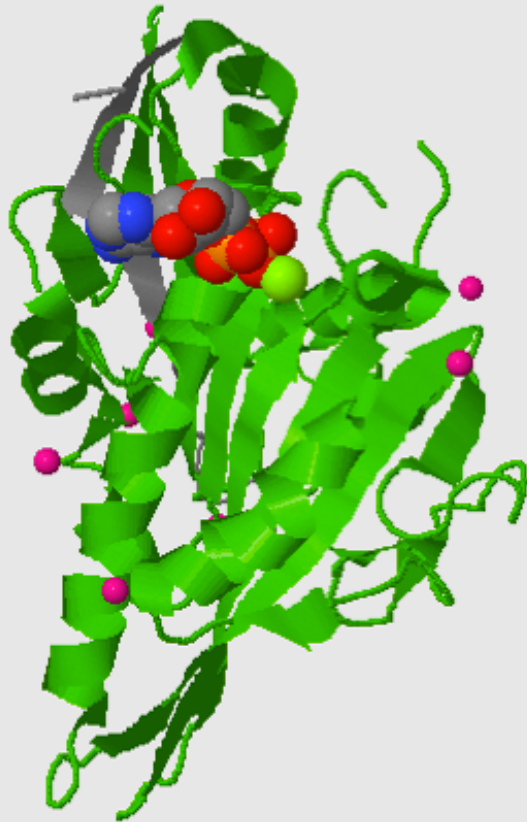
Jump to...

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		
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	
			1f9w	B	392 - 723	Jmol AstexViewer SPICE	
			3kar	A	392 - 723	Jmol AstexViewer SPICE	
KI13B_HUMAN	11 - 352	3qbj	A	11 - 352	Jmol AstexViewer SPICE		
			B	11 - 352	Jmol AstexViewer SPICE		
			C	11 - 352	Jmol AstexViewer SPICE		
		1ii6	A	24 - 359	Jmol AstexViewer SPICE		
			B	24 - 359	Jmol AstexViewer SPICE		
		1q0b	A	24 - 359	Jmol AstexViewer SPICE		
			B	24 - 359	Jmol AstexViewer SPICE		
		1x88	A	24 - 359	Jmol AstexViewer SPICE		
			B	24 - 359	Jmol AstexViewer SPICE		
					A	24 - 359	Jmol AstexViewer SPICE

PDB entry 3bfm



Jmol

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)	

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

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