

BIMM 143

Introduction to Bioinformatics

Barry Grant
UC San Diego

<http://thegrantlab.org/bimm143>

HELLO
my name is

BARRY

bjgrant@ucsd.edu

Office Hours:
[SignUp](#)

Location:
TATA, #2501

HELLO
HER — my name is

ALENA

amartsul@ucsd.edu

HELLO
HER — my name is

KELLY

kflander@ucsd.edu

Introduce Yourself!

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

Today's Menu

Course Logistics

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

Learning Objectives

What you need to learn to succeed in this course.

Course Structure

Major lecture topics and specific learning goals.

Introduction to Bioinformatics

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

Bioinformatics Database

Hands-on exploration of several major databases and their associated tools.

<http://thegrantlab.org/bimm143/>

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_F18/ in the address bar. The main content area displays the following information:

UC San Diego

BIMM 143

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

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

Instructional Assistant
Chao Shi (Email: bioshichao@gmail.com)

Course Syllabus
[Fall 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 upper division 4-unit course is designed for biology majors and provides an introduction to the principles and practical approaches of bioinformatics as applied to genes and proteins.

Navigation links on the left side of the page include:

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

Social media icons at the bottom left:

-
-
-
-

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

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_W18/goals/ in the address bar. The page content is as follows:

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 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.

Specific Learning Goals

The left sidebar of the website includes the following navigation links:

- UCSanDiego
- BIMM 143
- 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

Below the sidebar are social media sharing icons for Twitter, LinkedIn, Email, and RSS feed.

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/bimm143_W18/goals/. The page content is as follows:

UCSanDiego

BIMM 143

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

Navigation menu:

- Overview
- Lectures
- Computer Setup
- Learning Goals** (highlighted with a red border)
- Assignments & Grading
- Ethics Code

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 as 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 between global and local alignment along with their major application areas.	4, 5
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	5, 10

Course Structure

Derived from specific learning goals

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_S18/lectures/. The page title is "Lectures".
The left sidebar includes links for Overview, Lectures (which is highlighted with a red box), Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code.
The main content area displays a table of lectures for Spring 2018:

#	Date	Topics for Spring 2018
1	Tu, 04/03	Welcome to 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, Hands on with major Bioinformatics databases and key online NCBI and EBI resources
2	Th, 04/05	Sequence alignment fundamentals, algorithms and applications 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
		Advanced sequence alignment and database searching

Course Structure

Derived from specific learning goals

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_S18/lectures/. The page is titled "Lectures". A table lists the topics for Spring 2018, with the first topic, "Welcome to Bioinformatics", highlighted by a red box.

#	Date	Topics for Spring 2018
1	Tu, 04/03	Welcome to 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, Hands on with major Bioinformatics databases and key online NCBI and EBI resources
2	Th, 04/05	Sequence alignment fundamentals, algorithms and applications 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
		Advanced sequence alignment and database searching

Class Details

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

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_W18/lectures/#1. The page has a dark blue background with a glowing blue circular graphic on the left.

UCSanDiego

BIMM 143

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

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

Material:

- Pre class screen casts (also see below):
 - SC1: [Welcome to BIMM-143](#) [\[link\]](#),
 - SC2: [What is Bioinformatics?](#) [\[link\]](#) and
 - SC3: [How do we do Bioinformatics?](#) [\[link\]](#).
- Lecture Slides: Large PDF, Small PDF
- [Handout: Class Syllabus](#) [\[link\]](#)

Homework

Goals, Class material, Screencasts & Homework

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_W18/lectures/#1. The page content includes:

- UC San Diego BIMM 143**: A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.
- 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)

The video player displays a screen cast of a man speaking in front of a colorful molecular model. The video controls show it's at 2:05 / 4:05. The URL <http://thegrantlab.org/baan213> is visible at the bottom of the video frame.

1 Welcome to BIMM-143: Course introduction and logistics.

Homework

Goals, Class material, Screencasts & Homework

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_W18/lectures/#1. The page content includes:

- Homework:**
 - [Questions](#) (highlighted with a red box)
 - Readings:
 - PDF1: [What is bioinformatics? An introduction and overview](#),
 - PDF2: [Advancements and Challenges in Computational Biology](#),
 - Other: [For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights](#) New York Times, 2014.
- Screen Casts:**
 - Welcome to “Foundations of Bioinformatics” (BGGN-213)**: A video player showing a man speaking in front of a background of colorful 3D molecular models. The video is at 2:05 / 4:05. The URL <http://thegrantlab.org/baan213> is displayed below the video player.

On the left side of the browser window, there is a sidebar for the course BIMM 143, UC San Diego, featuring:

- BIMM 143**
- 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**

Homework

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

BIMM143 Lecture 1 Homework (W19)

Please answer the following questions including your main @ucsd.edu email address and UCSD PID number so you can receive credit for your responses.

* Required

Email address *

Your email

UCSD PID number (exam number)

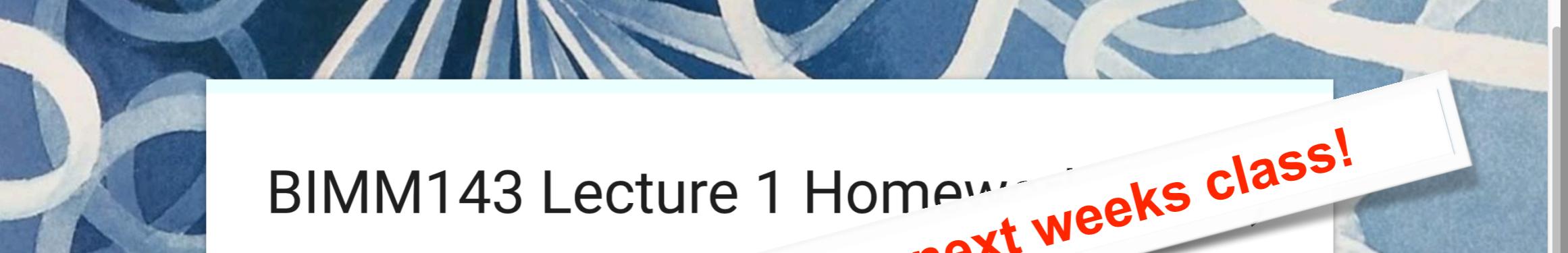
Your answer

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

1 point

Homework

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



BIMM143 Lecture 1 Homework

Please answer the following questions and include your UCSD PID number so you can receive credit.

Homework is due before the next weeks class!

Email address *

Your email

UCSD PID number (exam number)

Your answer

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

1 point

Homework

Goals, Class material, Screencasts & Homework

The screenshot shows a DataCamp exercise interface. The top navigation bar includes the DataCamp logo, course outline, and user notifications (5+). The main area has tabs for 'Exercise' (selected), 'R Documentation', and 'script.R'. The 'script.R' tab contains the following code:

```
1 # Transform the normalized counts
2 vsd_smoc2 <- vst(dds_smoc2, blind = TRUE)
3
4 # Plot the PCA of PC1 and PC2
5 ---(_____, intgroup=____)
```

The 'R Console' tab at the bottom shows the following session history:

```
> ?plotPCA
> plotPCA(vsd_smoc2)
Error: object 'vsd_smoc2' not found
> vsd_smoc2 <- vst(dds_smoc2, blind = TRUE)
+
> plotPCA(vsd_smoc2)
>
```

On the left, the 'Instructions' section says '1/2 50 XP' and lists two tasks:

- Run the code to transform the normalized counts.
- Perform PCA by plotting PC1 vs PC2 using the DESeq2 `plotPCA()` function on the DESeq2 transformed counts object, `vsd_smoc2` and specify the `intgroup` argument as the factor to color the plot.

A 'Take Hint (-15 XP)' button is also present.

Homework (35% of course grade)

Goals, Class material, Screencasts & Homework

The screenshot shows a DataCamp RStudio interface. On the left, there's an 'Exercise' panel titled 'PCA analysis'. It contains instructions: 'To continue with the quality assessment of our samples, in the first part of this exercise, we will perform PCA to look how our samples cluster and whether our condition of interest corresponds with the principal components explaining the most variation in the data. In the second part, we will answer questions about the PCA plot.' Below this, there's a code editor with a partially written R script:

```
1 # Transform the normalized counts
2 vsd_smoc2 <- vst(dds_smoc2, blind = TRUE)
3
4 # Plot the PCA of PC1 and PC2
5 ----, intgroup-
```

On the right, there's an 'R Console' tab showing the following R session:

```
> ?plotPCA
> plotPCA(vsd_smoc2)
Error: object 'vsd_smoc2' not found
> vsd_smoc2 <- vst(dds_smoc2, blind = TRUE)
+
> plotPCA(vsd_smoc2)
>
```

A large red diagonal banner across the interface reads 'Homework is due before the next weeks class!'

Projects

Week long **mini-projects** (x2),
and 1 five week main project

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_W19/lectures/#9. The page content is as follows:

UCSanDiego

BIMM 143

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

9: Unsupervised Learning Mini-Project

Topics: Longer hands-on session with unsupervised learning analysis of cancer cells, Practical considerations and best practices for the analysis and visualization of high dimensional datasets.

Goals:

- Be able to import data and prepare for unsupervised learning analysis.
- Be able to apply and test combinations of PCA, k-means and hierarchical clustering to high dimensional datasets and critically review results.

Material:

- Lecture Slides: [Large PDF](#), [Small PDF](#),
- Lab: [Hands-on section worksheet for PCA](#)
- Data file: [WisconsinCancer.csv](#), [new_samples.csv](#).
- Bio3D PCA App: <http://bio3d.ucsd.edu/pca-app/>
- Feedback: [Muddy point assessment](#)
- Bonus: [Kevin's StackExchange Link on PCA](#)

Projects (20% of course grade)

Week long mini-projects (x2),
and 1 five week **main project**

The screenshot shows a web browser window. The left side displays the UC San Diego BIMM 143 homepage with the title 'BIMM 143' and a description of the course as a hands-on introduction to computer-based analysis of genomic and biomolecular data. The right side shows a lecture page titled '10: (Project:) Find a Gene Assignment Part 1'. This page contains instructions for the assignment, including links to a project description and example report, and details about due dates. Below the main content, there is a section titled 'Bonus: Hands-on with Git' with a brief description of the topic.

UCSanDiego

BIMM 143

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

Twitter GitHub Email RSS

bioboot.github.io/bimm143_W19/lectures/#9

Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News +

10: (Project:) Find a Gene Assignment Part 1

The [find-a-gene project](#) is a required assignment for BIMM-143. The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered to date in class.

You may wish to consult the scoring rubric at the end of the above linked project description and the [example report](#) for format and content guidance.

Your responses to questions Q1-Q4 are due at the beginning of class **Thursday Nov 15th** (11/15/18).

The complete assignment, including responses to all questions, is due at the beginning of class **Thursday Dec 4th** (12/04/18).

Late responses will not be accepted under any circumstances.

Bonus: Hands-on with Git

Today's lecture and hands-on sessions introduce Git, currently the most popular version control system. We will learn how to perform common operations with Git and RStudio. We will also cover the popular social code-hosting platforms GitHub and BitBucket.

Final Exam

Open-book, open-notes 150-minute test
(45% of course grade)

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_W19/lectures/#20. The page content is as follows:

UC San Diego
BIMM 143
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

20: Final Exam

This open-book, open-notes 150-minute test consists of 35 questions. The number of points for each question is indicated in green font at the beginning of each question. There are 80 total points on offer.

Please remember to:

- Read all questions carefully before starting.
- Put your name, UCSD email and PID number on your test.
- Write all your answers on the space provided in the exam paper.
- Remember that concise answers are preferable to wordy ones.
- Clearly state any simplifying assumptions you make in solving a problem.
- No copies of this exam are to be removed from the class-room.
- No talking or communication (electronic to otherwise) with your fellow students once the exam has begun.
- **Good luck!**

At the bottom of the page are social media sharing icons for Twitter, LinkedIn, Email, and RSS.

Bonus:

Bioinformatics & Genomics in industry

The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_W19/lectures/#21 in the address bar. The browser interface includes standard controls like back, forward, and search. The main content area is a course page for BIMM 143 at UC San Diego. The page features a large blue header with the UC San Diego logo and the course name "BIMM 143". Below the header, there is a detailed description of the course: "A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD". A sidebar on the left lists navigation links: Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. At the bottom, there are social media icons for Twitter, GitHub, Email, and RSS. The central content area is titled "21: Bonus: Bioinformatics & Genomics in industry" and contains a text block about a guest lecture on March 15th.

21: Bonus: Bioinformatics & Genomics in industry

Friday March 15th at 1pm come and enjoy a set of short open ended guest lectures from leading genomic scientists at Illumina Inc., Synthetic Genomics Inc., Samumed and the La Jolla Institute for Allergy and Immunology. Come prepared for networking and to have your questions about industry careers in Bioinformatics and Genomics answered.

© 2019 Barry J. Grant. All rights reserved. A [UCSD](#) [Division of Biological Sciences](#) [Course](#)

Bonus:

Online portfolio of **your** bioinformatics work!

The screenshot shows a web browser window with the URL jasonpbennett.github.io/bimm143/. The page content is as follows:

Introduction to Bioinformatics Class S18

A repository to store and display my work completed during the Spring 2018 quarter in BIMM-143 at UCSD.

[View the Project on GitHub](#)
jasonPBennett/bimm143

Bioinformatics Class BIMM-143

This is my repository for my Bioinformatics class from UC San Diego in S18.

Index of Material

Introductory Material: Working With R

- [Class 5 - Basic Data Exploration and Visualization in R](#)
- [Class 6 - Creating R Functions](#)
- [Class 7 - R Packages, working with CRAN, and working with Bioconductor](#)

Using R and Other Tools for Bioinformatics Analysis

- [Class 8 - An Introduction to Machine Learning \(Hierarchical Clustering\)](#)
- [Class 9 - Analyzing High Dimensional Datasets and Unsupervised Learning](#)
- [Class 11 - Structural Bioinformatics: Analyzing Protein Structure and Function](#)
- [Class 12 - Drug Discovery: Techniques and Analysis](#)
- [Class 13 - Genome Informatics and High Throughput Sequencing \(NGS, RNA-Seq, and FastQC\)](#)
- [Class 14 - Transcriptomics and RNA-Seq Analysis](#)
- [Class 15 - Genome Annotation and Using Functional Databases \(KEGG and GO - Gene Ontology\)](#)
- [Class 16 - Transposons: A Sample Workflow](#)

This project is maintained by
JasonPBennett

Bonus:

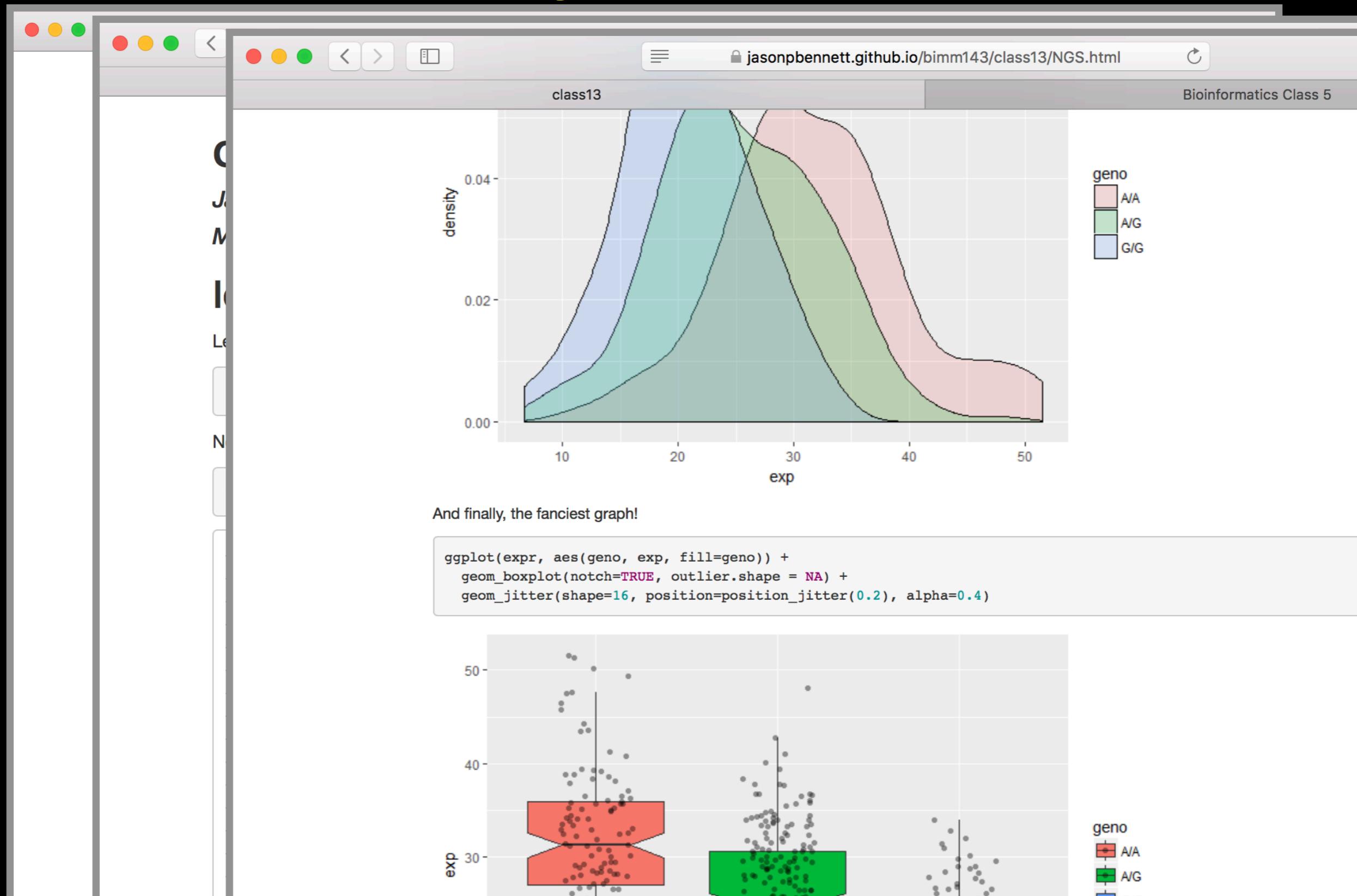
Online portfolio of **your** bioinformatics work!

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

- Address Bar:** jasonpbennett.github.io/bimm143/class13/NGS.html
- Page Title:** class13
- Page Subtitle:** Bioinformatics Class 5
- Content:**
 - # class13
 - Jason Patrick Bennett*
 - May 15, 2018*
 - ## Identifying SNP's in a Population
 - Lets analyze SNP's from the Mexican-American population in Los Angeles:
 - ```
genotype <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
```
  - Now lets look at a table of the data:
  - ```
table(genotype)
```
 - ```
, , Population.s. = ALL, AMR, MXL, Father = -, Mother = -
##
Genotype..forward.strand.
Sample..Male.Female.Unknown. A|A A|G G|A G|G
NA19648 (F) 1 0 0 0
NA19649 (M) 0 0 0 1
NA19651 (F) 1 0 0 0
NA19652 (M) 0 0 0 1
NA19654 (F) 0 0 0 1
NA19655 (M) 0 1 0 0
NA19657 (F) 0 1 0 0
NA19658 (M) 1 0 0 0
NA19661 (M) 0 1 0 0
NA19663 (F) 1 0 0 0
NA19664 (M) 0 0 1 0
NA19669 (F) 1 0 0 0
```

# Bonus:

## Online portfolio of **your** bioinformatics work!



## **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 biology majors 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!

## **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 biology majors 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!

# BIMM-143 Learning Goals....

## Data science R based learning goals

UCSanDiego

BIMM 143

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

bioboot.github.io/bimm143\_W18/goals/

BIMM 143 Home Gmail Gcal Bitbucket GitHub News Disqus BGGN-213 BIMM-143 GDocs

|    |                                                                                                                                                                                                                                                                                         |                          |
|----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------|
| 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 R to read and parse comma-separated (.csv) formatted files ready for subsequent analysis.                                                                                                                                                                                           | 8, 9, 10, 11, 13, 15, 16 |
| 7  | 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    |
| 8  | View and interpret the structural models in the PDB.                                                                                                                                                                                                                                    | 10, 11                   |
| 9  | Explain the outputs from structure prediction algorithms and small molecule docking approaches.                                                                                                                                                                                         | 11                       |
| 10 | Appreciate and describe in general terms the rapid advances in sequencing technologies and the new areas of investigation that these advances have made accessible.                                                                                                                     | 13, 14, 15               |
| 11 | Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.                                                                                                                                       | 13                       |
| 12 | 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                       |

# BIMM-143 Learning Goals....

Delve deeper into “real-world” bioinformatics

UC San Diego

**BIMM 143**

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**

|    | view and interpret the structural models in the PDB.                                                                                                                                         | 10, 11     |
|----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| 9  | Explain the outputs from structure prediction algorithms and small molecule docking approaches.                                                                                              | 11         |
| 10 | Appreciate and describe in general terms the rapid advances in sequencing technologies and the new areas of investigation that these advances have made accessible.                          | 13, 14, 15 |
| 11 | Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.                                            | 13         |
| 12 | 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         |
| 13 | 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     |
| 14 | 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         |
| 15 | Use the KEGG pathway database to look up interaction pathways.                                                                                                                               | 17         |
| 16 | Use graph theory to represent biological data networks.                                                                                                                                      | 17, 18     |
| 17 | Understand the challenges in integrating and interpreting large heterogeneous high throughput data sets into their functional                                                                | 19         |

# **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

Genomic data analysis

# IEEE 2016 Top Programming Languages

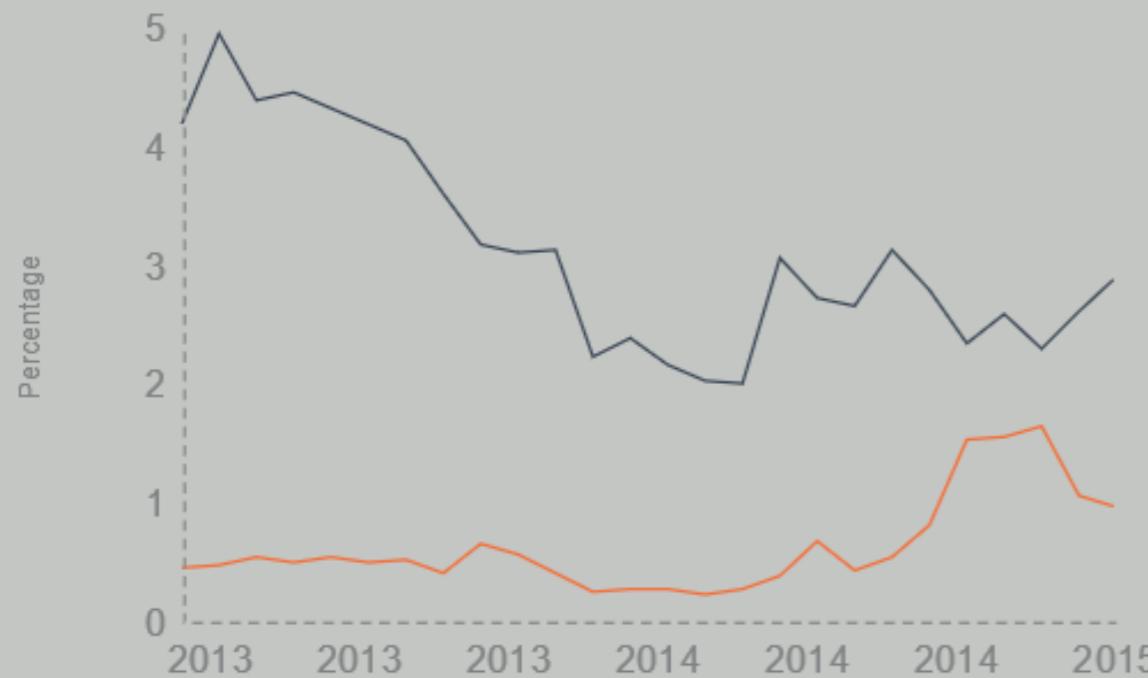
| Language Rank | Types | Spectrum Ranking |
|---------------|-------|------------------|
| 1. C          |       | 100.0            |
| 2. Java       |       | 98.1             |
| 3. Python     |       | 98.0             |
| 4. C++        |       | 95.9             |
| 5. R          |       | 87.9             |
| 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>

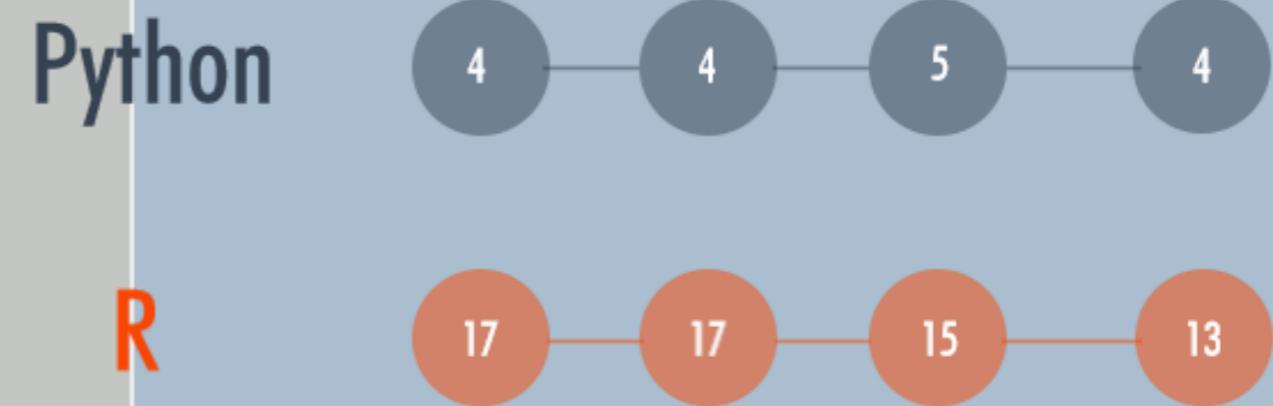
# R and Python: The Numbers

## Popularity Rankings

R and Pythons 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)



## Jobs And Salary?

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



\$ 115,531



Python

\$ 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 and was designed specifically for data analysis.
- Large friendly user and developer community.
  - As of Jan 6th 2019 there are 13,645 add on **R packages** on **CRAN** and 1,649 on **Bioconductor** - more on these later!
- Virtually every statistical technique is either already built into R, or available as a free package.
- Unparalleled data analysis environment for **high-throughput genomic data**.

# Past Student Opinions...

etherpad.net/p/bimm143\_f18

bimm143 f18 | etherpad.net Pad eGrade-BIMM143\_W19 - Google Sheets

B I U S 1 2 3 4 5 6 C Style A

7  
8 **Q1. Did you enjoy this course in relation to others you have experienced at UCSD?**  
9 Hell Yeah!  
10 Yes  
11 it was too lit  
12 Yes!  
13 Yes!  
14 yes  
15 yes!  
16 I do too!  
17 One of the best  
18 The best  
19 yes  
20 Ye  
21 Yes  
22 yes  
23 Yes  
24 Yes  
25 yes!  
26 yes, one of the most useful classes I've had  
27 no but im just really bad at coding so thats just me <—Don't be discouraged! It takes time. No one starts as a master. :)  
28

Chat 0

# Past Student Opinions...

The screenshot shows a web-based collaborative editing tool, Etherpad, with the URL [etherpad.net/p/bimm143\\_S18](http://etherpad.net/p/bimm143_S18). The pad is titled "bimm143 S18 | etherpad.net Pad". The interface includes standard browser controls at the top, followed by a toolbar with bold (B), italic (I), underline (U), strikethrough (S), and other styling options. Below the toolbar is a text area where students have typed their responses to a question. The responses are color-coded and numbered from 7 to 29.

**Q1. Did you enjoy this course in relation to others you have experienced at UCSD?**

7 Q1. Did yo  
8 Yes  
9 Hell Yeah!  
10 - Yes.  
11 Yes  
12 Yes  
13 yes, quite.  
14 yes  
15 - I enjoyed this lab course better than my other lab courses  
16 This is the best lab course I've taken at UCSD  
17 Yes  
18 Yes this course was very enjoyable and perhaps more relevant than others  
19 Yes even as a beginner +1  
20 Yes this course was interesting compared to other courses offered at UCSD+1  
21 This is one of the most enjoyable classes offered here! (:+1  
22 Yes  
23 Yes. I very much enjoyed this course.  
24 yes  
25 Yes!  
26 I enjoyed this course much more than many of my other courses at UCSD.  
27 This is one of the best and most useful courses I have taken at UCSD.  
28 Yes  
29 yes, it was a very relaxing course and I love how helpful and passionate the professor and the TA were.

# Past Student Opinions...

The image shows three separate etherpad.net pads side-by-side, each containing survey responses from students.

**bimm143 S18 | etherpad.net Pad**

**Q1. Did you**

- Yes
- Yes.
- Yes
- Yes
- yes, quite.
- yes
- I enjoyed this
- This is the best
- Yes
- Yes this course
- Yes even as a b
- Yes this course
- This is one of th
- Yes
- Yes. I very much
- yes
- Yes!
- I enjoyed this c
- This is one of th
- Yes
- yes, it was a ve

**bggn213 S18 | etherpad.net Pad**

**Q1. Did you enjoy this course in relation to others you have experience**

- Yes, very much
- Yes, absolutely!
- Yes
- Yes, I like the focus on applying R to real world biological datasets
- Yes
- yes
- Yes
- It was a lot harder than I was expecting
- yes
- Yes!
- yes
- Yes!
- yes
- Yes, I learned lots of things that are very useful in research but hard to learn ourselves
- Yes this class was awesome!
- Yes, this course was amazingly put together in a logical way and was extremely thorough.



| Instructor     | Course                                                   | Term | Rcmnd Class | Rcmnd Instr | Study Hrs/wk | Avg Grade Expected | Avg Grade Received |
|----------------|----------------------------------------------------------|------|-------------|-------------|--------------|--------------------|--------------------|
| Grant, Barry J | <a href="#">BIMM 143 - Bioinformatics Laboratory (A)</a> | FA18 | 100.0 %     | 100.0 %     | 4.50         | B+ (3.53)          | N/A                |
| Grant, Barry J | <a href="#">BIMM 143 - Bioinformatics Laboratory (A)</a> | SP18 | 94.7 %      | 94.7 %      | 5.66         | B+ (3.63)          | B+ (3.35)          |
| Grant, Barry J | <a href="#">BIMM 143 - Bioinformatics Laboratory (A)</a> | WI18 | 100.0 %     | 100.0 %     | 5.64         | B+ (3.64)          | N/A                |
| Grant, Barry J | <a href="#">BIMM 194 - Adv Topics-Molecular Bio (C)</a>  | WI18 | 92.9 %      | 100.0 %     | 1.30         | A (4.00)           | A (4.00)           |

Average for BIMM143: 98.2 % 98.2 % 5.27

# Today's Menu

## Course Logistics

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

## Learning Objectives

What you need to learn to succeed in this course.

## Course Structure

Major lecture topics and specific learning goals.

## Introduction to Bioinformatics

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

## Bioinformatics Database

Hands-on exploration of several major databases and their associated tools.

**Q. What is Bioinformatics?**

## **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...

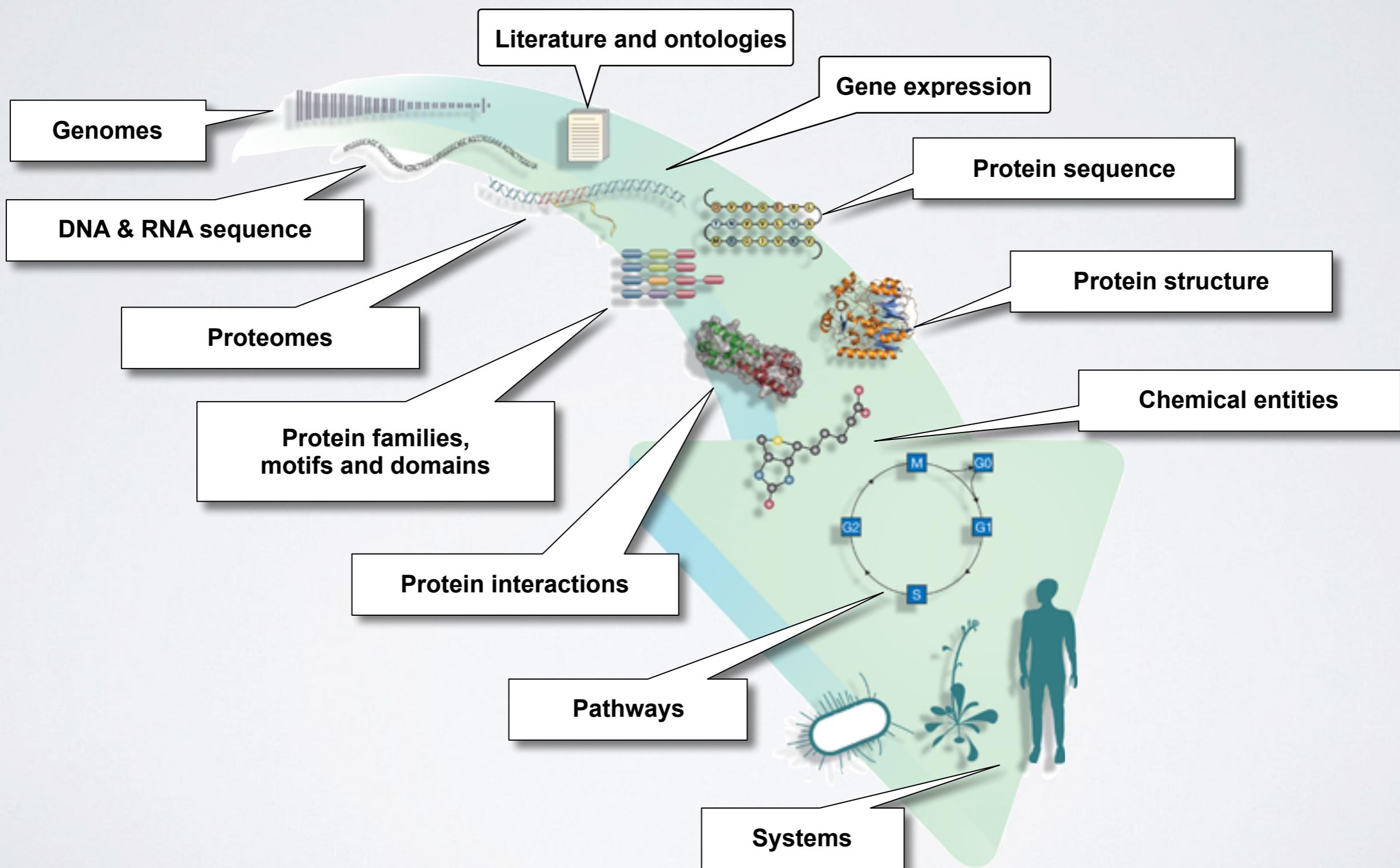
# MORE DEFINITIONS

- ▶ “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> )

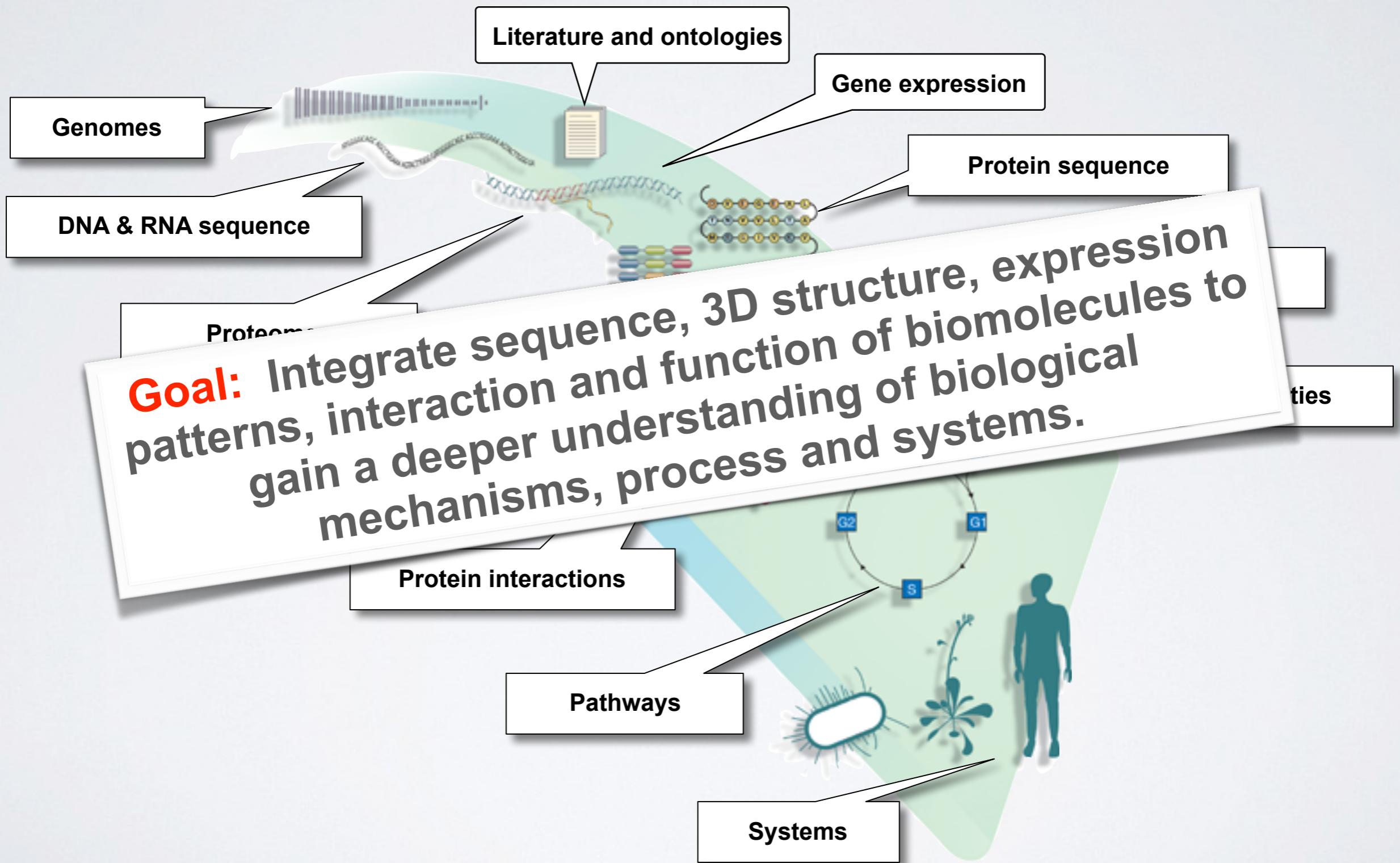
# MORE DEFINITIONS

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

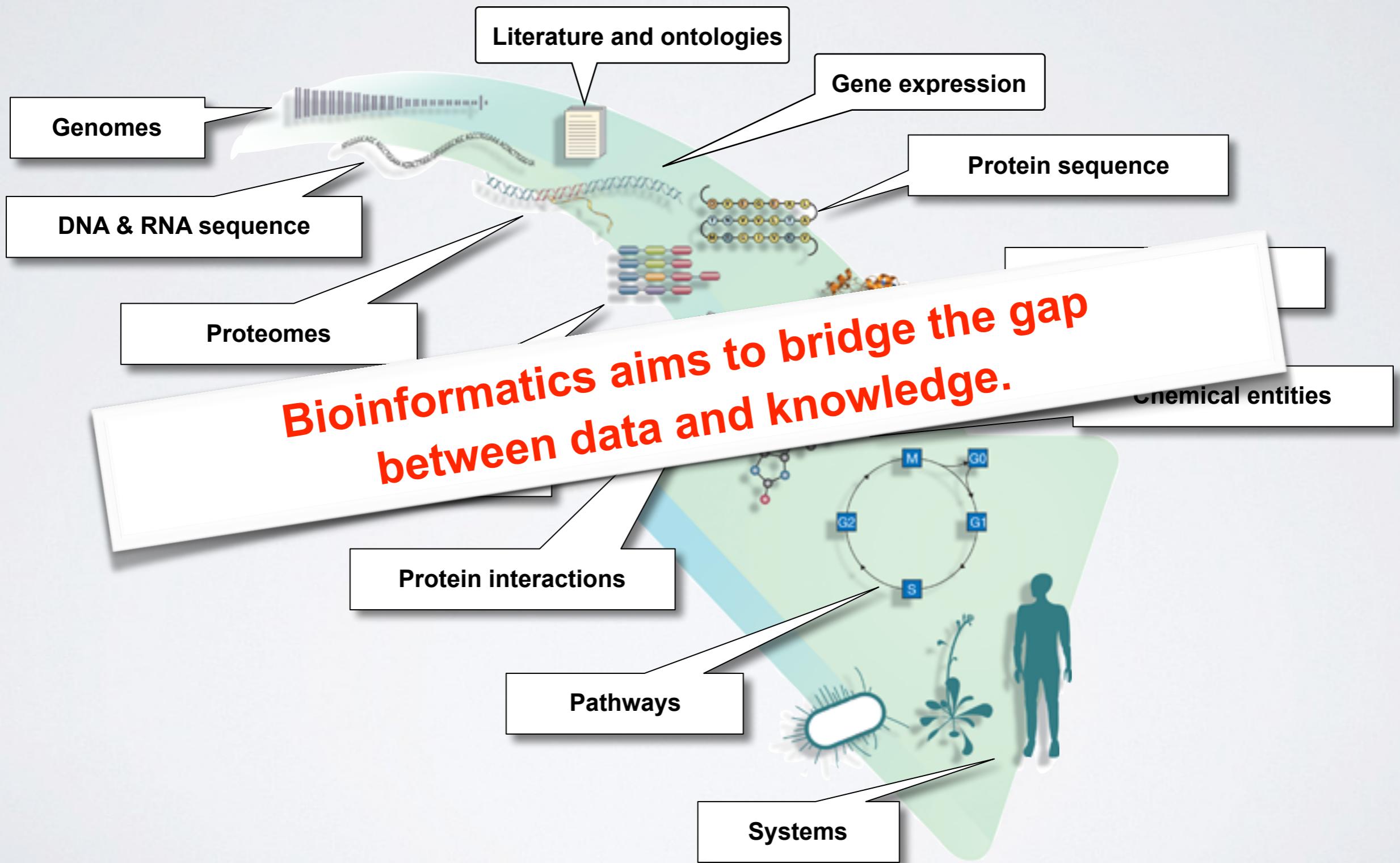
# Major types of Bioinformatics Data



# 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**

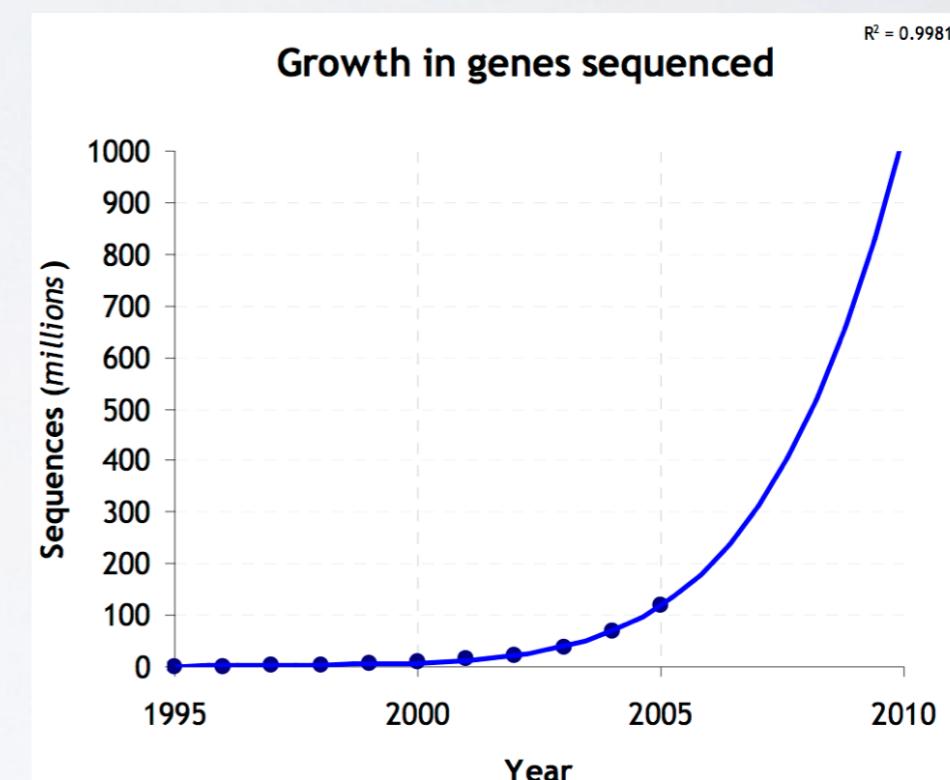
- *DNA sequence determines protein sequence.*
- *Protein sequence determines protein structure.*
- *Protein structure determines protein function.*
- *Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.*

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**

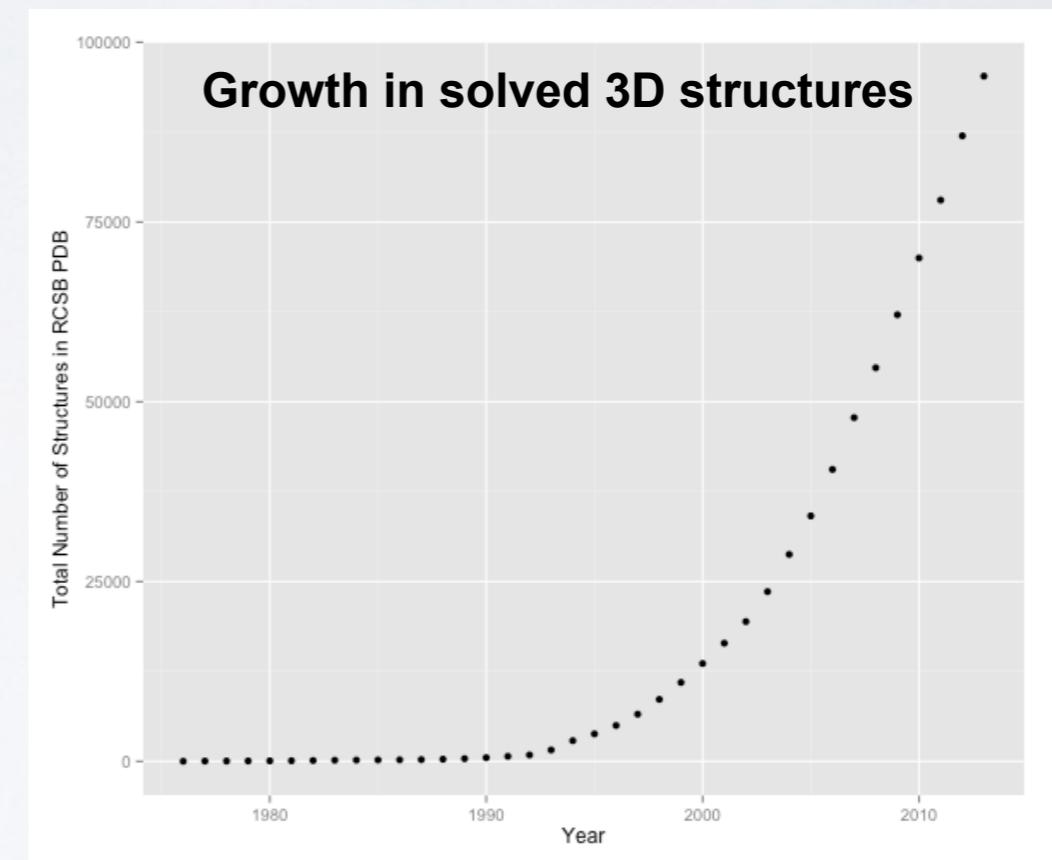


E.G. data from sequencing, structural genomics, proteomics, new high throughput assays, etc...

# Why do we need Bioinformatics?

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

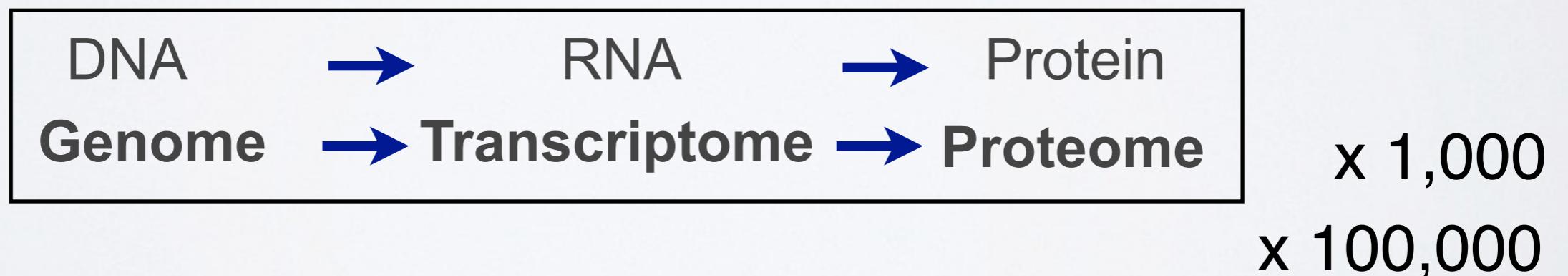
- Bioinformatics provides methods for the efficient:
  - **storage**
  - **annotation**
  - **search and retrieval**
  - **data integration**
  - **data mining and analysis**



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?

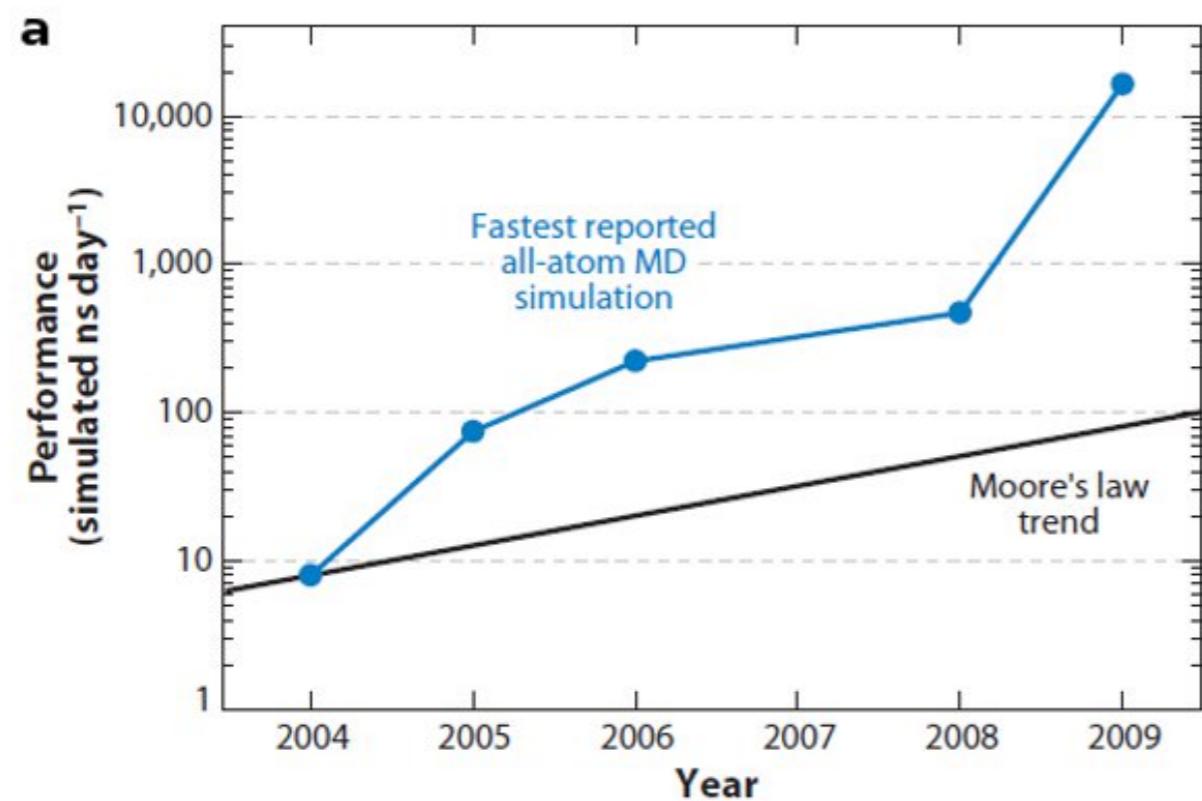
## Pre-packaged tools and databases

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

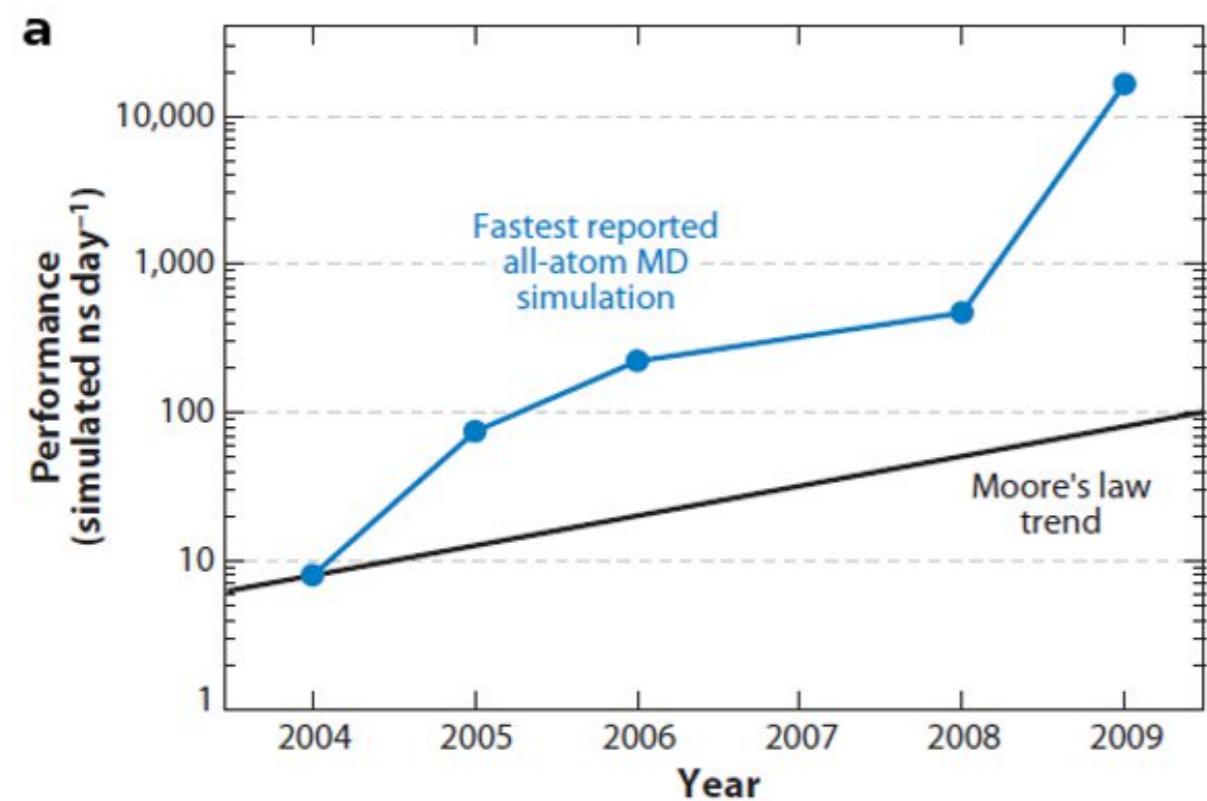
## Advanced tool application & development

- 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 ANDGPUS



# 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 shadow.



# 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 “*Sequence Analysis in Molecular Biology*” states:

→“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                                                        | 500                                                                                                             |
| Select the maximum number of aligned sequences to display <a href="#">?</a> |                                                                                                                 |
| Short queries                                                               | <input checked="" type="checkbox"/> Automatically adjust parameters for short input sequences <a href="#">?</a> |
| Expect threshold                                                            | 10                                                                                                              |
| Word size                                                                   | 3                                                                                                               |
| Max matches in a query range                                                | 0                                                                                                               |

### Scoring Parameters

|                           |                               |
|---------------------------|-------------------------------|
| Matrix                    | BLOSUM62                      |
| Gap Costs                 | Existence: 11 Extension: 1    |
| Compositional adjustments | Conditional compositional sco |

### Filters and Masking

|        |                                                                                                                           |
|--------|---------------------------------------------------------------------------------------------------------------------------|
| Filter | <input type="checkbox"/> Low complexity regions <a href="#">?</a>                                                         |
| Mask   | <input type="checkbox"/> Mask for lookup table only<br><input type="checkbox"/> Mask lower case letters <a href="#">?</a> |

### PSI/PHI/DELTA BLAST

|                      |                                                             |
|----------------------|-------------------------------------------------------------|
| Upload PSSM Optional | <input type="button" value="Choose File"/> no file selected |
| PSI-BLAST Threshold  | 0.005                                                       |
| Pseudocount          | 0                                                           |

Even Blast has many settable parameters

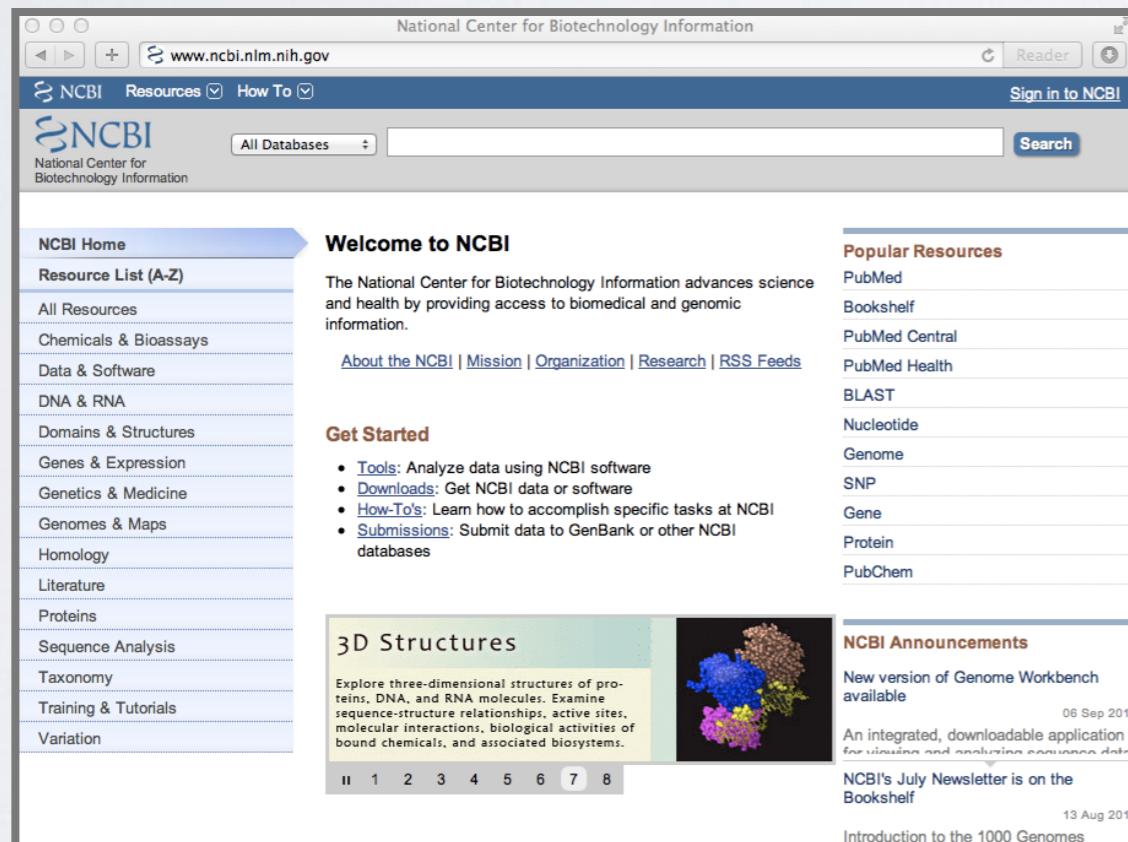
Related tools with different terminology

STEP 3 - Set your PROGRAM FASTA

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

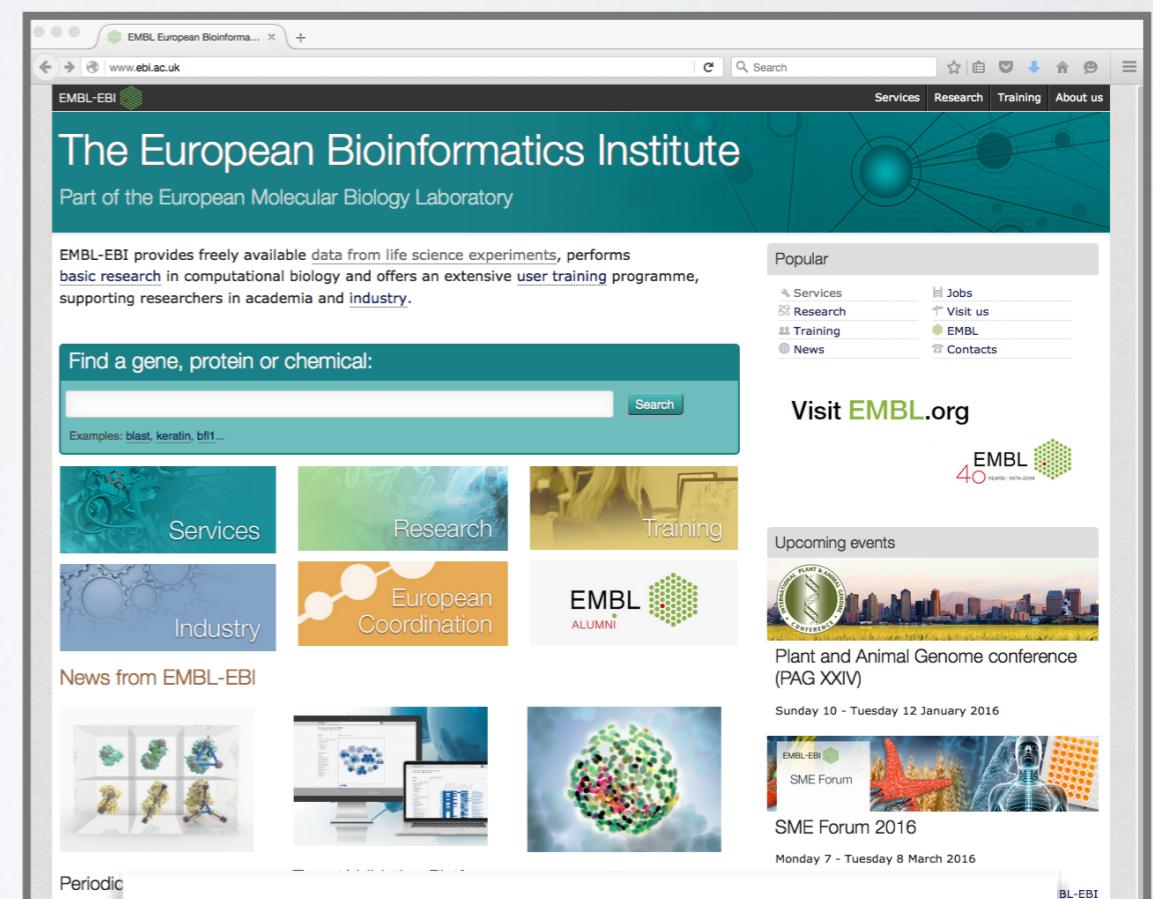
# Key Online Bioinformatics Resources: NCBI & EBI

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



The screenshot shows the homepage of the National Center for Biotechnology Information (NCBI). The top navigation bar includes links for 'Resources' (with a dropdown menu), 'How To', 'Sign in to NCBI', and a search bar. The main content area features a 'Welcome to NCBI' section, a 'Get Started' section with links to tools, downloads, and how-to guides, and a '3D Structures' section showing a molecular model. On the right, there's a sidebar titled 'Popular Resources' listing links to PubMed, Bookshelf, PubMed Central, and other databases. A 'NCBI Announcements' section highlights the new version of the Genome Workbench.

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



The screenshot shows the homepage of the European Bioinformatics Institute (EMBL-EBI). The top navigation bar includes links for 'Services', 'Research', 'Training', and 'About us'. The main content area features a search bar for finding genes, proteins, or chemicals, and sections for 'Services', 'Research', 'Training', 'Industry', 'European Coordination', and 'EMBL ALUMNI'. A 'News from EMBL-EBI' section shows images of scientific work. On the right, there's a sidebar titled 'Popular' with links to 'Services', 'Research', 'Training', and 'News'. Below the sidebar, there are sections for 'Visit EMBL.org', 'Upcoming events' (including the Plant and Animal Genome conference and SME Forum), and 'SME Forum 2016'.

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

# National Center for Biotechnology Information (NCBI)

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



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

National Center for Biotechnology Information

NCBI Resources How To Sign in to NCBI

All Databases Search

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

Welcome to NCBI

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

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

Get Started

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

3D Structures

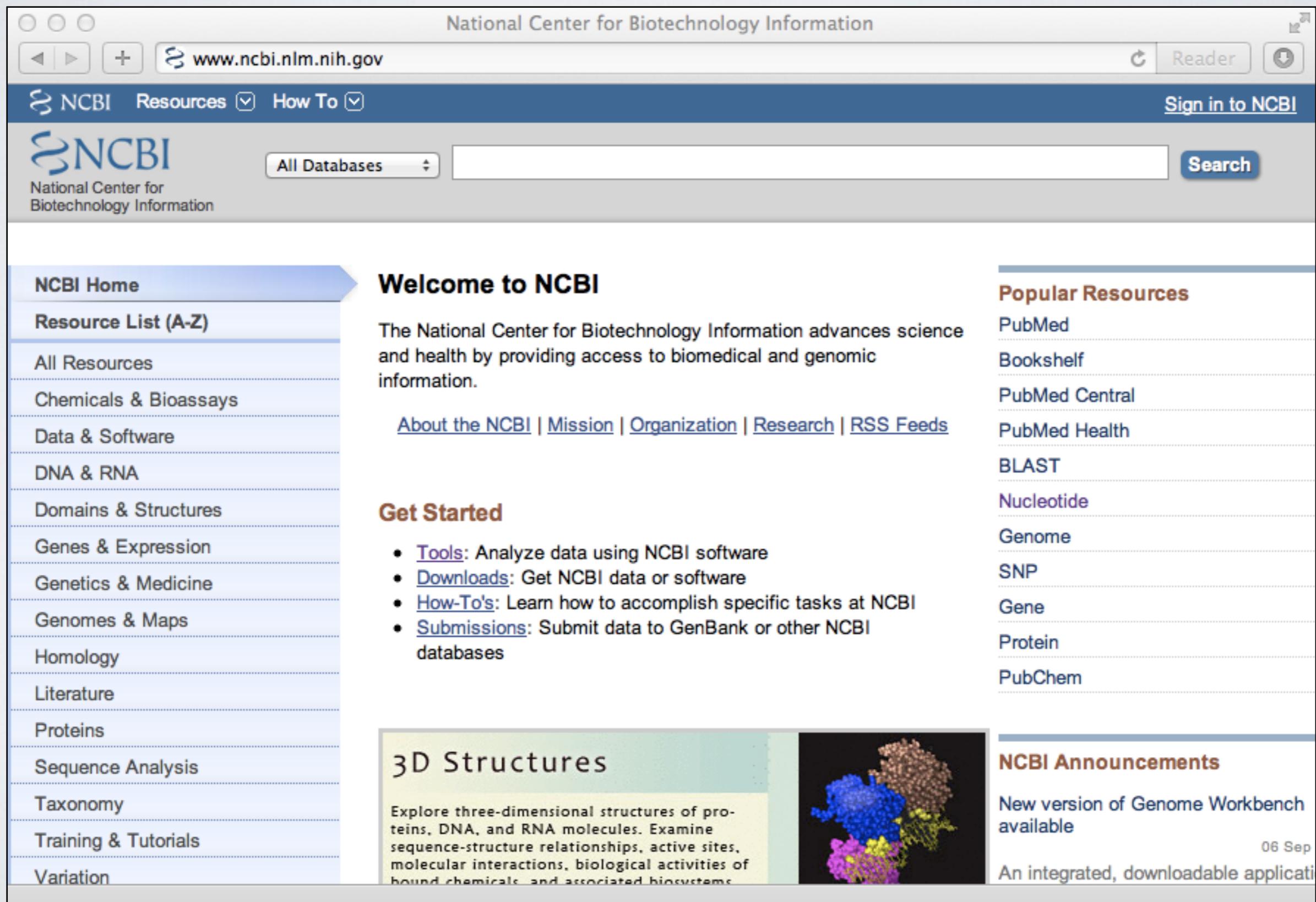
Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals, and associated biosystems.

Popular Resources

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

NCBI Announcements

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



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

National Center for Biotechnology Information

NCBI Resources How To Sign in to NCBI

All Databases

Search

Popular Resources

PubMed ←

Bookshelf

PubMed Central

PubMed Health

BLAST ←

Nucleotide

Genome

SNP ←

Gene

Protein

PubChem

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 unique information, tools and resources to support basic research and health by providing access to its databases and information.

About the NCBI | Mission | Our History

Get Started

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

3D Structures

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

New version of Genome Workbench available

06 Sep

An integrated, downloadable application

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

National Center for Biotechnology Information

NCBI Resources How To Sign in to NCBI

All Databases Search

NCBI Home Resource List (A-Z)

Welcome to NCBI  
The National Center for Biotechnology Information advances science

Popular Resources PubMed

Notable NCBI databases include:  
**GenBank**, **RefSeq**, **PubMed**, **dbSNP**

and the search tools **ENTREZ** and **BLAST**

Homology Literature Proteins Sequence Analysis Taxonomy Training & Tutorials Variation

databases

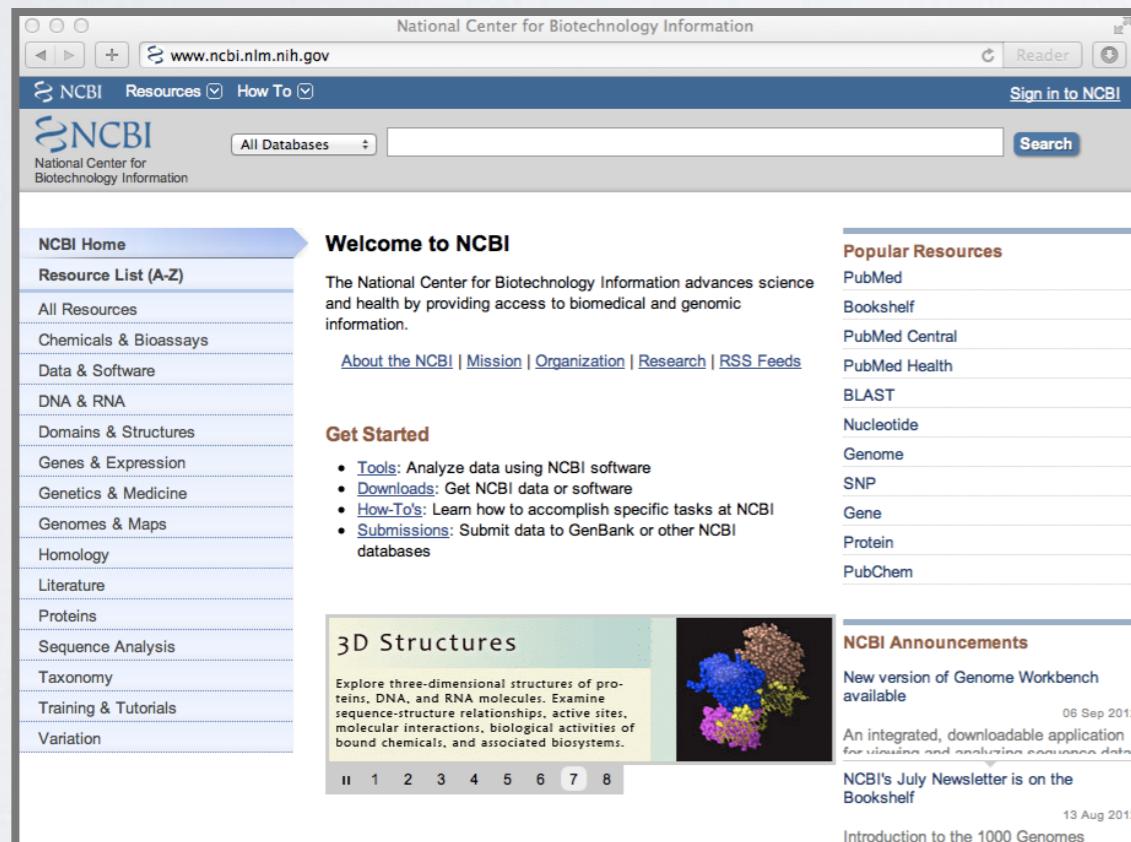
3D Structures  
Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals and associated biosystems

Protein PubChem

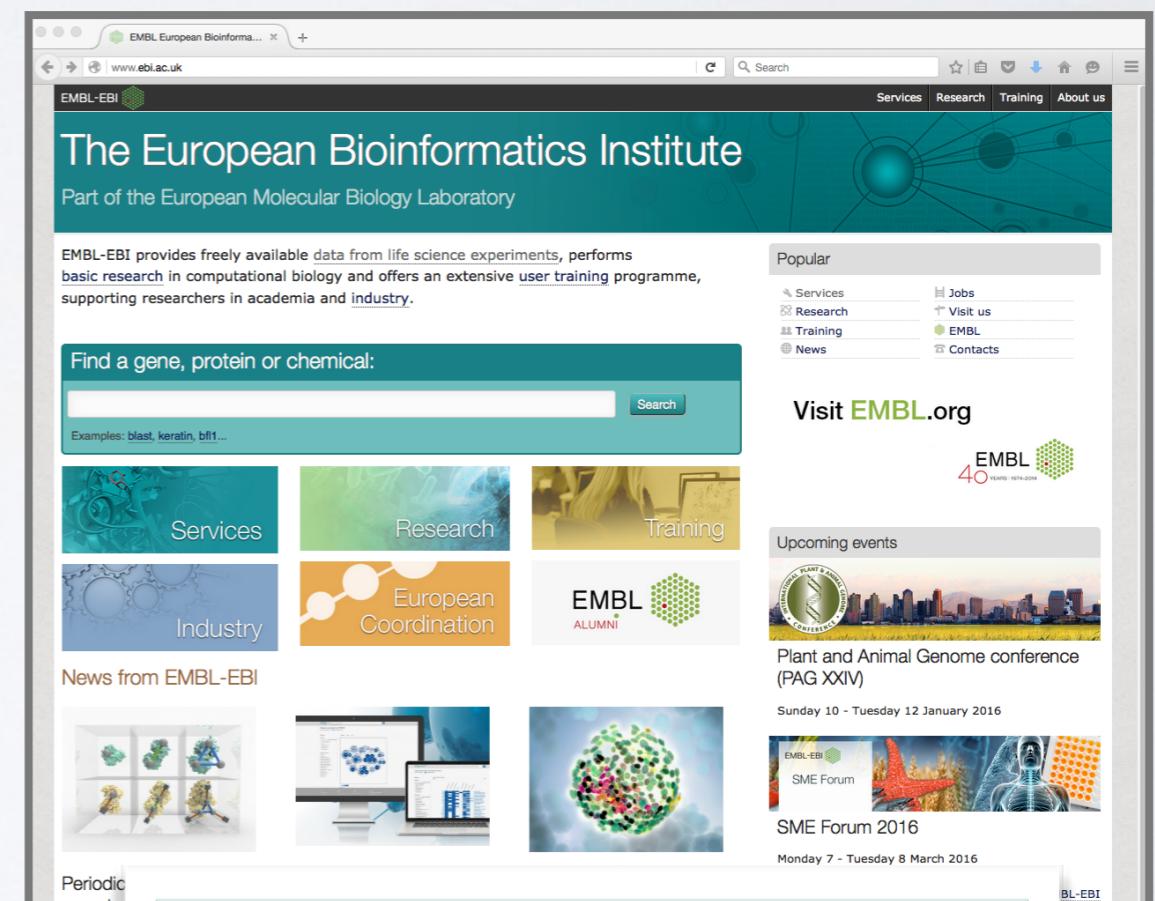
NCBI Announcements  
New version of Genome Workbench available 06 Sep  
An integrated, downloadable applicati

# Key Online Bioinformatics Resources: NCBI & EBI

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



The screenshot shows the NCBI homepage with a blue header bar containing the NCBI logo, a search bar, and links for "Resources", "How To", and "Sign in to NCBI". Below the header is a navigation menu with links to "NCBI Home", "Resource List (A-Z)", "All Resources", "Chemicals & Bioassays", "Data & Software", "DNA & RNA", "Domains & Structures", "Genes & Expression", "Genetics & Medicine", "Genomes & Maps", "Homology", "Literature", "Proteins", "Sequence Analysis", "Taxonomy", "Training & Tutorials", and "Variation". A "Popular Resources" sidebar on the right lists "PubMed", "Bookshelf", "PubMed Central", "PubMed Health", "BLAST", "Nucleotide", "Genome", "SNP", "Gene", "Protein", and "PubChem". The main content area features a "Welcome to NCBI" section, a "Get Started" section with links to tools, downloads, how-to's, and submissions, and a "3D Structures" section showing a molecular model.



The screenshot shows the EMBL-EBI homepage with a teal header bar containing the EMBL-EBI logo, a search bar, and links for "Services", "Research", "Training", and "About us". Below the header is a main content area with a teal banner stating "The European Bioinformatics Institute Part of the European Molecular Biology Laboratory". It features a "Find a gene, protein or chemical:" search bar, several colored boxes for "Services", "Research", "Training", "Industry", "European Coordination", and "EMBL ALUMNI", and a "News from EMBL-EBI" section with images of scientific data. On the right, there's a "Popular" sidebar with links to "Services", "Research", "Training", and "News", and a "Visit EMBL.org" section with the EMBL 40th anniversary logo and information about the Plant and Animal Genome conference (PAG XXIV) and the SME Forum 2016.

<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 homepage of the EMBL European Bioinformatics Institute (EBI) at [www.ebi.ac.uk](http://www.ebi.ac.uk). The page features a dark blue header with the EMBL-EBI logo and navigation links for Services, Research, Training, and About us. Below the header is a teal banner with the text "The European Bioinformatics Institute" and "Part of the European Molecular Biology Laboratory". A search bar is located above a main content area. The content area includes a section about EMBL-EBI's mission, a search bar for finding genes, proteins, or chemicals, and several promotional boxes for Services, Research, Training, and EMBL ALUMNI. On the right side, there are sections for Popular links (Services, Research, Training, News, Jobs, Visit us, EMBL, Contacts), a "Visit EMBL.org" link with the EMBL 40th anniversary logo, an "Upcoming events" section featuring the Plant and Animal Genome conference (PAG XXIV), and a "News from EMBL-EBI" section with three thumbnail images.

EMBL European Bioinforma... [www.ebi.ac.uk](#) Search Services Research Training About us

# The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

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

Find a gene, protein or chemical:

Examples: blast, keratin, bfl1...

Search

Services

Research

Training

EMBL ALUMNI

Upcoming events

Plant and Animal Genome conference (PAG XXIV)

Sunday 10 - Tuesday 12 January 2016

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

The screenshot shows the EBI Services website ([www.ebi.ac.uk/services](http://www.ebi.ac.uk/services)) with a teal header and a sidebar on the left listing services like DNA & RNA, Gene expression, Proteins, Systems, Chemical biology, Ontologies, Literature, and Cross domain. A red box highlights the 'Proteins' service. On the right, a 'Popular' section lists Ensembl, UniProt, PDB, ArrayExpress, and ChEMBL, with a red box around Ensembl. Below this is a banner featuring a monarch butterfly and the word 'Training'.

Services < EMBL-EBI

www.ebi.ac.uk/services

EMBL-EBI

Services | Research | Training | About us

## Services

Overview | A to Z | Data submission | Support

### Bioinformatics services

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

**DNA & RNA**  
genes, genomes & variation

**Gene expression**  
RNA, protein & metabolite expression

**Proteins**  
sequences, families & motifs

**Structures**  
Molecular & cellular structures

**Systems**  
reactions, interactions & pathways

**Chemical biology**  
chemogenomics & metabolomics

**Ontologies**  
taxonomies & controlled vocabularies

**Literature**  
Scientific publications & patents

**Cross domain**  
cross-domain tools & resources

### Popular

**Ensembl**

**UniProt**

**PDB**

**ArrayExpress**

**ChEMBL**

Training

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

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

## Proteins

### Popular services



#### UniProt: The Universal Protein Resource

The gold-standard, comprehensive resource for protein sequence and functional annotation data.



#### InterPro

A database for the classification of proteins into families, domains and conserved sites.



#### PRIDE: The Proteomics Identifications Database

An archive of protein expression data determined by mass spectrometry.



#### Pfam

A database of hidden Markov models and alignments to describe conserved protein families and domains.



#### Clustal Omega

Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.



#### HMMER - protein homology search

Fast sensitive protein homology searches using profile hidden Markov models (HMMs). Variety of different search methods for querying against both sequence and HMM target databases.



#### InterProScan 5

InterProScan 5 searches sequences against InterPro's predictive protein signatures. Please note that [InterProScan 4.8 has been retired](#).

### Quick links

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

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

The screenshot shows the homepage of the EMBL European Bioinformatics Institute (EBI) at [www.ebi.ac.uk](http://www.ebi.ac.uk). The page features a dark teal header with the EMBL-EBI logo and navigation links for Services, Research, Training, and About us. Below the header is a large banner with the text "The European Bioinformatics Institute" and "Part of the European Molecular Biology Laboratory". A search bar and a "Popular" sidebar are also visible.

**Popular**

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

**Find a gene, protein or chemical:**

Examples: blast, keratin, bfl1...

**Services**

**Research**

**Training**

**EMBL ALUMNI**

**Industry**

**European Coordination**

**News from EMBL-EBI**

**Visit EMBL.org**

**Upcoming events**

**Plant and Animal Genome conference (PAG XXIV)**

Sunday 10 - Tuesday 12 January 2016

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

The screenshot shows a web browser displaying the EBI Training online course page. The URL in the address bar is [www.ebi.ac.uk/training/online/course/using-sequence-similarity-searching-tools-embl-ebi](http://www.ebi.ac.uk/training/online/course/using-sequence-similarity-searching-tools-embl-ebi). The page title is "Using sequence similarity searching tools at EMBL-EBI: webinar". The main content area shows a thumbnail of the webinar video, which features a blue background with white text and a portrait of a man. Below the thumbnail, a video player interface shows the start time as 0:00 / 37:42. To the left of the video, there's a sidebar with "Course content" sections for "Using sequence similarity searching tools at EMBL-EBI: webinar" and "Contributors". A "Print Course" link is also present. On the right side, there are "Popular" links for "Train online", "Find us", and "Funding", and a "Find us at..." section with links for "Open days and career days", "Conference exhibitions", "EMBL courses and events", "Genome campus events", and "Science for schools". The top navigation bar includes links for "Services", "Research", "Training" (which is highlighted), and "About us".

Using sequence similarity searching tools at EMBL-EBI: webinar

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

EMBL-EBI

0:00 / 37:42

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

A screenshot of a web browser displaying the EBI Train online website. The title bar reads "Train online | EBI Train online". The address bar shows the URL "www.ebi.ac.uk/training/online/". The page header includes the EMBL-EBI logo, a search bar, and links for "Find", "Help", and "Feedback". A red "Beta" badge is visible in the top right corner. The main menu bar has links for "Databases", "Tools", "Research", "Training", "Industry", "About Us", and "Help". A secondary navigation bar on the left is titled "Navigation" and includes a link to "Train online Home". The main content area features a large heading "Notable EBI databases include:" followed by a list of databases: ENA, UniProt, Ensembl.

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](#)

# Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, BiolImage, 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, RCCP, Beanref, TKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZy, ChickGBASE, Colibri, COPE, CottonDB, dbSTS, DDBJ, DGP, DictyDb, ECGC, EC02DBASE, FlyBase, GDB, HEPDB, KEGG, MHCDB, MycoDB, PDBe, PDB, 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 ..!!!!

**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.

# Today's Menu

## Course Logistics

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

## Learning Objectives

What you need to learn to succeed in this course.

## Course Structure

Major lecture topics and specific learning goals.

## Introduction to Bioinformatics

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

## Bioinformatics Database

**Hands-on** exploration of several major databases and their associated tools.

# Your Turn!

[https://bioboot.github.io/bimm143\\_W19/lectures/#1](https://bioboot.github.io/bimm143_W19/lectures/#1)

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

- Title Bar:** Shows the URL [https://bioboot.github.io/bimm143\\_W18/lectures/#1](https://bioboot.github.io/bimm143_W18/lectures/#1).
- Header:** BIMM 143 Home Gmail Gcal Bitbucket GitHub News Disqus BGGN-213 BIMM-143 GDocs
- Left Sidebar (UC San Diego BIMM 143):**
  - UCSanDiego**
  - BIMM 143**
  - A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.
  - Navigation links: Overview, Lectures (highlighted with a red box), Computer Setup, Learning Goals, Assignments & Grading, Ethics Code.
- Main Content (Section 1: Welcome to Foundations of Bioinformatics):**
  - Topics:** Course introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Student 30-second introductions, Student computer setup.
  - Goals:**
    - Understand course scope, expectations, logistics and [ethics code](#).
    - Understand the increasing necessity for computation in modern life sciences research.
    - Get introduced to how bioinformatics is practiced.
    - Complete the [pre-course questionnaire](#).
    - Setup your [laptop computer](#) for this course.
  - Material:**
    - Lecture Slides: [Large PDF](#), [Small PDF](#),
    - Lab: [Hands-on section worksheet](#)
    - Feedback: [Muddy Point Assessment](#) (highlighted with a red box)
    - Handout: [Class Syllabus](#)
    - Computer [Setup Instructions](#).

## BIMM-143: INTRODUCTION TO BIOINFORMATICS (Lecture 1)

### **Bioinformatics Databases and Key Online Resources**

[https://bioboot.github.io/bimm143\\_W18/lectures/#1](https://bioboot.github.io/bimm143_W18/lectures/#1)

Dr. Barry Grant

Jan 2018

**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
ATGGTGCATCTGACTCCTGTGGAGAAGTCTGCCGTTACTGCCCTGTGGGCAAGGTGAACGTGGATGAAG
TTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCCTGGACCCAGAGGTTCTTGAGTCCTTGG
GGATCTGTCCACTCCTGATGCAGTTATGGGCAACCCCTAACGGTAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTAGTGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTGCCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGCAACGTGCTGGTCTGTGTGCTGGCCA
TCACTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCCCACAAGTATCACTAAGCTCGCTTCTTGCTGTCCAATTT
```

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 | [10:35 am]   |
| 2. GENE database @ NCBI           | [10:55 am]   |
| — BREAK —                         | — 11:05 am — |
| 3. UniProt & Muscle @ EBI         | [11:30 am]   |
| 4. PFAM, PDB & NGL                | [12:00 pm]   |
| — BREAK —                         | — 12:10 am — |
| 5. Extension exercises            | [12:40 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.
- 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/bimm143\\_W19/lectures/#1](https://bioboot.github.io/bimm143_W19/lectures/#1)

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

