



**BGGN 213**

**Foundations of Bioinformatics**

**Barry Grant**  
**UC San Diego**

<http://thegrantlab.org/bggn213>

HELLO  
my name is

*BARRY*

[bjgrant@ucsd.edu](mailto:bjgrant@ucsd.edu)

HELLO  
*HIS* — my name is

*YUANSHENG*

[yuz461@ucsd.edu](mailto:yuz461@ucsd.edu)

# Introduce Yourself!

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

# Today's Menu

<b>Course Logistics</b>	Website, screencasts, survey, ethics, assessment and grading.
<b>Learning Objectives</b>	What you need to learn to succeed in this course.
<b>Course Structure</b>	Major lecture topics and specific leaning goals.
<b>Introduction to Bioinformatics</b>	<b>Introducing the <i>what, why</i> and <i>how</i> of bioinformatics?</b>
<b>Bioinformatics Database</b>	<b>Hands-on</b> exploration of several major databases and their associated tools.

<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

Course Director  
[Prof. Barry J. Grant](#) (Email: [bjgrant@ucsd.edu](mailto:bjgrant@ucsd.edu))

Instructional Assistant  
Yuansheng Zhou (Email: [yuz461@ucsd.edu](mailto:yuz461@ucsd.edu))

Course Syllabus  
[Spring 2018 \(PDF\)](#)

### Overview

Bioinformatics - the application of computational and analytical methods to biological problems - is a rapidly maturing field that is driving the collection, analysis, and interpretation of the avalanche of data in modern life sciences and medical research.

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

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

The screenshot shows a web browser window with the URL `binboot.github.io/bgggn213_f17/goals/`. The browser's address bar and navigation buttons are visible at the top. Below the browser window, the page content is displayed. On the left side, there is a navigation menu for the course 'BGGN 213' at UC San Diego. The menu items are: Overview, Lectures, Computer Setup, Learning Goals (highlighted with a red box), Assignments & Grading, Ethics Code, and Screen Cast Videos. The main content area on the right is titled 'Learning Goals' and contains the following text and list:

**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:**

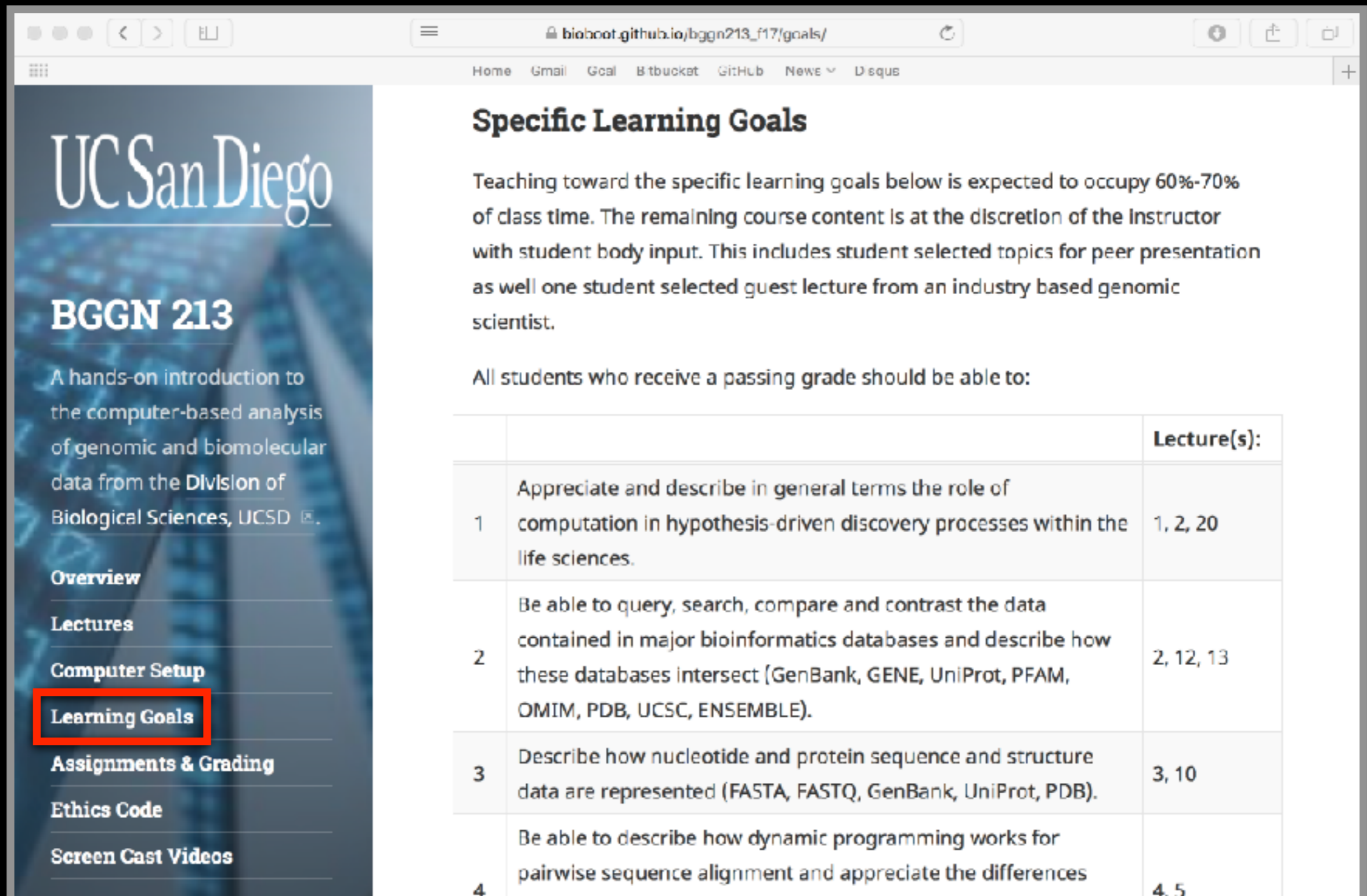
- Understand the increasing necessity for computation in modern life sciences research.
- Be able to use and evaluate online bioinformatics resources and analysis tools to solve problems in the biological sciences.
- Be able to use the R environment to analyze bioinformatics data at scale.
- Be familiar 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/bgggn213_f17/goals/`. The page is titled "Specific Learning Goals" and is part of the BGGN 213 course materials at UC San Diego. The left sidebar contains a navigation menu with the following items: Overview, Lectures, Computer Setup, Learning Goals (highlighted with a red box), Assignments & Grading, Ethics Code, and Screen Cast Videos. The main content area includes an introductory paragraph and a table of learning goals.

## 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

The screenshot shows a web browser window displaying the course page for BGGN 213 at UC San Diego. The page is titled "Lectures" and provides information about the course structure. The "Lectures" link in the left sidebar is highlighted with a red box. The main content area features a table of topics for Spring 2018, with the first row, "Welcome to Bioinformatics", also highlighted with a red box.

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

### Lectures

All Lectures are Wed/Fri 1:00-4:00 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 Spring 2018
1	Wed, 04/04	<b>Welcome to Bioinformatics</b> Course introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Hands on with major Bioinformatics databases and key online NCBI and EBI resources
2	Fri, 04/06	<b>Sequence alignment fundamentals, algorithms and applications</b> Homology, Sequence similarity, Local and global alignment, classic Needleman-Wunsch, Smith-Waterman and BLAST heuristic approaches, Hands on with dot plots, Needleman-Wunsch and BLAST algorithms highlighting their utility and limitations

# Class Details

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 navigation bar includes links for Home, Gmail, Gcal, Bitbucket, GitHub, News, and Disqus. The main content area is titled "1: Welcome to Foundations of Bioinformatics".

**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 [\[i\]](#).

- 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](#) [\[i\]](#).
- Setup your [laptop computer](#) for this course.

**Material:**

- [Pre class screen cast](#) [\[i\]](#),
- Lecture Slides: Large PDF, [Small PDF](#) [\[i\]](#), (To be updated!)
- [Handout: Class Syllabus](#) [\[i\]](#)
- [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_117/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**

**Homework:**

- [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:**

Welcome to "Foundations of Bioinformatics" (BGGN-21...  

**BGGN 213**

**Foundations of Bioinformatics**

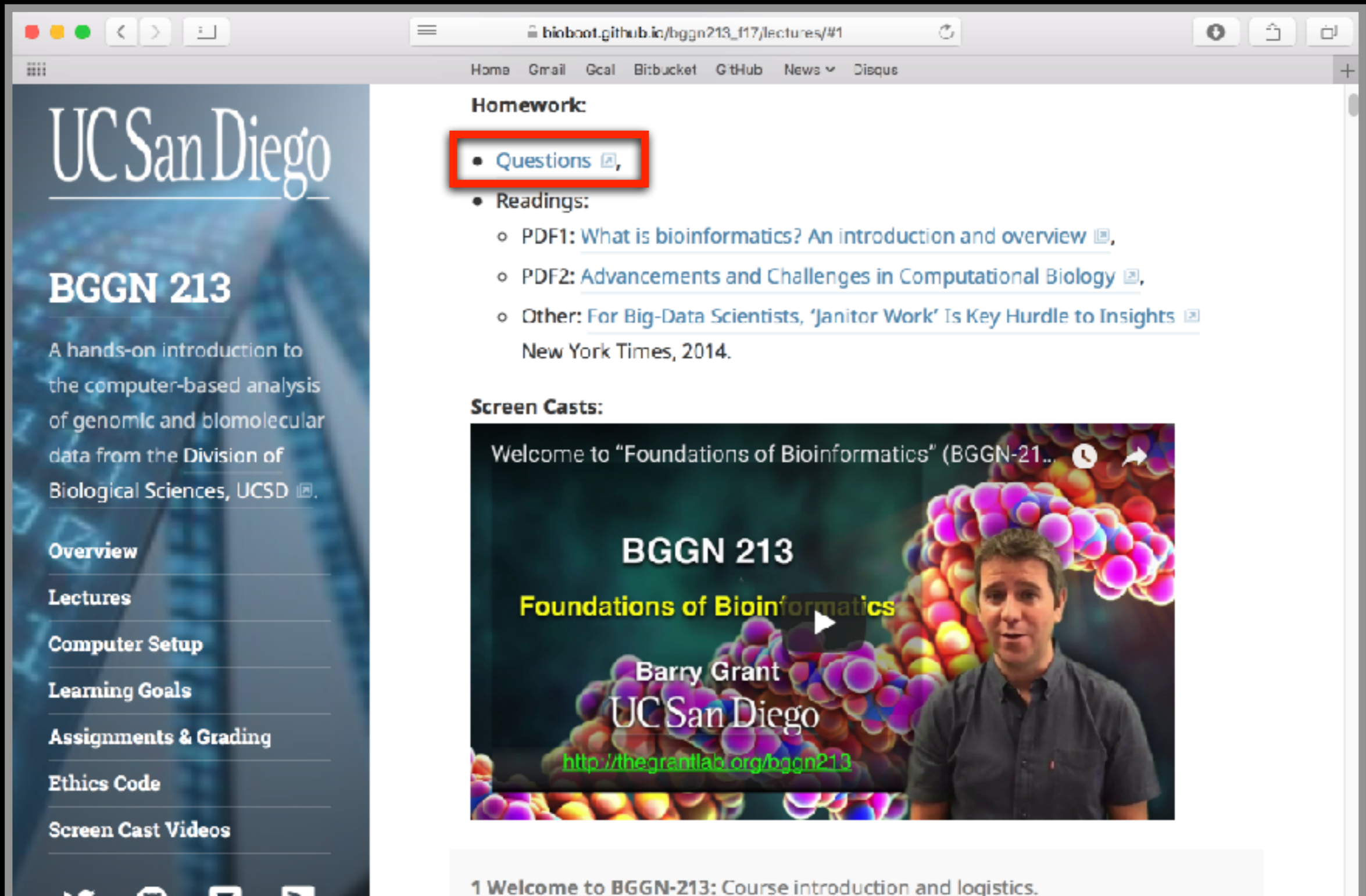
Barry Grant  
UC San Diego

<http://thegrantlab.org/bgggn213>

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

# Homework

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



The screenshot shows a web browser window with the URL `bioboot.github.io/bgggn213_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**

**Homework:**

- **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:**

Welcome to "Foundations of Bioinformatics" (BGGN-213)

**BGGN 213**

**Foundations of Bioinformatics**

Barry Grant  
UC San Diego

<http://thegrantlab.org/bgggn213>

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

# Homework

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

docs.google.com/forms/d/e/1FAIpQLSeN3pg-AaRg5laSPxZuqSj1

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 primarily protein**

# Homework

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

The image shows a screenshot of a Google Forms page titled "BGGN213 Lecture 1 Homework". The browser's address bar shows the URL "docs.google.com/forms/d/e/1FAIpQLSeN3pg-AaRg5laSPxZuqSj1". The page has a purple header and a white content area. A red banner with white text is overlaid diagonally across the page, reading "Homework is due before the next weeks class!". Below the banner, the form asks the user to "Please answer the following questions". A required field is labeled "name/email address" with a red asterisk. Below this field is a text input area labeled "Your answer". 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. The second question is partially visible at the bottom: "Which of the following databases contains primarily protein".

## Side Note: **Why stick with this course?**

**Provides a hands-on practical introduction to major bioinformatics concepts and resources.**

Covers modern hot topics and the intimate coupling of informatics with biology - **highlighting the impact of computing advances and 'big data' on biology!**

Designed for graduates in the biosciences with no programming experience or high level math skills.

Provides a hook for increasing computational and data science competencies in the biosciences - **valuable high demand translational skills!**



# BGGN-213 Learning Goals....

Advanced UNIX and R based learning goals

The screenshot shows a web browser window displaying the BGGN 213 Learning Goals page. The browser address bar shows the URL `bioboot.github.io/bgggn213_f17/goals/`. The page header includes navigation links: Home, Gmail, Gcal, bitbucket, GitHub, News, Disqus, and BGGN-213. The main content area features a table of learning goals. A green box highlights goals 6, 7, 8, and 9. A red arrow points to the bottom right corner of the page.

**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

5	Calculate the alignment score between two nucleotide or protein sequences using a provided scoring matrix and be able to perform BLAST, PSI-BLAST, HMMER and protein structure based database searches and interpret the results in terms of the biological significance of an e-value.	5, 10
6	Use UNIX command-line tools for file system navigation and text file manipulation.	6, 7, 10, 11, 24, 15
7	Use existing programs at the UNIX command line to analyze bioinformatics data.	7, 10, 11, 13, 14, 15, 16
8	Use R to read and parse comma-separated (.csv) formatted files ready for subsequent analysis.	8, 9, 10, 11, 13, 15, 16
9	Perform elementary statistical analysis on biomolecular and "omics" datasets with R and produce informative graphical displays and data summaries.	9, 10, 11, 13, 15, 16
10	View and interpret the structural models in the PDB.	10, 11
11	Explain the outputs from structure prediction algorithms and small molecule docking approaches.	11
12	Appreciate and describe in general terms the rapid advances in sequencing technologies and the new areas of investigation that	13, 14, 15

# BGGN-213 Learning Goals....

Delve deeper into “real-world” bioinformatics

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 [\[i\]](#).

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

13	sequenced and the bioinformatics processing and analysis required for their interpretation.	13
14	For a genomic region of interest (e.g. the neighborhood of a particular gene), use a genome browser to view nearby genes, transcription factor binding regions, epigenetic information, etc.	14
15	Given an RNA-Seq data file, find the set of significantly differentially expressed genes and use online tools to interpret gene lists and annotate potential gene functions.	15, 16
16	Perform a GO analysis to identify the pathways relevant to a set of genes (e.g. identified by transcriptomic study or a proteomic experiment).	16
17	Use the KEGG pathway database to look up interaction pathways.	17
18	Use graph theory to represent biological data networks.	17, 18
19	Understand the challenges in integrating and interpreting large heterogenous high throughput data sets into their functional context.	19
20	Have an appreciation for the social impacts and ethical implications of how genomic sequence information is used in our society	20

# These support a major learning objective

## 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 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.



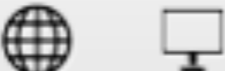





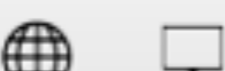

# Why use R?

Productivity

Flexibility

Designed for data analysis

# IEEE 2016 Top Programming Languages

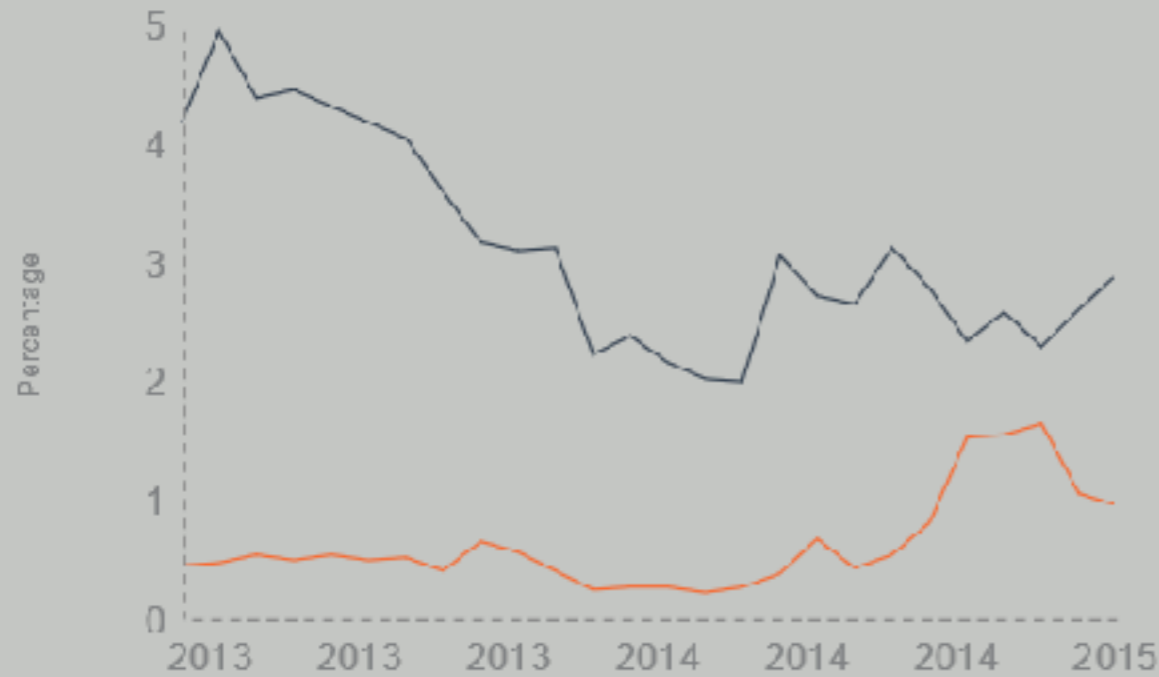
Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

<http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages>

# R and Python: The Numbers

## Popularity Rankings

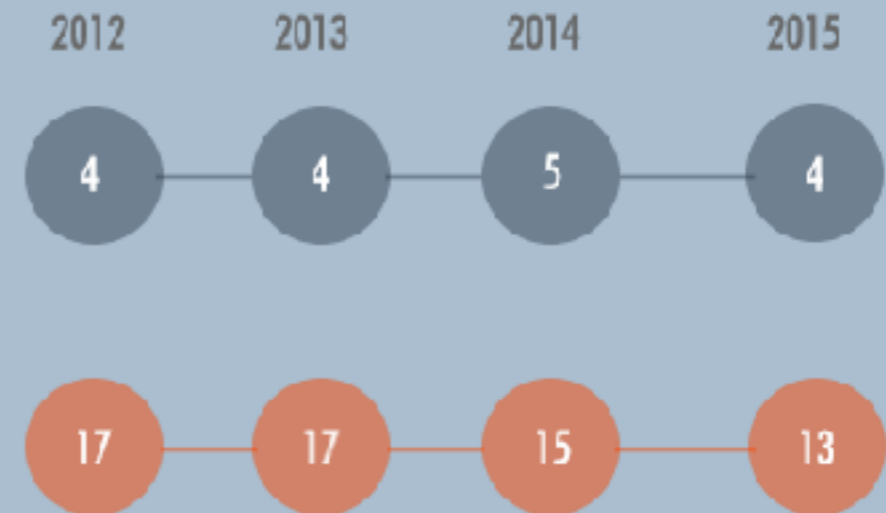
R and Python's popularity between 2013 and February 2015 (TIOBE Index)



Redmonk ranking, comparing the relative performance of programming languages on GitHub and Stack Overflow (September 2012 and January 2013, 2014, 2015)

Python

R



## Jobs And Salary?

2014 Dice Tech Salary Survey:  
Average Salary For High Paying Skills and Experience



\$115,531



\$94,139

[http://www.kdnuggets.com/2015/05/r-vs-python-data-science.html?utm\\_medium=email&utm\\_source=flipboard](http://www.kdnuggets.com/2015/05/r-vs-python-data-science.html?utm_medium=email&utm_source=flipboard)

- R is the “lingua franca” of data science in industry and academia.
- Large user and developer community.
  - As of Jan 8th 2018 there are 12,039 add on **R packages** on CRAN and 1,473 on Bioconductor - more on these later!
- Virtually every statistical technique is either already built into R, or available as a free package.
- Unparalleled exploratory data analysis environment.

< <https://www.datacamp.com/> >

The screenshot shows the DataCamp website interface. At the top, there is a navigation bar with the DataCamp logo, links for 'Learn', 'Groups', and 'About', and a user profile section showing '1,250 XP' and a notification icon with a red badge. A dropdown menu is open from the notification icon, listing several notifications: 'You have a new assignment: Conditionals and Con...', 'You have a new assignment: Working with the RSt...', 'You have a new assignment: Introduction to R', 'bjgrant invited you to the group Foundations o...', and 'You have a new assignment: Orientation'. Below the navigation bar, the main content area features a section titled 'Your Latest Activity' with a card for 'Introduction to Spark in R using'. The card includes a progress indicator and a message: 'You are doing awesome barryus! So far you've earned 250 XP!'. Below this, it mentions the last chapter worked on: 'Light My Fire: Starting To Use Spark With dplyr Syntax'. At the bottom, there is a 'DAILY PRACTICE' section with a sub-header and a paragraph: 'Learning data science requires practice every day. Build your data science fluency with DataCamp practice mode.'



< <https://www.datacamp.com/> >

The image shows a screenshot of a web browser displaying a DataCamp course page on the left and an RStudio terminal window on the right.

**Left Panel (DataCamp Course Page):**

- Header:** DataCamp logo and navigation buttons (back, Course Outline, forward).
- Course Title:** "What is an IDE anyway?" with a "50xp" badge.
- Description:** "RStudio is an IDE that makes R easier to use by combining a set of tools into a single environment. What does IDE stand for?"
- Possible Answers:** A list of radio button options:
  - Intensive Design Environment
  - Integrated Document Environment
  - Independent Developer Ecosystem
  - Integrated Development Environment
- Buttons:** "Take Hint (-15xp)" and "Submit Answer" (circled in red).

**Right Panel (RStudio Terminal):**

```
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help,
or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

< <https://www.datacamp.com/> >

The screenshot shows a web browser window displaying a DataCamp course page. The browser's address bar shows the URL: <https://campus.datacamp.com/courses/working-with-the-rstudio-ide-part-1/orientation?ex=2>. The DataCamp logo is visible in the top left corner, and a 'Course Outline' button is in the top right. The main content area is a dark grey panel with the following elements:

- A title: "What is an IDE anyway?"
- A large heading: "Exercise Completed" with a blue checkmark and "50xp" next to it.
- Text: "Nice job! Move onto the next video to start learning more about the RStudio IDE!"
- A yellow "Continue" button.
- A box titled "Become a power user!" containing the text "Submit Answer" followed by three buttons: "Ctrl", "+", "Shift", "+", "Enter". Below this is a link: "See all keyboard shortcuts".

Overlaid on the right side of the browser is a screenshot of the RStudio IDE interface. The R logo is in the top left of the IDE. The menu bar includes: File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, barryus, Sessions. The toolbar contains icons for file operations and a search bar. The main console window displays the following text:

```
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help,
or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

On the right side of the IDE, there are two panels: "Environment" and "History". The "Environment" panel shows "Global Environment" and "Environment is empty". Below it is a "Files" panel with a tree view showing "Home" and a table with columns "Name" and "Size".

< <https://www.datacamp.com/> >

Back to My Dashboard

## Foundations of Bioinformatics (BGGN-213)

Leaderboard | My Assignments

30 Days | [90 Days](#) | [Last Year](#)

	Member	XP ↕	Courses ↕	Chapters ↕
1	Angela Nicholser	22450	4	20
2	Ben Song	12850	2	11
3	Ana Grant	12120	2	9
4	Delaney Pagluso	12085	2	11
5	oehernan	11055	2	10
6	Erin Schikaris	10350	2	9
7	Zachary Warburg	9110	1	8
8	Alexander Weitzel	6950	1	6

< <https://www.datacamp.com/> >

Back to My Dashboard

## Foundations of Bioinformatics (BGGN-213)

Leaderboard **My Assignments**

Name	Assigned At	Due By	
Conditionals and Control Flow	Oct 2, 2017	Nov 2, 2017	In progress
Introduction to R	Oct 2, 2017	Oct 26, 2017	In progress
Working with the RStudio IDE (Part 1)	Oct 2, 2017	Oct 26, 2017	In progress

LEARN RESOURCES GROUPS ABOUT

# Today's Menu

<b>Course Logistics</b>	Website, screencasts, survey, ethics, assessment and grading.
<b>Learning Objectives</b>	What you need to learn to succeed in this course.
<b>Course Structure</b>	Major lecture topics and specific learning goals.
<b>Introduction to Bioinformatics</b>	Introducing the <i>what, why</i> and <i>how</i> of bioinformatics?
<b>Computer Setup</b>	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> )

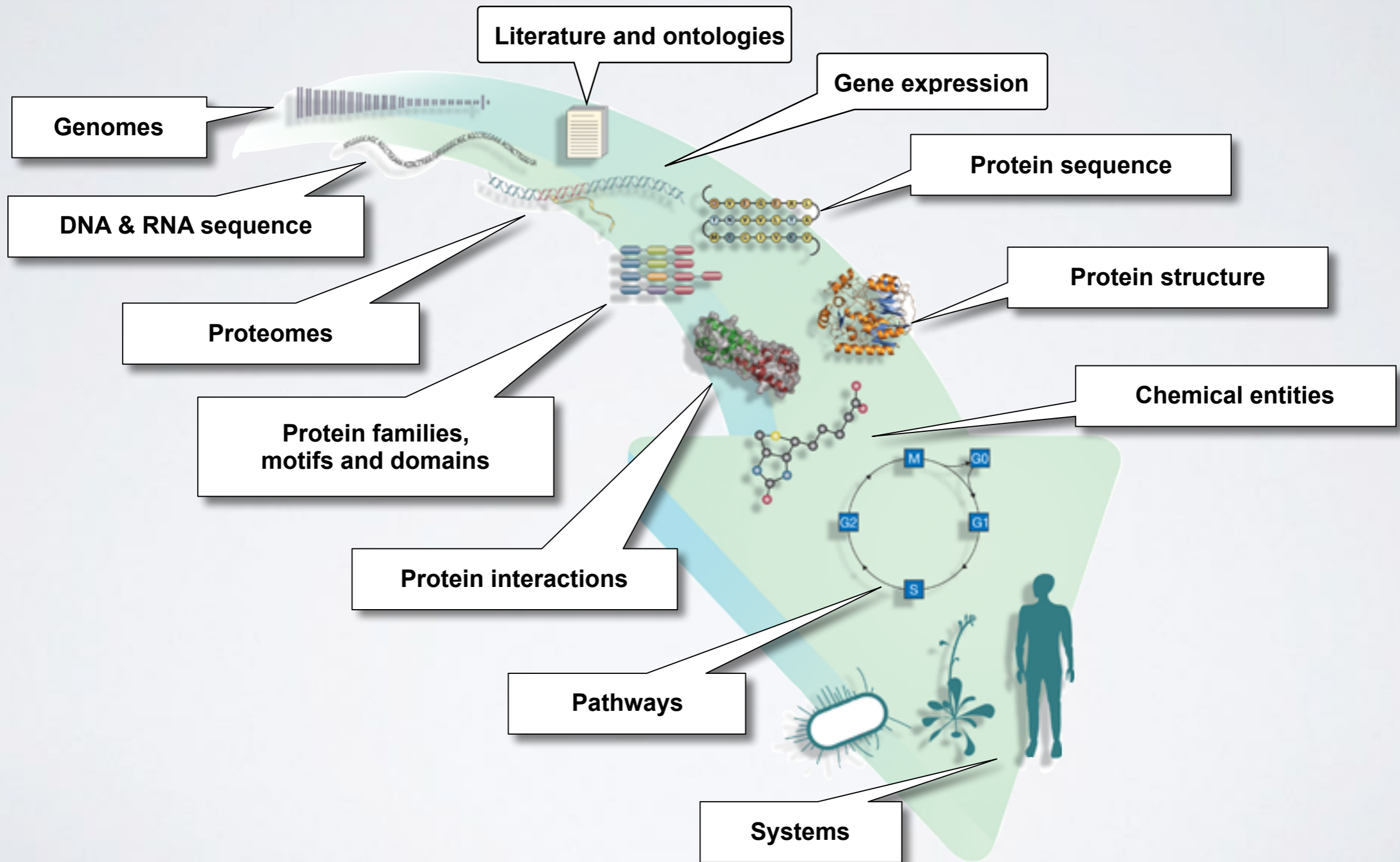
# MORE DEFINITIONS

- ▶ “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.

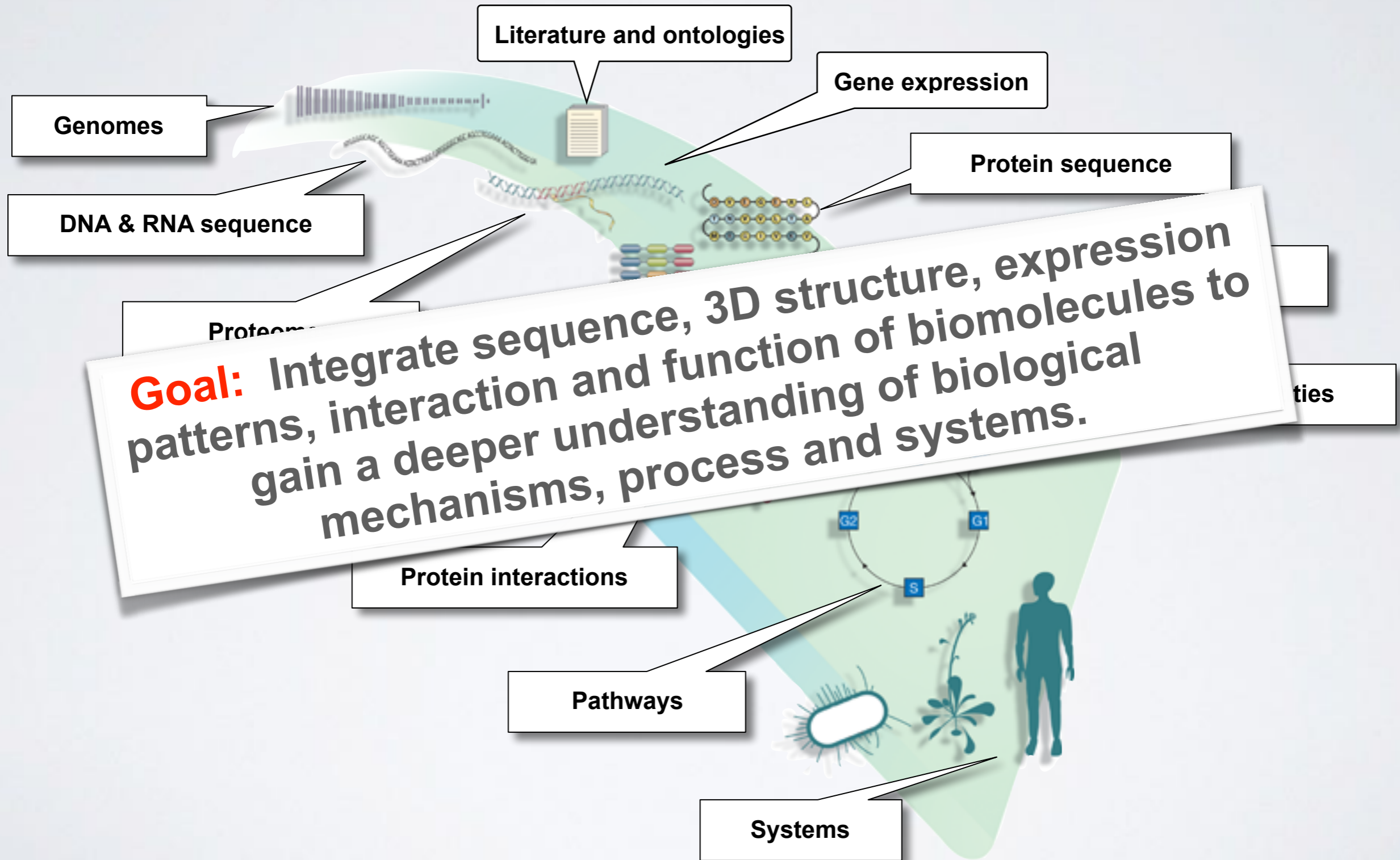
- ▶ “Bioinformatics is the research, development, or application of **computational approaches** for expanding the use of **biological, medical, behavioral or health data**, including those to **acquire, store, organize and analyze** such data.”  
National Institutes of Health (NIH) ( <http://tinyurl.com/l3gxr6b> )

**Key Point: Bioinformatics is Computer Aided Biology**

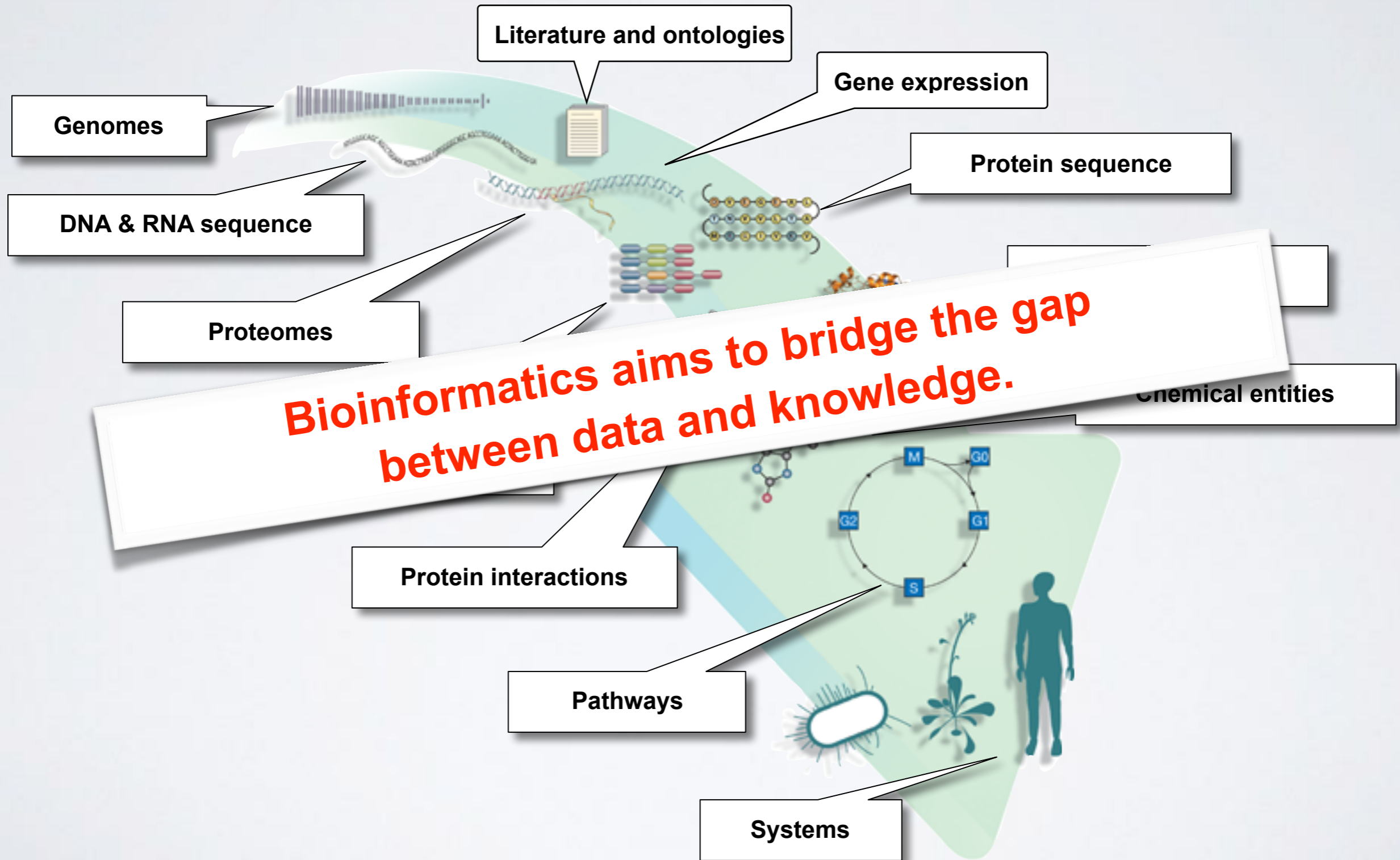
# Major types of Bioinformatics Data



# 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.*

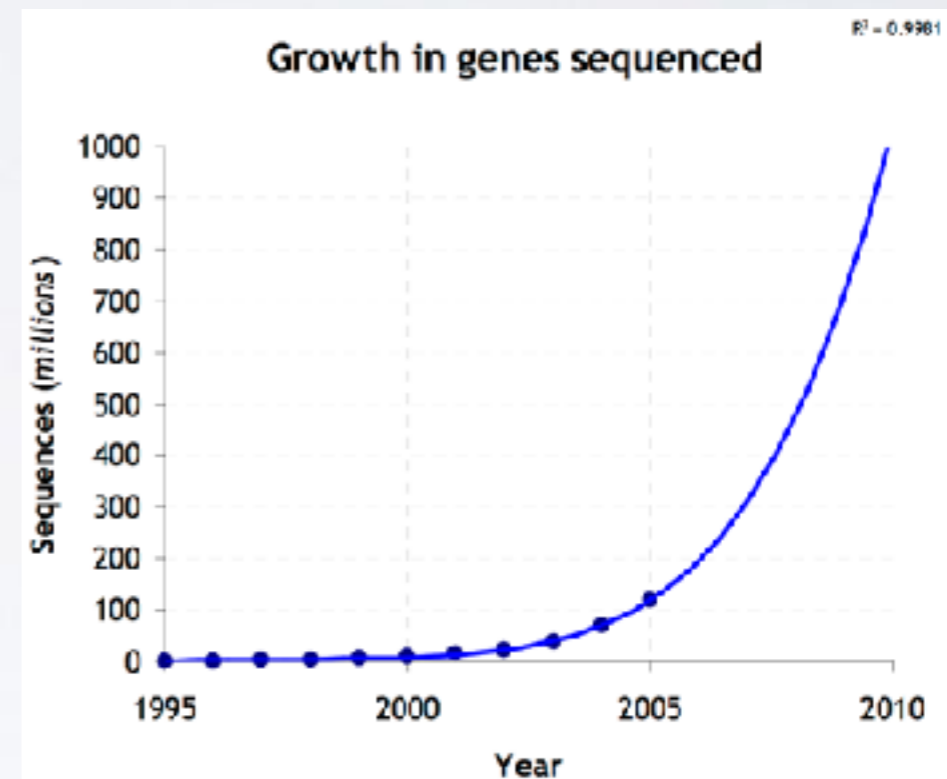
Bioinformatics is now essential for the archiving, organization and analysis of data related to all these processes.



# 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**

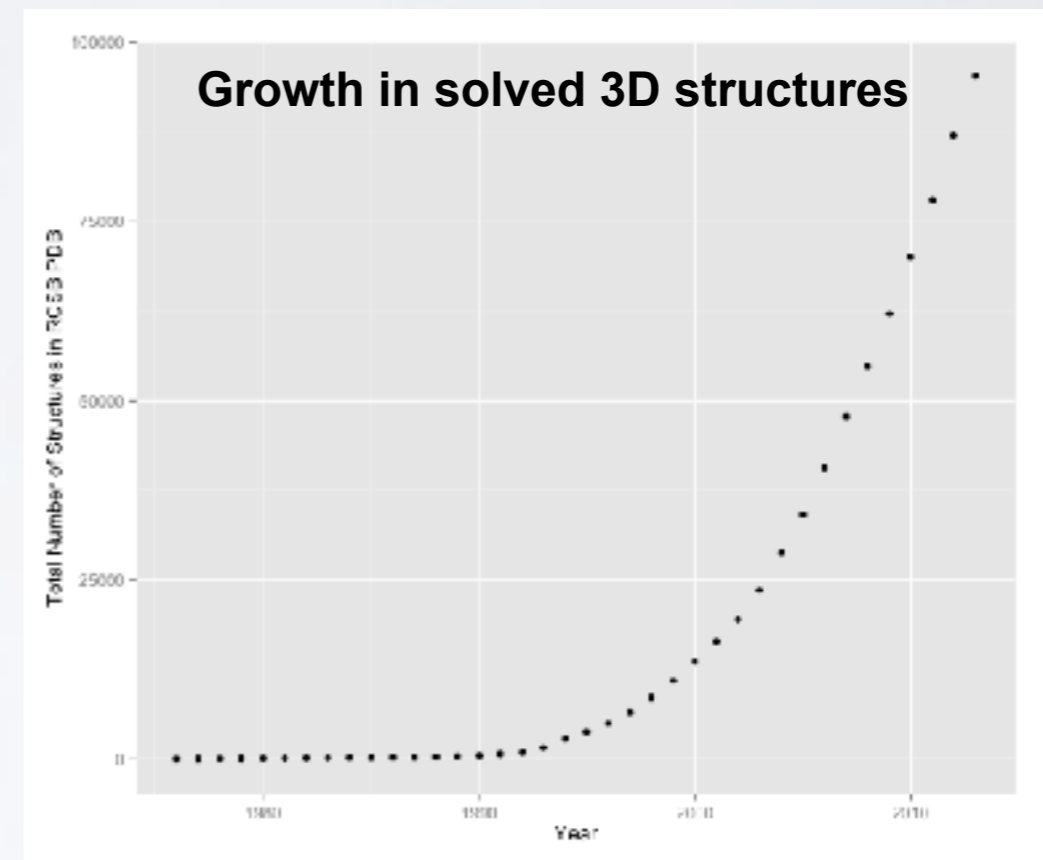


E.G. data from sequencing, structural genomics, proteomics, new high throughput assays, *etc...*

# 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
- Most are free to use
- Time consuming methods require downloading...

## Advanced tool application & development

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

# How do we *actually* do Bioinformatics?

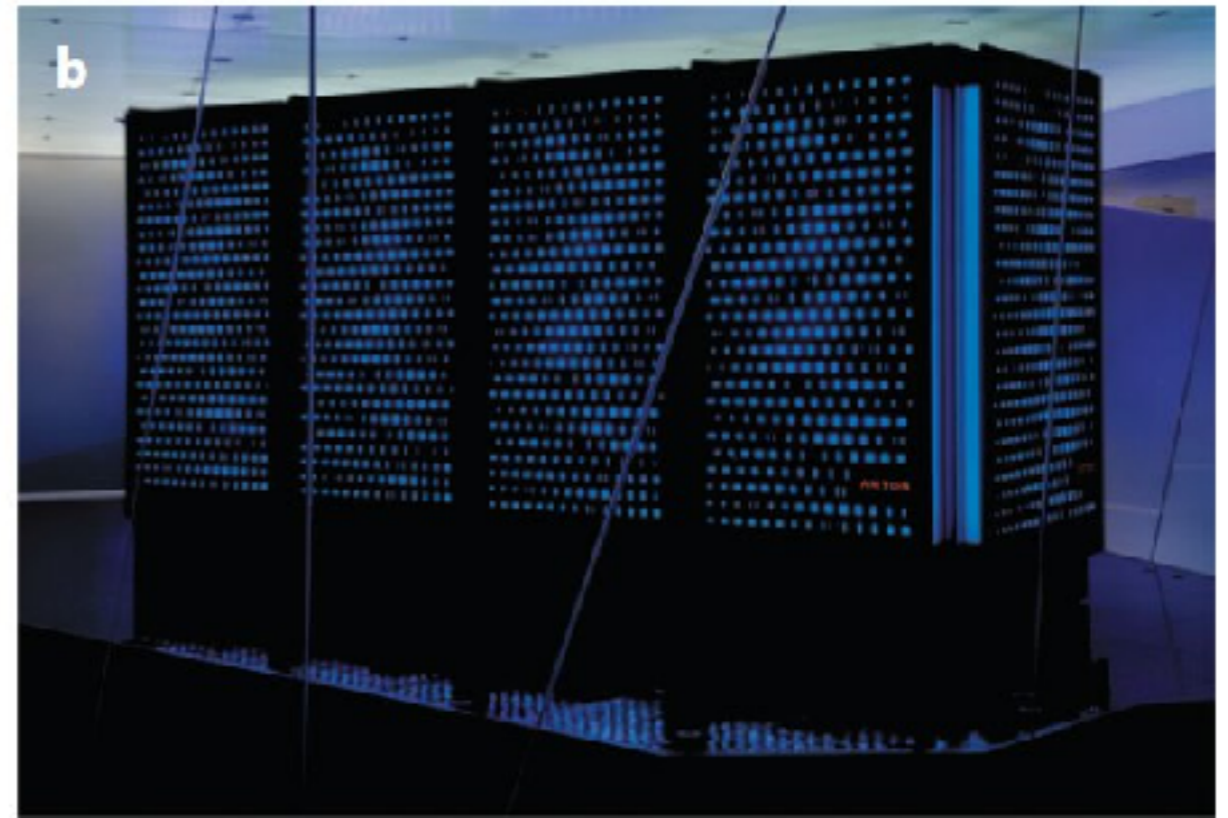
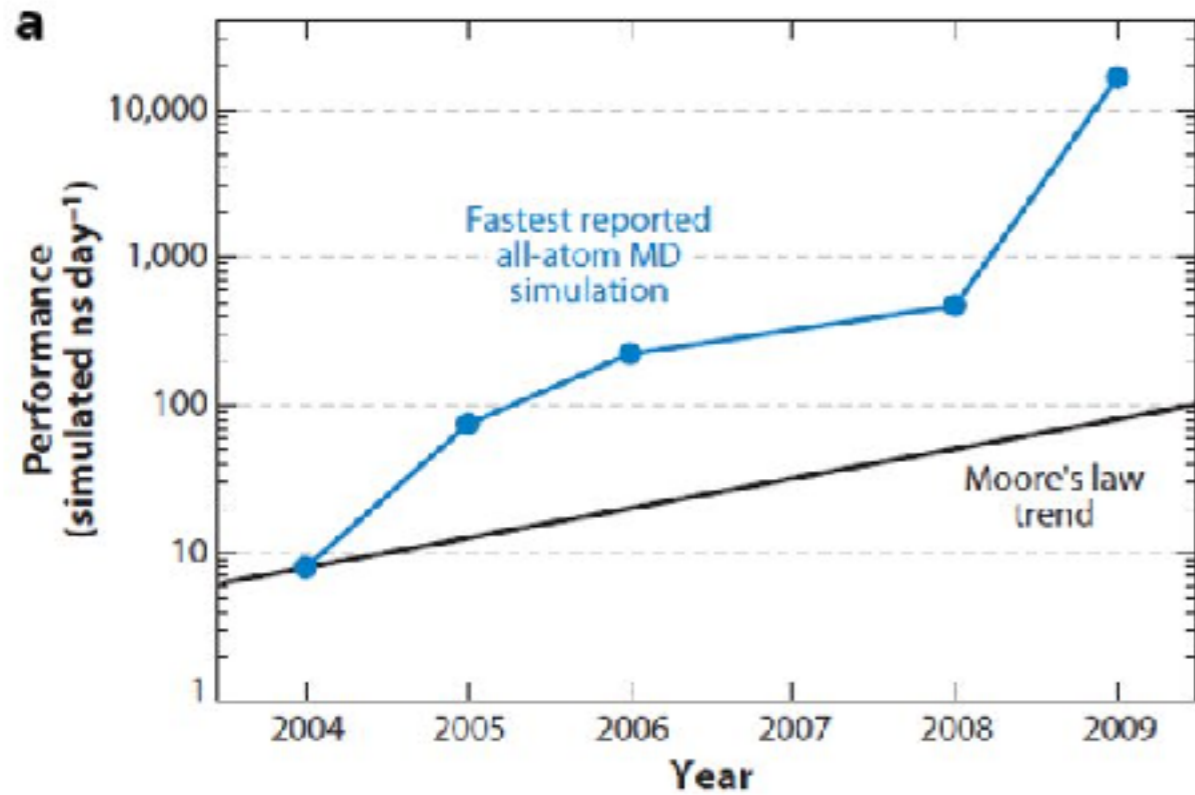
## Pre-packaged tools and databases

- Many online
- Most are free to use
- Time consuming methods require downloading...

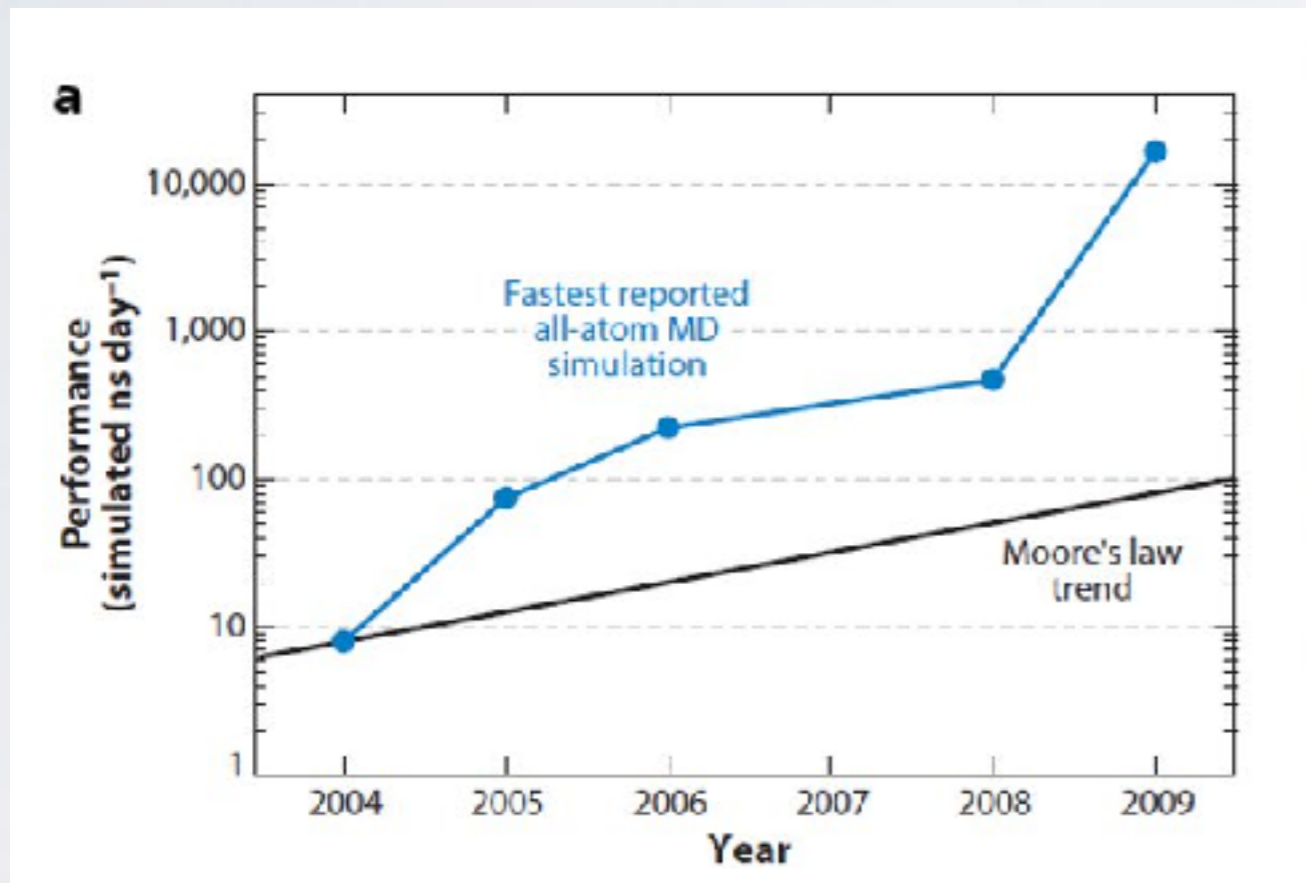
## Advanced tool application & development

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

# SIDE-NOTE: SUPERCOMPUTERS AND GPUS



# SIDE-NOTE: SUPERCOMPUTERS AND GPUS



## HOW COMPUTERS HAVE CHANGED

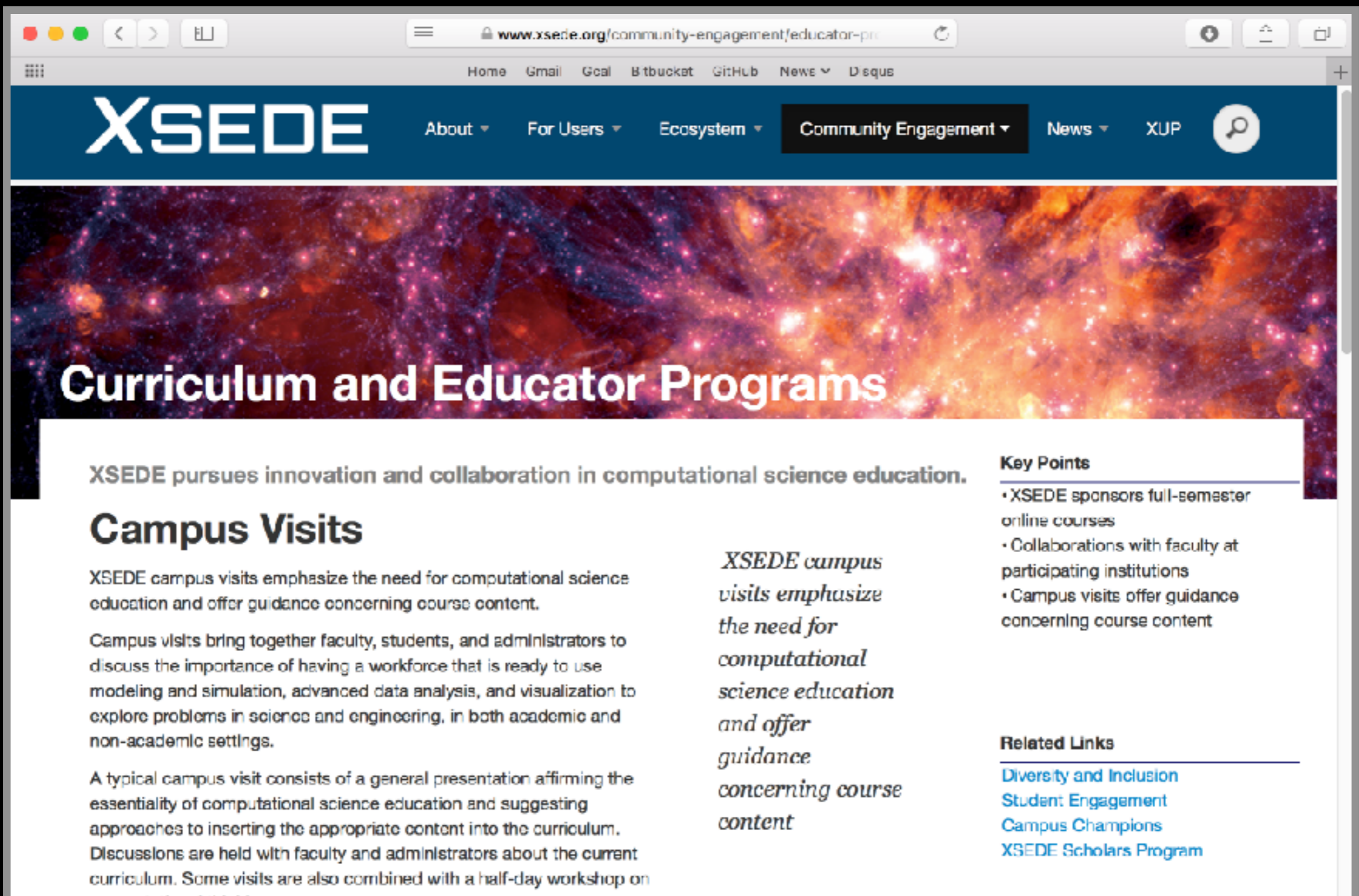
DATE	COST	SPEED	MEMORY	SIZE
1967	\$40M	0.1 MHz	1 MB	WALL
2013	\$4,000	1 GHz	10 GB	LAPTOP
CHANGE	10,000	10,000	10,000	10,000

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 shedbox





# NSF Extreme Science and Engineering Discovery Environment (XSEDE)



The image is a screenshot of a web browser displaying the XSEDE website. The browser's address bar shows the URL [www.xse.de.org/community-engagement/educator-pr](http://www.xse.de.org/community-engagement/educator-pr). The website's navigation menu includes links for Home, Gmail, Gcal, Btbucket, GitHub, News, and Disqus. The main header features the XSEDE logo and a menu with options: About, For Users, Ecosystem, Community Engagement (highlighted), News, and XUP. A search icon is also present. The main content area has a background image of a colorful nebula and is titled "Curriculum and Educator Programs". Below the title, there is a paragraph stating "XSEDE pursues innovation and collaboration in computational science education." followed by a section titled "Campus Visits". This section contains three paragraphs of text. To the right of the main text, there are two sidebars: "Key Points" with a bulleted list and "Related Links" with a list of hyperlinks.

www.xse.de.org/community-engagement/educator-pr

Home Gmail Gcal Btbucket GitHub News Disqus

**XSEDE** About For Users Ecosystem **Community Engagement** News XUP

## Curriculum and Educator Programs

XSEDE pursues innovation and collaboration in computational science education.

### Campus Visits

XSEDE campus visits emphasize the need for computational science education and offer guidance concerning course content.

Campus visits bring together faculty, students, and administrators to discuss the importance of having a workforce that is ready to use modeling and simulation, advanced data analysis, and visualization to explore problems in science and engineering, in both academic and non-academic settings.

A typical campus visit consists of a general presentation affirming the essentiality of computational science education and suggesting approaches to inserting the appropriate content into the curriculum. Discussions are held with faculty and administrators about the current curriculum. Some visits are also combined with a half-day workshop on

*XSEDE campus visits emphasize the need for computational science education and offer guidance concerning course content*

#### Key Points

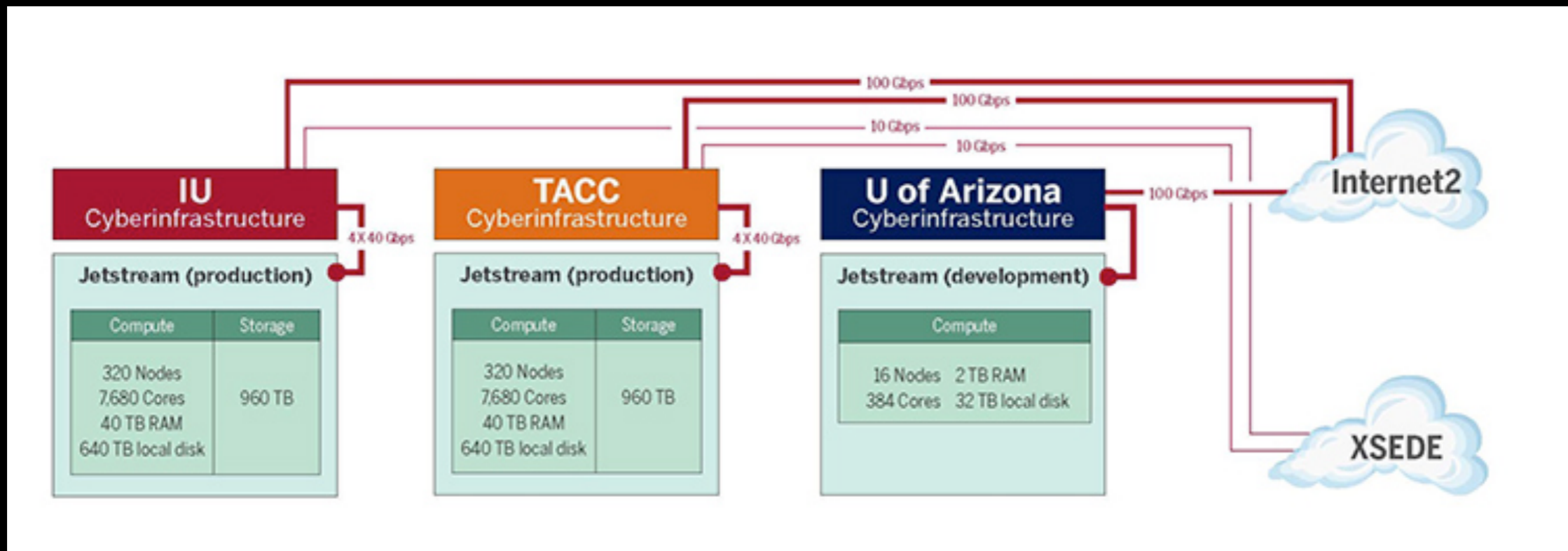
- XSEDE sponsors full-semester online courses
- Collaborations with faculty at participating institutions
- Campus visits offer guidance concerning course content

#### Related Links

- [Diversity and Inclusion](#)
- [Student Engagement](#)
- [Campus Champions](#)
- [XSEDE Scholars Program](#)

# What is *Jetstream*?

- A new cloud computing environment based at Indiana University and the Texas Advanced Computing Center (TACC) providing on-demand access to interactive computing and data analysis resources.



# Jetstream tutorials

Developed *user friendly* labs for Jetstream basics



The screenshot shows a web browser window with the URL `biohoot.github.io/bggn213_f17/jetstream/boot/`. The page title is "Starting a Jetstream Computer Instance!". The main content area contains a paragraph: "Here we describe the process of starting up and managing a [Jetstream](#) service virtual machine instance." Below this is a note in a light gray box: "Note: Jetstream is a cloud-based on-demand virtual machine system funded by the National Science Foundation. It will provide us with computers (what we call 'virtual machine instances') that look and feel just like a regular Linux workstation but with thousands of times the computing power!" The text continues: "What we're going to do here is walk through starting up an running computer (an 'instance') on the Jetstream service." and "Below we walk through the process of starting up and accessing one of these instances. To begin with, just think of it like requesting and logging-in to a brand new remote computer. We have provided screenshots of the whole process that you can click on to see a larger version. The important areas to fill in are circled in red." At the bottom, another note states: "Note Some of the details may vary - for example, if you have your own XSEDE account, you may want to log in with that - and the name of the operating system or 'image' may also vary from 'Ubuntu 16.04' depending".

The left sidebar features the UC San Diego logo and the course title "BGGN 213". Below the title is a description: "A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD". The sidebar menu includes "Overview", "Lectures" (highlighted with a red box), "Computer Setup", "Learning Goals", "Assignments & Grading", "Ethics Code", and "Screen Cast Videos".

On the right side of the page, there is a vertical red arrow pointing downwards, composed of a series of small red squares followed by a larger red triangle.

# Jetstream tutorials

Developed *user friendly* labs for Jetstream basics

The image shows two overlapping browser windows. The background window displays the UC San Diego BGGN 213 course page. The foreground window shows a 'Request to log in to the Jetstream Portal' page with instructions and a screenshot of the Jetstream application interface.

**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**

**Request to log in to the Jetstream Portal**

First, go to the Jetstream application at:  
<https://use.jetstream-cloud.org/application>.

Now click the **login** link in the upper right.

The screenshot of the Jetstream application shows a search interface with a 'Login' link circled in red in the top right corner.



# 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'

# Skepticism & Bioinformatics

Gunnar von Heijne in his old but quite readable treatise, *Sequence Analysis in Molecular Biology; Treasure Trove or Trivial Pursuit*, provides a very appropriate conclusion:

- “Think about what you’re doing; use your knowledge of the molecular system involved to guide both your interpretation of results and your direction of inquiry; use as much information as possible; and do not blindly accept everything the computer offers you”.
- Key-Point: **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

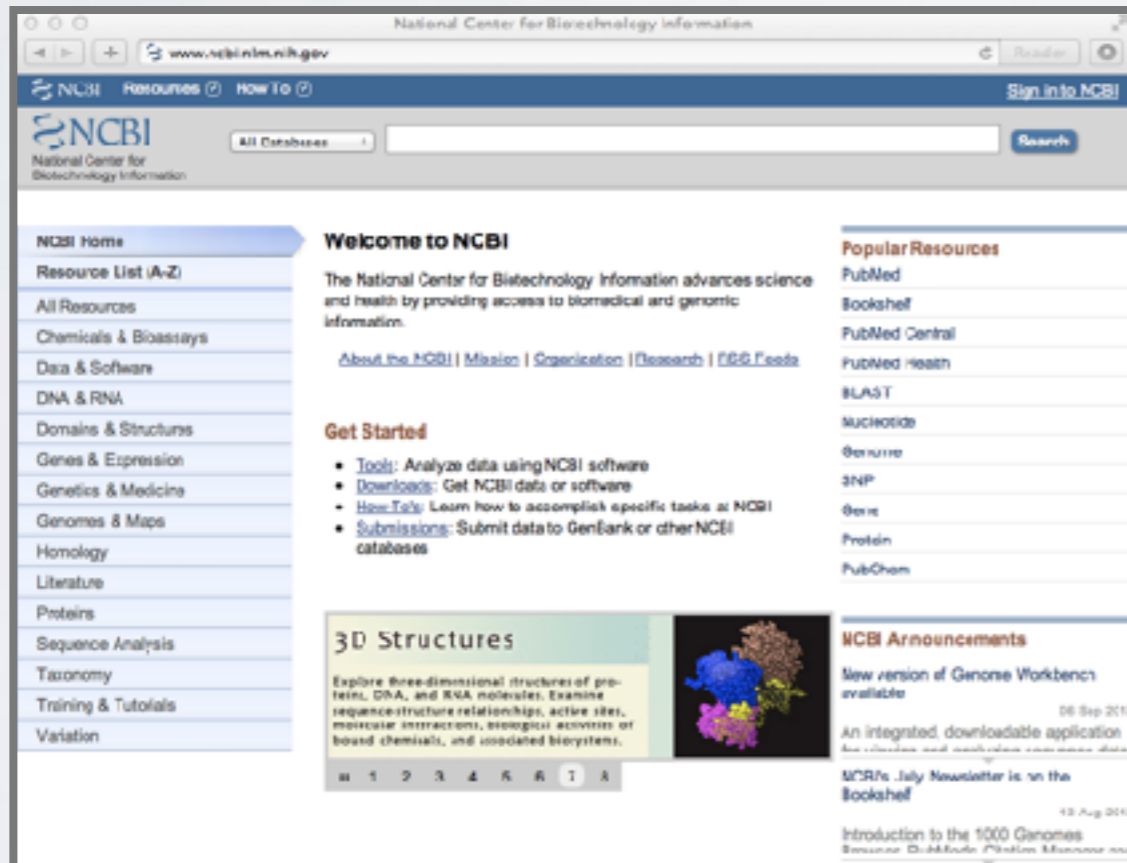
Related tools with different terminology

**STEP 3 - Set your PROGRAM**

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



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



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

# National Center for Biotechnology Information (NCBI)

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



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

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.

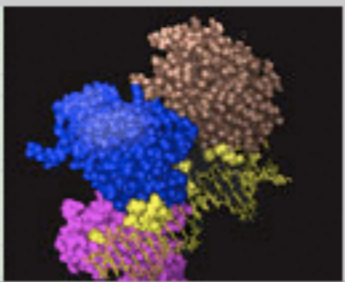
[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

### 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 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 and a 'Get Started' section with links to 'Tools', 'Downloads', 'How-To's', and 'Submissions'. A 'Popular Resources' box is overlaid on the right side of the page, listing various services: PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. Red arrows point to PubMed, BLAST, and SNP in this list. The background shows a search bar and a 'Sign in to NCBI' link.

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

**Popular Resources**

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

**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 provides access to a wide range of biological and health information.

[About the NCBI](#) | [Mission](#) | [Our Services](#)

**Get Started**

- [Tools](#): Analyze data using NCBI tools
- [Downloads](#): Get NCBI data
- [How-To's](#): Learn how to access 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.

**Resources**

Central Health

**Announcements**

New version of Genome Workbench available

06 Sep

An integrated, downloadable application

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

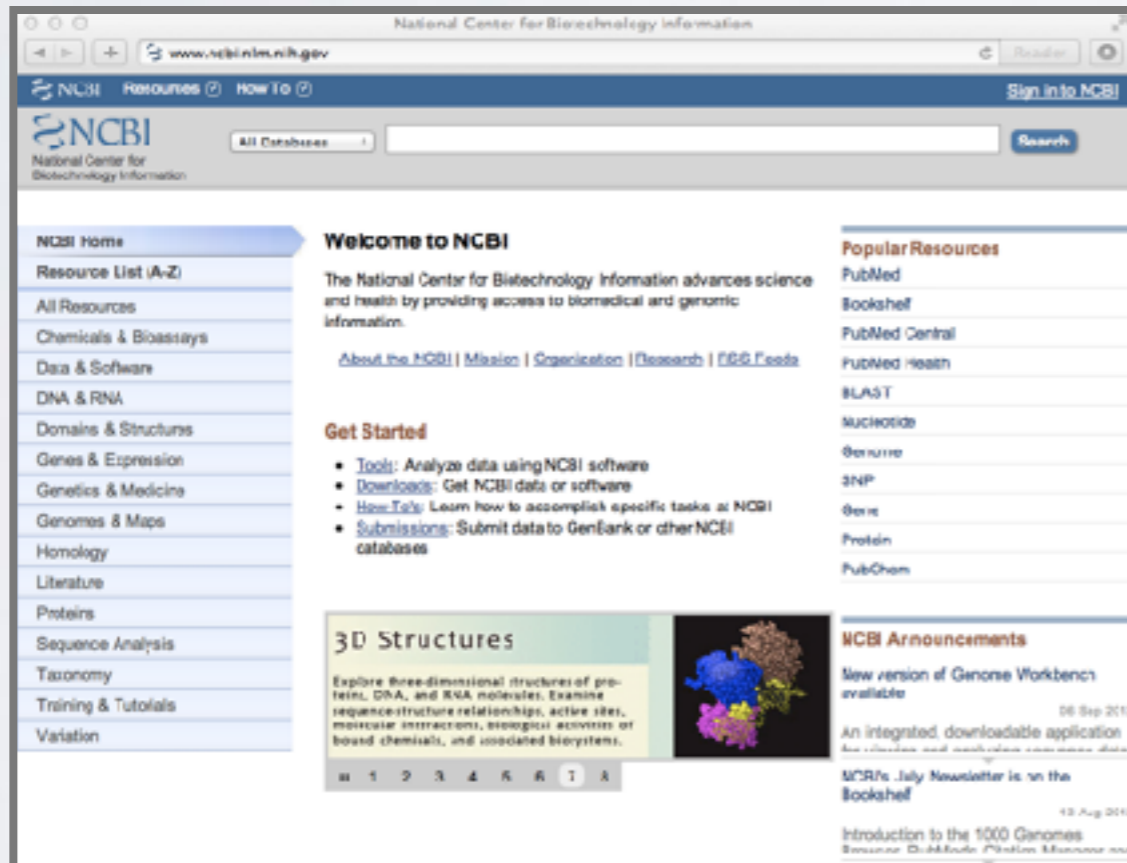
The screenshot shows the NCBI website homepage. At the top, there is a navigation bar with "NCBI", "Resources", and "How To" menus, along with a "Sign in to NCBI" link. Below this is a search bar with a dropdown menu set to "All Databases" and a "Search" button. The main content area features a "Welcome to NCBI" message, a "Resource List (A-Z)" link, and a "Popular Resources" section with a "PubMed" link. The tagline "The National Center for Biotechnology Information advances science" is also visible.

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

This screenshot shows a section of the NCBI website with a sidebar on the left containing links to "Homology", "Literature", "Proteins", "Sequence Analysis", "Taxonomy", "Training & Tutorials", and "Variation". The main content area is titled "databases" and features a "3D Structures" section with a description: "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." To the right of this text is a 3D molecular model. Further right, there are links for "Protein" and "PubChem". At the bottom right, there is an "NCBI Announcements" section with the headline "New version of Genome Workbench available" and a date of "06 Sep".

# Key Online Bioinformatics Resources: NCBI & EBI

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



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



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

# European Bioinformatics Institute (EBI)

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





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

The screenshot shows the EMBL-EBI website homepage. At the top, the browser address bar displays 'www.ebi.ac.uk'. The main header features the EMBL-EBI logo and the text 'The European Bioinformatics Institute' and 'Part of the European Molecular Biology Laboratory'. A navigation menu includes 'Services', 'Research', 'Training', and 'About us'. 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.' Below this is a search bar with the prompt 'Find a gene, protein or chemical:' and a 'Search' button. A row of six colored buttons is visible: 'Services' (teal, highlighted with a red border), 'Research' (green), 'Training' (yellow), 'Industry' (blue), 'European Coordination' (orange), and 'EMBL ALUMNI' (white with green logo). 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). An '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 'News from EMBL-EBI' section with several 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 links for 'Services', 'Research', 'Training', and 'About us'. Below the header, there are tabs 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. A grid of service categories is presented, including DNA & RNA, Gene expression, Proteins, Structures, Systems, Chemical biology, Ontologies, Literature, and Cross domain. A 'Popular' section lists tools like Ensembl, UniProt, PDBc, ArrayExpress, ChEMBL, BLAST, Europe PMC, Reactome, Train online, and Support. A 'Service news' section features a butterfly and molecular structure image. At the bottom, a 'Training' section is partially visible.

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
- PDBc
- 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 heading is "Services" with sub-navigation for "Overview", "A to Z", "Data submission", and "Support". The "Bioinformatics services" section describes the availability of molecular databases. A grid of service categories includes "Proteins" (sequences, families & motifs), "Gene expression", "DNA & RNA", "Structures", "Systems", "Chemical biology", "Ontologies", "Literature", and "Cross domain". A "Popular" sidebar lists Ensembl, UniProt, PDBe, ArrayExpress, and ChEMBL. 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
- PDBe
- ArrayExpress
- ChEMBL

#### Training








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

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

The screenshot shows the 'Proteins' section of the EBI website. It features a 'Popular services' table with icons and descriptions for UniProt, InterPro, PRIDE, Pfam, Clustal Omega, HMMER, and InterProScan 5. To the right, a 'Quick links' box contains three links: 'Popular services in this category', 'All services in this category', and 'Project websites in this category'.

## Proteins

### Popular services

	<b>UniProt: The Universal Protein Resource</b> The gold-standard, comprehensive resource for protein sequence and functional annotation data.
	<b>InterPro</b> A database for the classification of proteins into families, domains and conserved sites.
	<b>PRIDE: The Proteomics Identifications Database</b> An archive of protein expression data determined by mass spectrometry.
	<b>Pfam</b> A database of hidden Markov models and alignments to describe conserved protein families and domains.
	<b>Clustal Omega</b> Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.
	<b>HMMER - protein homology search</b> 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.
	<b>InterProScan 5</b> 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

EMBL-EBI

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:

Search

Examples: [blast](#), [keratin](#), [bff1](#)...

Services Research **Training** Industry European Coordination

EMBL ALUMNI

Popular

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

Visit **EMBL.org**

EMBL 40th anniversary logo

Upcoming events

Plant and Animal Genome conference (PAG XXIV)

Sunday 10 - Tuesday 12 January 2016

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

The screenshot shows a web browser window with the URL [www.ebi.ac.uk/training/online/course/using-sequence-similarity-searching-tools-ebi-ebi](http://www.ebi.ac.uk/training/online/course/using-sequence-similarity-searching-tools-ebi-ebi). The page features the EMBL-EBI logo and navigation menus for Services, Research, Training, and About us. A prominent yellow banner reads "Train online". Below this, a breadcrumb trail indicates the current page: training > online > course-list > using-sequence-similarity-searching-tools-ebi-ebi. The main content area is titled "Using sequence similarity searching tools at EMBL-EBI: webinar" and contains a video player. The video player shows a slide with the following text: "Using sequence similarity search tools at EMBL-EBI. Finding homologous sequences with BLAST, FASTA, PSI-Search etc." and a portrait of Andrew Cowley, with contact information: andrew.cowley@ebi.ac.uk and support@ebi.ac.uk. The video player controls show a progress bar at 0:05 / 37:42. To the left of the video, there is a "Course content" sidebar with a highlighted item "Using sequence similarity searching tools at EMBL-EBI: webinar" and a "Contributors" section. Below the sidebar is a "Print Course" link. To the right of the video, there are two sections: "Popular" with links for "Train online", "Find us", and "Funding"; and "Find us at..." with links for "Open days and career days", "Conference exhibitions", "EMBL courses and events", "Genome campus events", and "Science for schools".

EMBL-EBI

Services Research Training About us

# Train online

Training Train online Home Course list Glossary Support & Feedback Log in / Register

training > online > course-list > using-sequence-similarity-searching-tools-ebi-ebi

## Course content

Using sequence similarity searching tools at EMBL-EBI: webinar

Contributors

[Print Course](#)

### Using sequence similarity searching tools at EMBL-EBI: webinar

Using sequence similarity search tools at EMBL-EBI  
Finding homologous sequences with BLAST, FASTA, PSI-Search etc.

Andrew Cowley  
andrew.cowley@ebi.ac.uk  
support@ebi.ac.uk

0:05 / 37:42

Popular

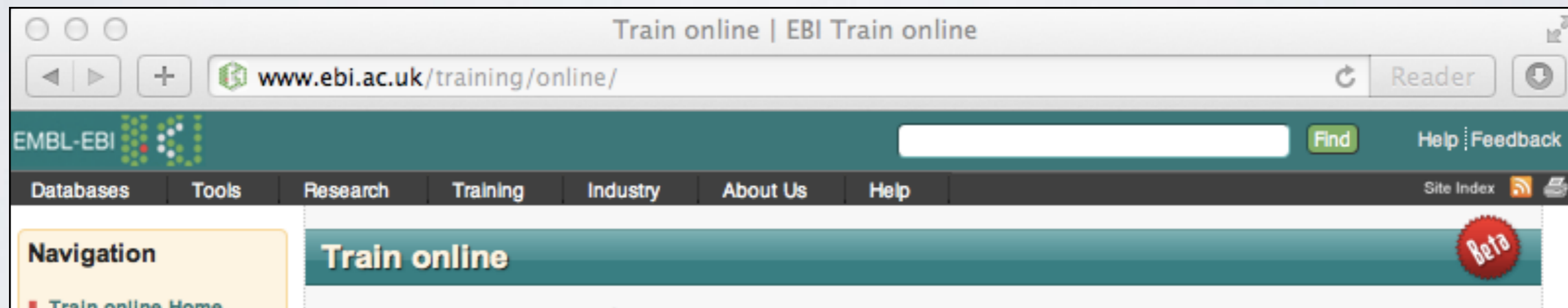
- [Train online](#)
- [Find us](#)
- [Funding](#)

Find us at...

- [Open days and career days](#)
- [Conference exhibitions](#)
- [EMBL courses and events](#)
- [Genome campus events](#)
- [Science for schools](#)

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, DPlInteract, 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-U's, 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, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc ..... !!!!

# Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCCP, Beanref, Biolmage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVM, TKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, AP, ChickGBASE, Colibri, COPE, CottonDB, bEST, dbSTS, DDBJ, DGP, DictyDb, CDC, ECGC, EC02DBASE, OTHER, FlyBase, Link, G, HAEMB, HotMolecBase, H, K, MZRGbase, IMGT, Kabat, KDNA, MHC, Medline, Mendel, MEROPS, MGDB, MGI, Myc, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PD, P, 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**

**[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.

# Today's Menu

<b>Course Logistics</b>	Website, screencasts, survey, ethics, assessment and grading.
<b>Learning Objectives</b>	What you need to learn to succeed in this course.
<b>Course Structure</b>	Major lecture topics and specific learning goals.
<b>Introduction to Bioinformatics</b>	Introducing the <i>what</i> , <i>why</i> and <i>how</i> of bioinformatics?
<b>Bioinformatics Database</b>	<b>Hands-on</b> exploration of several major databases and their associated tools.

# Your Turn!

[https://bioboot.github.io/bgggn213\\_S18/lectures/#1](https://bioboot.github.io/bgggn213_S18/lectures/#1)

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 [\[R\]](#).

Overview

**Lectures**

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

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](#) [\[R\]](#).
- Setup your [laptop computer](#) for this course.
- The goals of the hands-on session is to introduce a range of core bioinformatics databases and associated online services whilst actively investigating the molecular basis of several common human disease.

Material:

- Lecture Slides: [Large PDF](#) [\[R\]](#), [Small PDF](#) [\[R\]](#),
- Lab: [Hands-on section worksheet](#) [\[R\]](#)
- Feedback: [Muddy Point Assessment](#) [\[R\]](#),
- Feedback: [Results](#) [\[R\]](#).
- Handout: [Class Syllabus](#) [\[R\]](#)
- Computer [Setup Instructions](#).

Homework:

- [Questions](#) [\[R\]](#),
- Readings:
  - PDF1: [What is bioinformatics? An introduction and overview](#) [\[R\]](#),
  - PDF2: [Advancements and Challenges in Computational Biology](#) [\[R\]](#),

## BGGN-213: FOUNDATIONS OF BIOINFORMATICS (Lecture 1)

### Bioinformatics Databases and Key Online Resources

[https://bioboot.github.io/bggn213\\_S18/lectures/#1](https://bioboot.github.io/bggn213_S18/lectures/#1)

Dr. Barry Grant

**Overview:** The purpose of this lab session is to introduce a range of bioinformatics databases and associated services available on the Web whilst investigating the molecular basis of a common human disease.

Sections 1 and 2 deal with querying and searching GenBank, GENE and OMIM databases at NCBI. Sections 3 and 4 provide exposure to EBI resources for comparing proteins and visualizing protein structures. Finally, section 5 provides an opportunity to explore these and other databases further with additional examples.

**Side-note:** The Web is a dynamic environment, where information is constantly added and removed. Servers "go down", links change without warning, etc. This can lead to "broken" links and results not being returned from services. Don't give up - give it a second go and try a search engine using terms related to the page you are trying to access.

### Section 1

The following transcript was found to be abundant in a human patient's blood sample.

>example1

```
ATGGTGCATCTGACTCCTGTGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG
TTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGG
GGATCTGTCCACTCCTGATGCAGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCA
TCACCTTGGCAAAGAATTCACCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCCCAAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTT
```

The only information you are given is the above sequence so you must begin your investigation with a sequence search - for this example we will use NCBI's **BLAST** service at: <http://blast.ncbi.nlm.nih.gov/>

*Note that there are several different "basic BLAST" programs available at NCBI (including nucleotide BLAST, protein BLAST, and BLASTx).*

# YOUR TURN!

- There are five major hands-on sections including:
  1. BLAST, GenBank and OMIM @ **NCBI** [~35 mins]
  2. GENE database @ **NCBI** [~15 mins]  
— BREAK —
  3. UniProt & Muscle @ **EBI** [~25 mins]
  4. PFAM, PDB & NGL [~30 mins]  
— BREAK —
  5. Extension exercises [~30 mins]
- ▶ Please do answer the last review question (**Q19**).
- ▶ We encourage discussion and exploration!



# YOUR TURN!

- There are five major hands-on sections including:

End times:

1. BLAST, GenBank and OMIM @ **NCBI**

[10:45 am]

2. GENE database @ **NCBI**

[11:00 am]

— BREAK —

— 11:10 am —

3. UniProt & Muscle @ **EBI**

[11:35 am]

4. PFAM, PDB & NGL

[12:05 pm]

— BREAK —

— 12:15 am —

5. Extension exercises

[12:45 pm]

- ▶ Please do answer the last review question (**Q19**).
- ▶ We encourage discussion and exploration!

# 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.

# HOMework

[https://bioboot.github.io/bggn213\\_S18/lectures/#1](https://bioboot.github.io/bggn213_S18/lectures/#1)

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

THANK YOU