



**BIMM 143**

**Introduction to Bioinformatics**

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

<http://thegrantlab.org/bimm143>

HELLO  
my name is

*BARRY*

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

HELLO  
*HIS* — my name is

*ALEX*

[arsharp@ucsd.edu](mailto:arsharp@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/bimm143/>

bioboot.github.io/bimm143\_w18/

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

# 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 [\[a\]](#).

**Overview**

**Lectures**

**Computer Setup**

**Learning Goals**

**Assignments & Grading**

**Ethics Code**

## Bioinformatics (BIMM 143, Winter 2018)

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

**Instructional Assistant**  
Alexzander Sharp (Email: [arsharp@ucsd.edu](mailto:arsharp@ucsd.edu))

**Course Syllabus**  
[Winter 2018 \(PDF\)](#) [\[a\]](#)

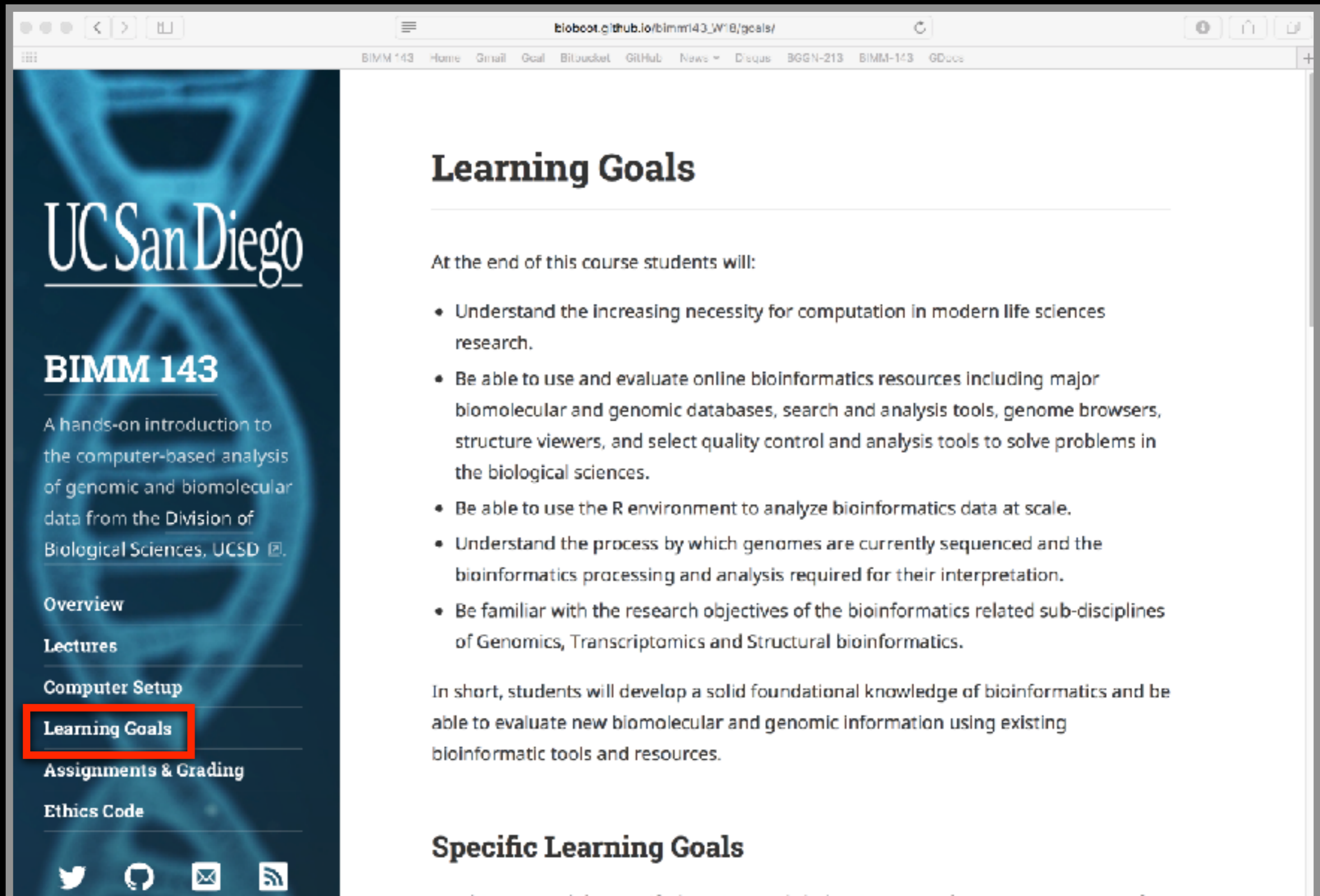
## Overview

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.

An integrated lecture/lab structure with hands-on exercises and small-scale projects emphasizes modern developments in genomics and proteomics. A detailed listing of

# 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/`. The browser's address bar and tabs are visible at the top. The page content is divided into a left sidebar and a main content area.

**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**

**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**

## **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!

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

### Specific Learning Goals

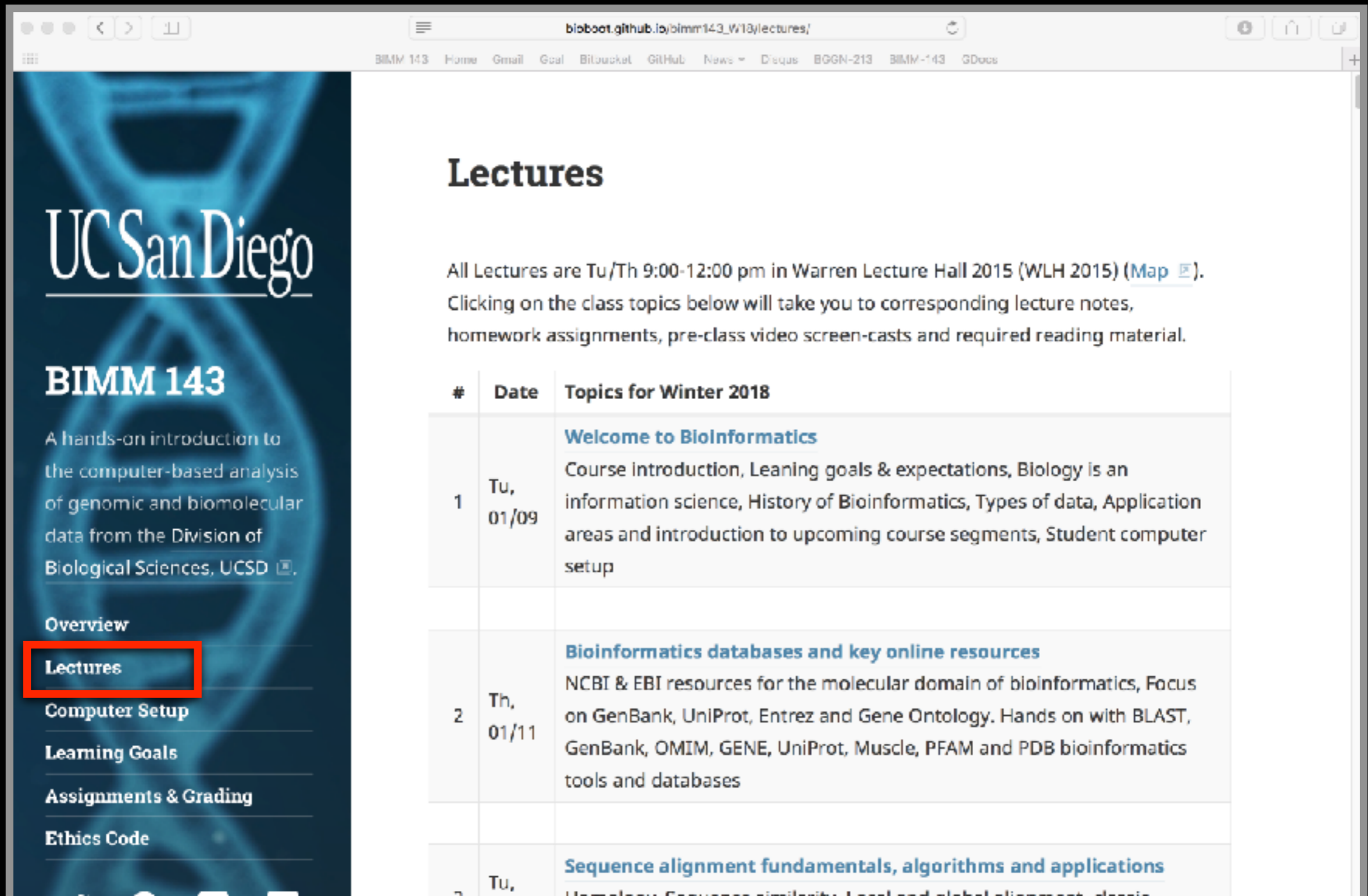
Teaching toward the specific learning goals below is expected to occupy 60%-70% of class time. The remaining course content is at the discretion of the instructor with student body input. This includes student selected topics for peer presentation, as well as one student selected guest lecture from an 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_w18/lectures/`. The page is titled "Lectures" and contains the following text:

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

| # | Date      | Topics for Winter 2018  |
|---|-----------|---|
| 1 | Tu, 01/09 | <a href="#">Welcome to Bioinformatics</a><br>Course Introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Student computer setup                                  |
| 2 | Th, 01/11 | <a href="#">Bioinformatics databases and key online resources</a><br>NCBI & EBI resources for the molecular domain of bioinformatics, Focus on GenBank, UniProt, Entrez and Gene Ontology. Hands on with BLAST, GenBank, OMIM, GENE, UniProt, Muscle, PFAM and PDB bioinformatics tools and databases |
| 3 | Tu,       | <a href="#">Sequence alignment fundamentals, algorithms and applications</a><br>Homology, Sequence similarity, Local and global alignment, classic  |

The left sidebar of the page contains the following navigation links:

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

# Course Structure

Derived from specific learning goals

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

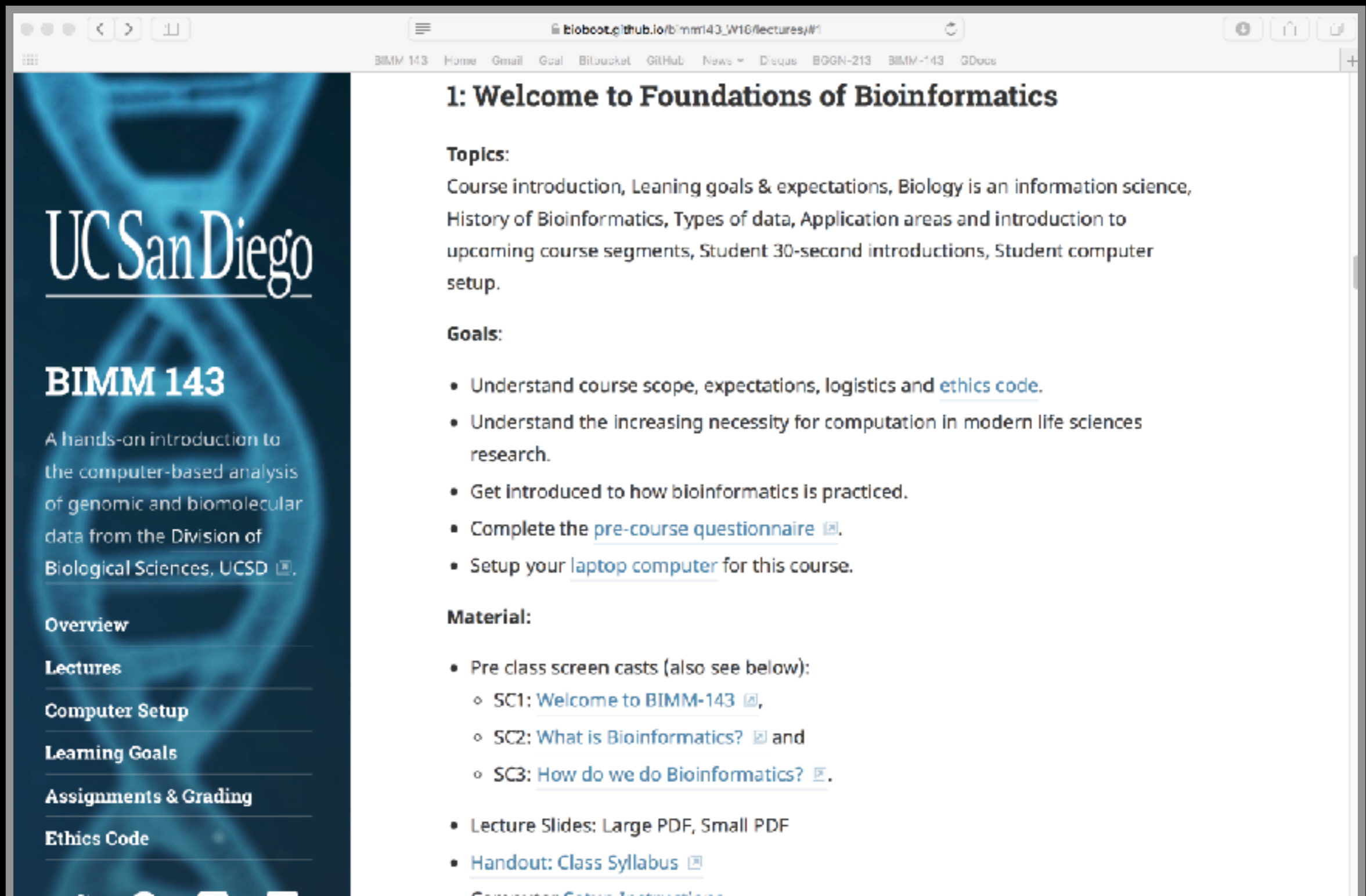
### Lectures

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

| # | Date      | Topics for Winter 2018   |
|---|-----------|--|
| 1 | Tu, 01/09 | <b>Welcome to Bioinformatics</b><br>Course Introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Student computer setup                                  |
| 2 | Th, 01/11 | <b>Bioinformatics databases and key online resources</b><br>NCBI & EBI resources for the molecular domain of bioinformatics, Focus on GenBank, UniProt, Entrez and Gene Ontology. Hands on with BLAST, GenBank, OMIM, GENE, UniProt, Muscle, PFAM and PDB bioinformatics tools and databases |
| 3 | Tu,       | <b>Sequence alignment fundamentals, algorithms and applications</b><br>Homology, Sequence similarity, Local and global alignment, classic  |

# Class Details

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



The screenshot shows a web browser window with the address bar displaying `bioboot.github.io/bimm143_w16/lectures/#1`. The browser tabs include BIMM 143, Home, Gmail, Goal, Bitbucket, GitHub, News, Disqus, BGGN-213, BIMM-143, and GDocs. The page content is as follows:

### 1: Welcome to Foundations of Bioinformatics

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

**Goals:**

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

**Material:**

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

**UC San Diego**

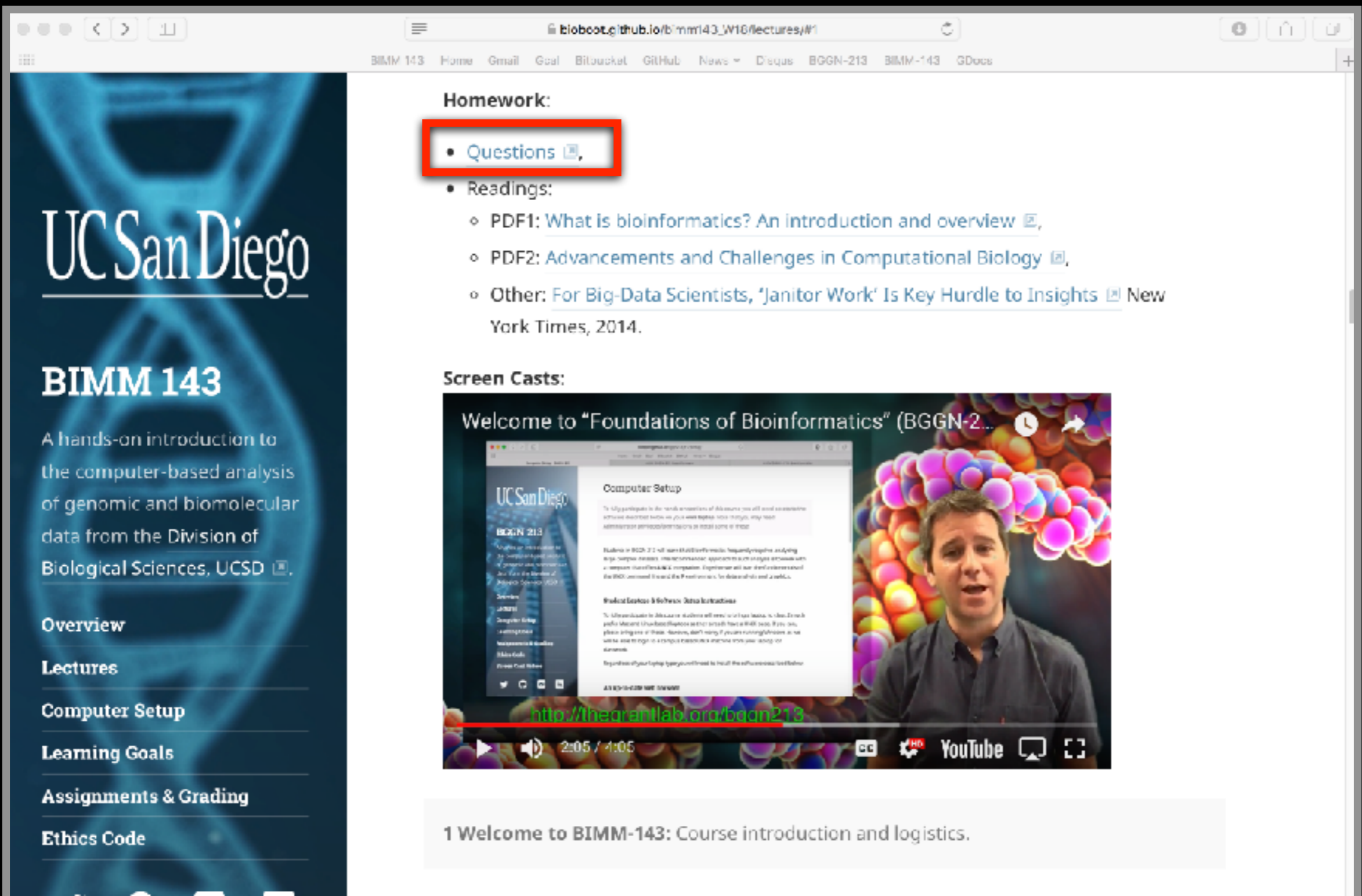
## BIMM 143

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

# Homework

## Goals, Class material, Screencasts & Homework



The screenshot shows a web browser window with the address bar displaying `bioboot.github.io/bimm143_w16/lectures/#1`. The browser tabs include BIMM 143, Home, Gmail, Goal, Bitbucket, GitHub, News, Disqus, BGGN-213, BIMM-143, and GDocs.

**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

**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-2...)

<http://theoranalab.org/bggm213>

2:05 / 4:05

YouTube

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

# Homework

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

The image shows a screenshot of a Google Forms page titled "BIMM143 Lecture 1 Homework (W18)". The browser's address bar shows the URL "docs.google.com/forms/d/e/1FAIpQLSeqDGQCYYIWKovPsC3Unk4SsgAEdTHRJp...". The page has a teal header and a light blue background. The form content includes:

- A title: "BIMM143 Lecture 1 Homework (W18)"
- A prompt: "Please answer the following questions"
- A red asterisk indicating a required field: "\* Required"
- A text input field labeled "Email address \*" with the placeholder text "Your email".
- A multiple-choice question: "Which of the following operating systems is most frequently used for bioinformatics tool development" (1 point). The options are Windows, iOS, Unix, and Perl.
- A multiple-choice question: "Which of the following databases contains primarily protein sequences" (1 point). The visible option is GenBank.

# Homework

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

The image shows a screenshot of a Google Forms page titled "BIMM143 Lecture 1 Homework (W18)". The page is displayed in a browser window with a teal header. A prominent red banner with white text is overlaid diagonally across the center of the page, stating "Homework is due before the next weeks class!". Below the banner, the form content is visible, including the title, a request to answer questions, a "Required" indicator, and two multiple-choice questions. The first question asks for the most frequently used operating system for bioinformatics tool development, with options: Windows, iOS, Unix, and Perl. The second question asks for the database containing primarily protein sequences, with the option "GenBank" visible.

docs.google.com/forms/d/e/1FAIpQLSeqDGQCYYWkVpsC3Unk4SsgAEdTHRJp-  
BIMM 143 Home Gmail Goal Bitbucket GitHub News Disqus BGGN-213 BIMM-143 GDocs

## BIMM143 Lecture 1 Homework (W18)

Please answer the following questions

\* Required

Which of the following operating systems is most frequently used for bioinformatics tool development 1 point

- Windows
- iOS
- Unix
- Perl

Which of the following databases contains primarily protein sequences 1 point

- GenBank

## 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

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

|    |   |                          |
|----|---|--------------------------|
| 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                       |
|    | Given an RNA-Seq data file, find the set of significantly differentially  |                          |

# 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

| Goal Number | Goal Description   | Associated Lectures |
|-------------|--|---------------------|
| 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                  |
| 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 heterogenous 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.



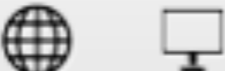





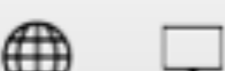

# Why use R?

Productivity

Flexibility

Designed for data analysis

# IEEE 2016 Top Programming Languages

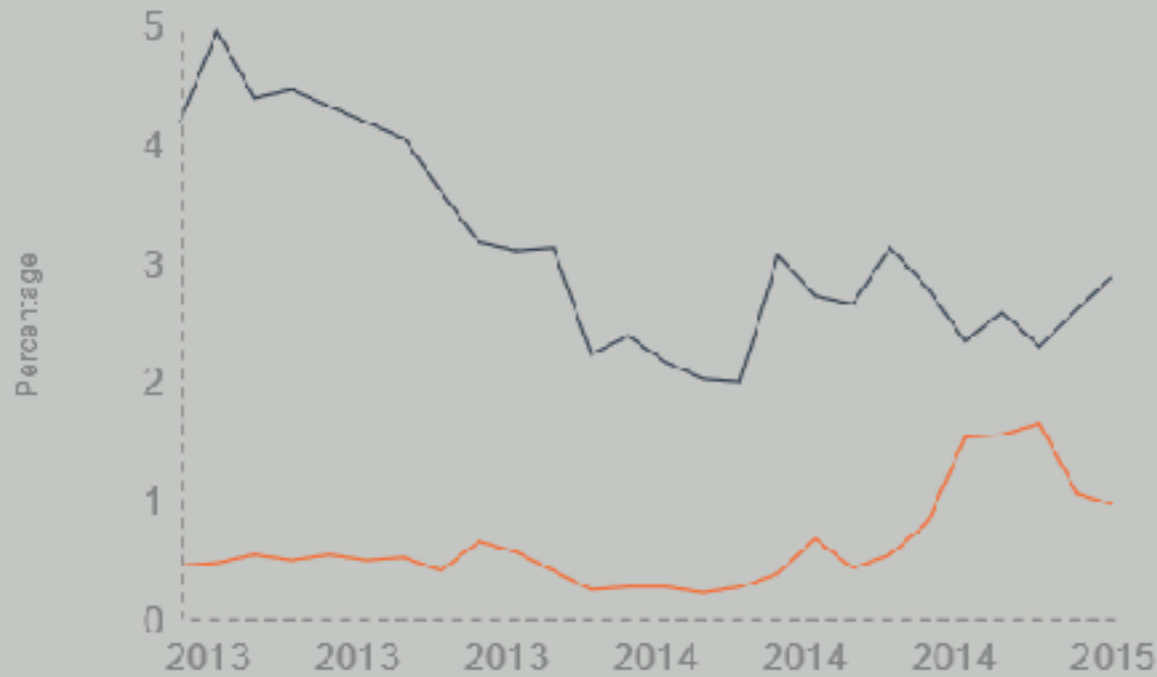
| Language Rank | Types  | Spectrum Ranking |
|---------------|--|------------------|
| 1. C          |    | 100.0            |
| 2. Java       |    | 98.1             |
| 3. Python     |    | 98.0             |
| 4. C++        |    | 95.9             |
| 5. R          |  | 87.9             |
| 6. C#         |  | 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

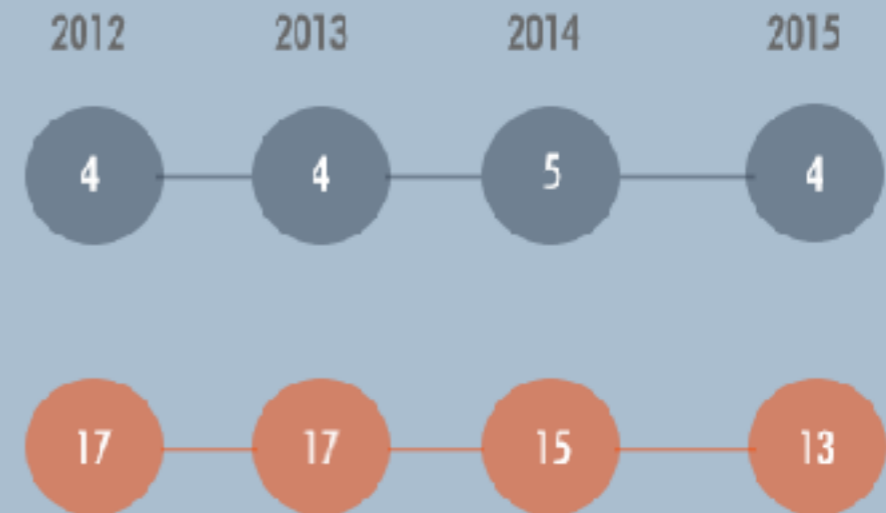
R and Python's popularity between 2013 and February 2015 (TIOBE Index)



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

- 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.



# 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

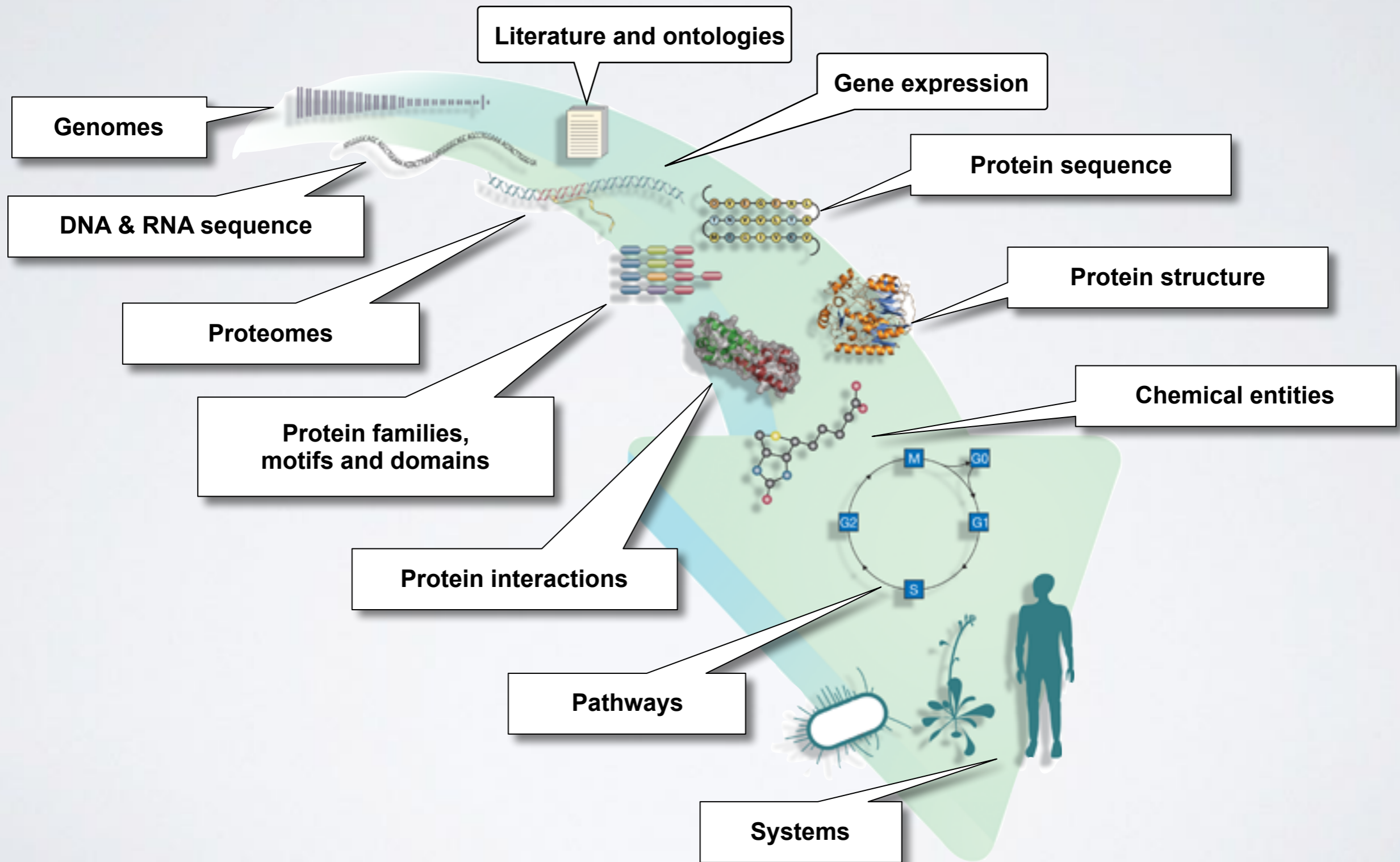
- ▶ “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “**informatics**” techniques (derived from disciplines such as applied maths, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a **large-scale**.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

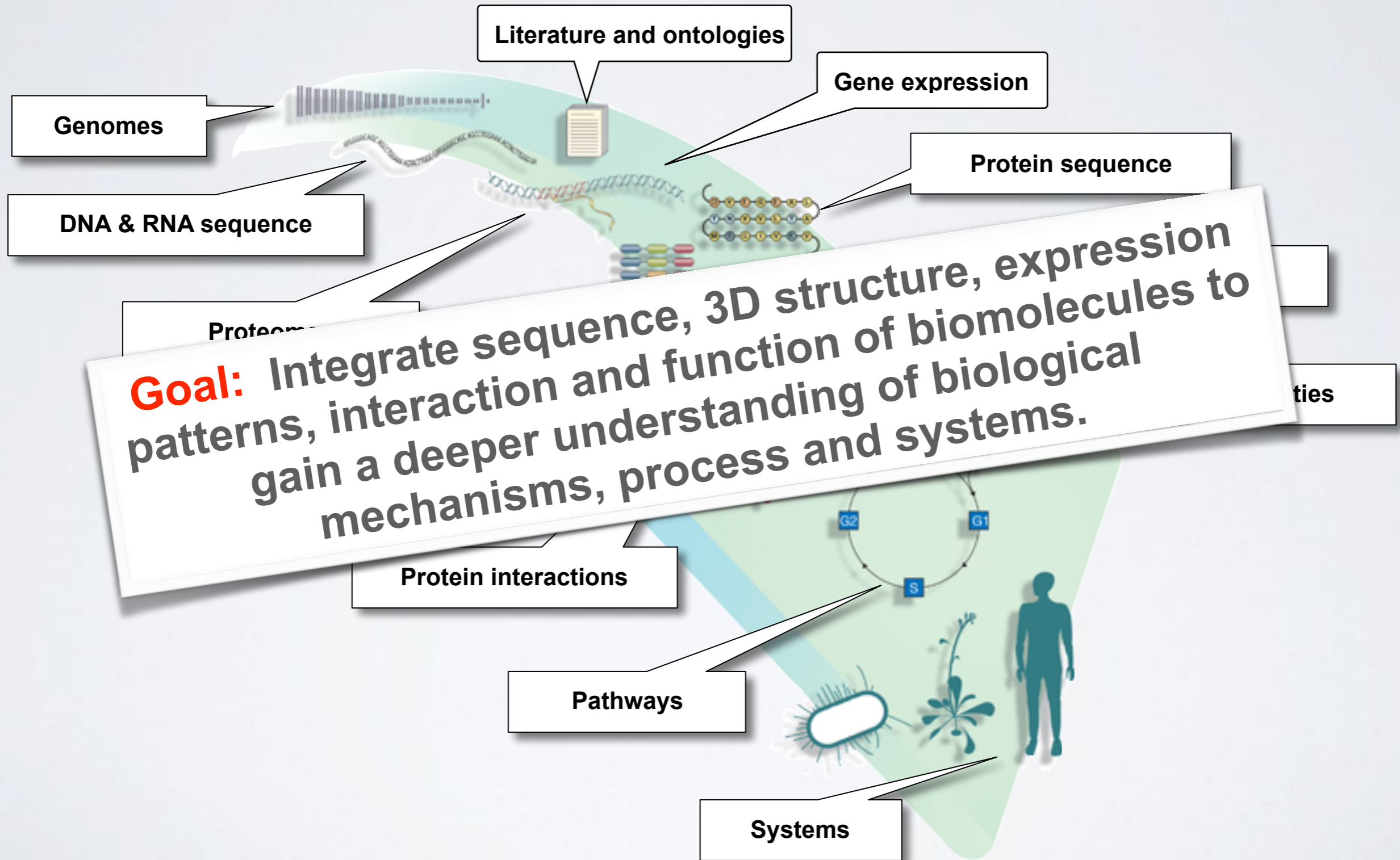
**Key Point: Bioinformatics is Computer Aided Biology**

# Major types of Bioinformatics Data

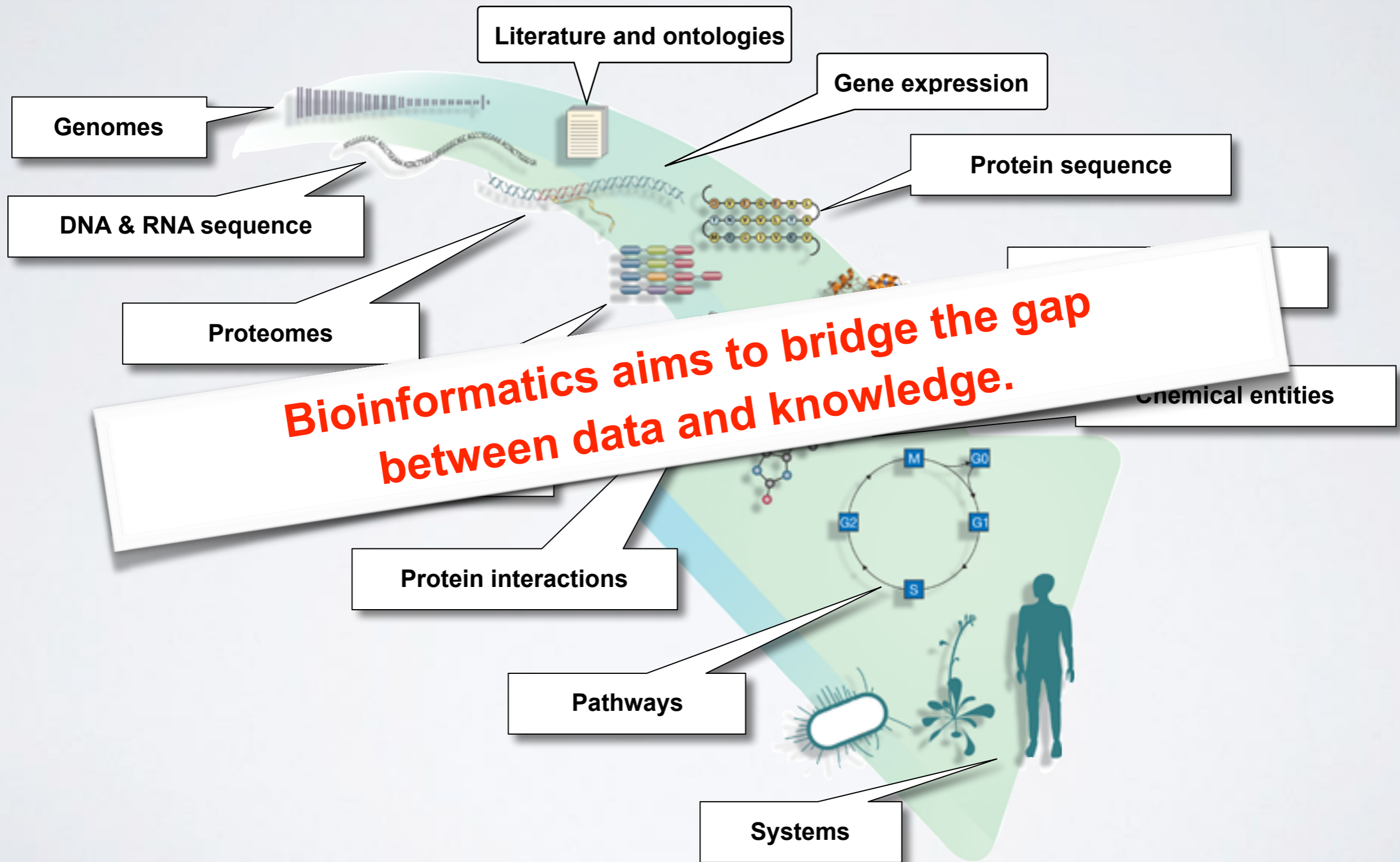




# Major types of Bioinformatics Data



# 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**

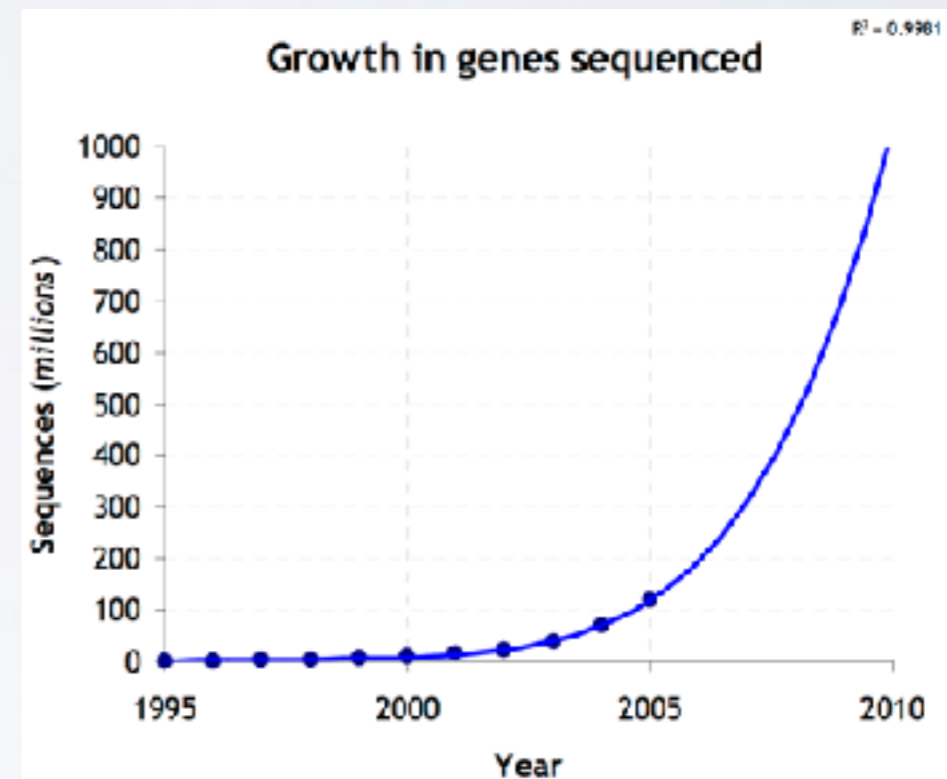
- *DNA sequence determines protein sequence.*
- *Protein sequence determines protein structure.*
- *Protein structure determines protein function.*
- *Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.*

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...*

# 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

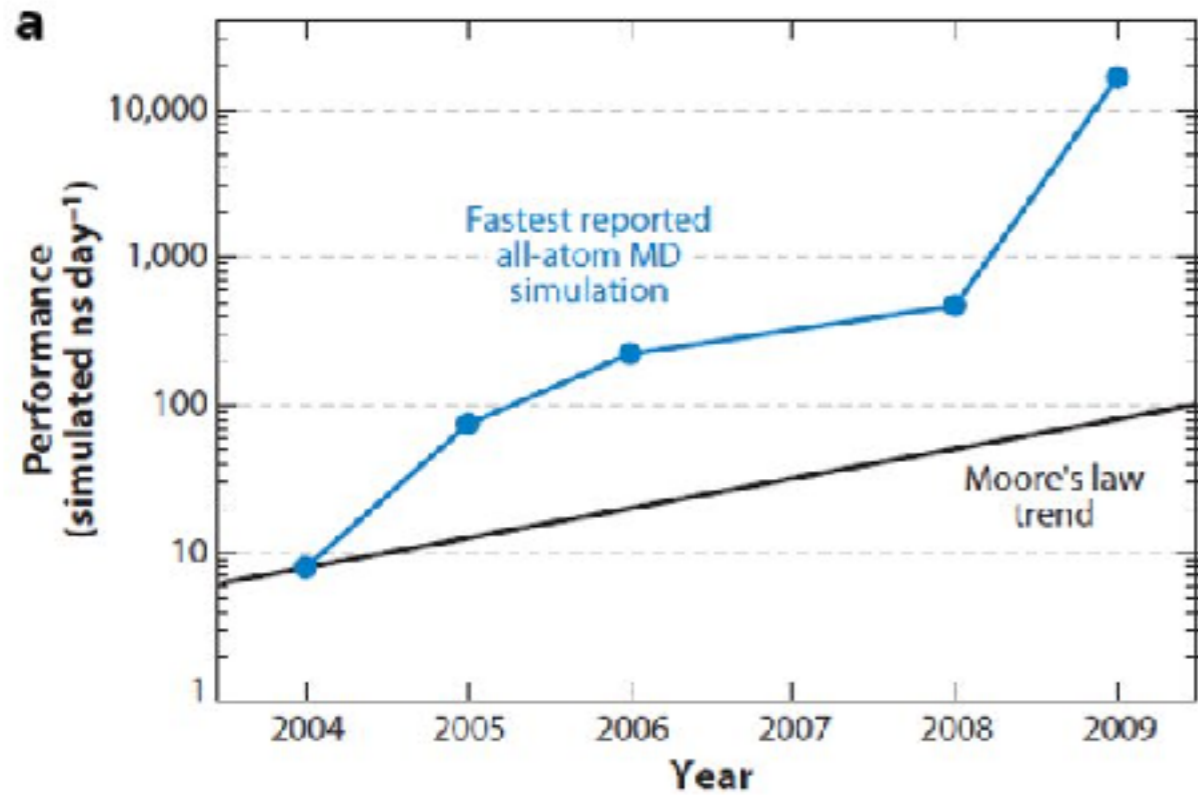
- Many online
- Most are free to use
- Time consuming methods require downloading...

## Advanced tool application & development

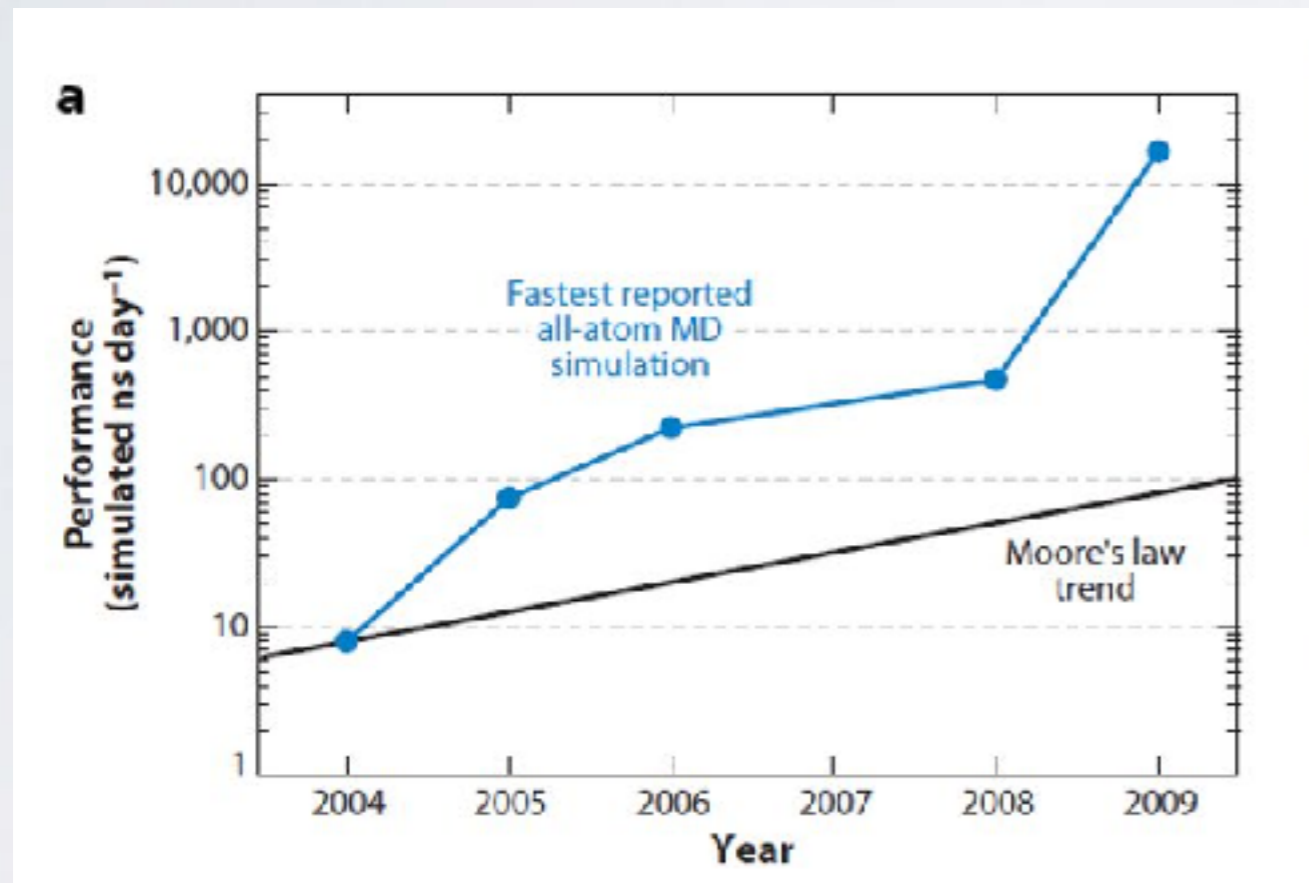
- Mostly on a UNIX environment
- Knowledge of programming languages frequently required (*e.g.* R, Python, Perl, C, Java, Fortran)
- May require specialized high performance computing...



# 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



# 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



The screenshot shows the NCBI website homepage. The URL is www.ncbi.nlm.nih.gov. The page features a navigation menu on the left with categories like 'All Resources', 'Chemicals & Bioassays', 'Data & Software', 'DNA & RNA', 'Domains & Structures', 'Genes & Expression', 'Genetics & Medicine', 'Genomes & Maps', 'Homology', 'Literature', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Variation'. The main content area includes a 'Welcome to NCBI' message, a 'Get Started' section with links to 'Tools', 'Downloads', 'How To's', and 'Submissions', and a 'Popular Resources' list containing 'PubMed', 'Bookshelf', 'PubMed Central', 'PubMed Health', 'BLAST', 'Nucleotide', 'Genome', 'SNP', 'Gene', 'Protein', and 'PubChem'. There is also a '3D Structures' section and 'NCBI Announcements'.

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



The screenshot shows the EBI website homepage. The URL is www.ebi.ac.uk. The page features a navigation menu on the right with categories like 'Services', 'Research', 'Training', 'Industry', 'European Coordination', and 'EMBL Alumni'. The main content area includes a 'The European Bioinformatics Institute' header, a search bar, and a 'Find a gene, protein or chemical' section. There are also sections for 'Popular Resources', 'Visit EMBL.org', and 'Involving events'.

<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 various biological topics. 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 several key services: PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. Red arrows point to PubMed, BLAST, and SNP, while a red bracket groups Nucleotide, Genome, SNP, Gene, and Protein.

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 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>

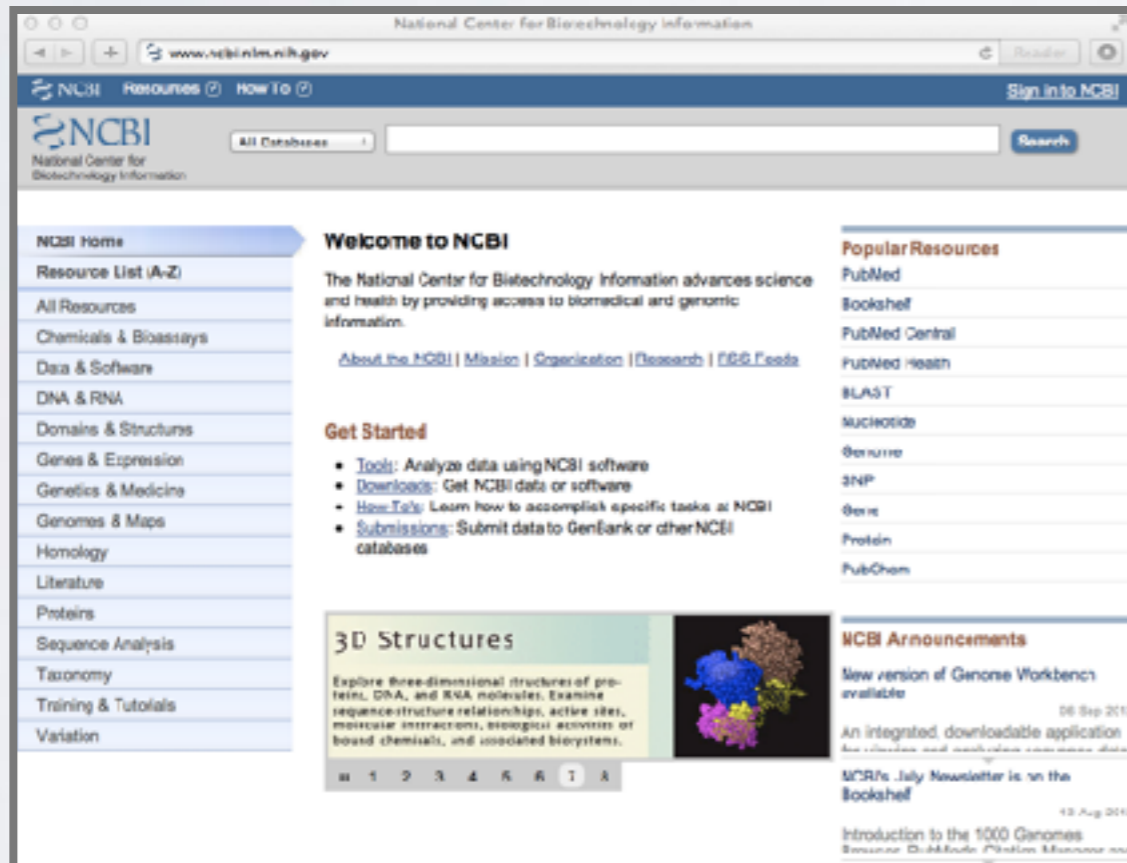
The screenshot shows the NCBI website homepage. At the top, there is a navigation bar with "NCBI", "Resources", and "How To" menus, along with a "Sign in to NCBI" link. Below this is a search bar with a dropdown menu set to "All Databases" and a "Search" button. The main content area features a "Welcome 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 the "databases" section of the NCBI website. On the left is a vertical navigation menu with links for "Homology", "Literature", "Proteins", "Sequence Analysis", "Taxonomy", "Training & Tutorials", and "Variation". The main content area 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 of a protein complex. Further right, there are links for "Protein" and "PubChem". At the bottom right, there is an "NCBI Announcements" section with a headline: "New version of Genome Workbench available" dated "06 Sep" and a sub-headline: "An integrated, downloadable applicati".

# 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 navigation links for 'Services', 'Research', 'Training', and 'About us'. The main title is 'The European Bioinformatics Institute', with the subtitle 'Part of the European Molecular Biology Laboratory'. A descriptive paragraph states: 'EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.' Below this is a search bar with the prompt 'Find a gene, protein or chemical:' and a 'Search' button. A row of six colored tiles 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 that is a 'Visit EMBL.org' section with the EMBL 40th anniversary logo (1974-2014). The 'Upcoming events' section features a banner for the 'Plant and Animal Genome conference (PAG XXIV)' on Sunday 10 - Tuesday 12 January 2016. The bottom of the page shows a '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](http://www.ebi.ac.uk/services). The page features a teal header with the EMBL-EBI logo and navigation links for Services, Research, Training, and About us. Below the header, a secondary navigation bar includes Overview, A to Z, Data submission, and Support. The main content area is titled "Bioinformatics services" and contains a paragraph describing the website's offerings. To the right, a "Popular" section lists various tools and databases. The main content is organized into a grid of service categories, and a "Service news" section features a butterfly image. At the bottom right, a "Training" section is visible.

Services < EMBL-EBI

www.ebi.ac.uk/services

Services Research Training About us

# Services

Overview A to Z Data submission Support

## Bioinformatics services

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

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

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

**Proteins**  
sequences, families & motifs

**Structures**  
Molecular & cellular structures

**Systems**  
reactions, interactions & pathways

**Chemical biology**  
chemogenomics & metabolomics

**Ontologies**  
taxonomies & controlled vocabularies

**Literature**  
Scientific publications & patents

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

Popular

- Ensembl
- UniProt
- 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 and tools. A grid of service categories includes DNA & RNA, Gene expression, Proteins, 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

## Proteins

### Popular services

|   |  |
|---|--|
|    | <b>UniProt: The Universal Protein Resource</b><br>The gold-standard, comprehensive resource for protein sequence and functional annotation data.   |
|    | <b>InterPro</b><br>A database for the classification of proteins into families, domains and conserved sites.   |
|  | <b>PRIDE: The Proteomics Identifications Database</b><br>An archive of protein expression data determined by mass spectrometry.  |
|  | <b>Pfam</b><br>A database of hidden Markov models and alignments to describe conserved protein families and domains.   |
|  | <b>Clustal Omega</b><br>Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.  |
|  | <b>HMMER - protein homology search</b><br>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><br>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 40 YEARS 1974-2014

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 his email address 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 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

0:05 / 37:42

Popular

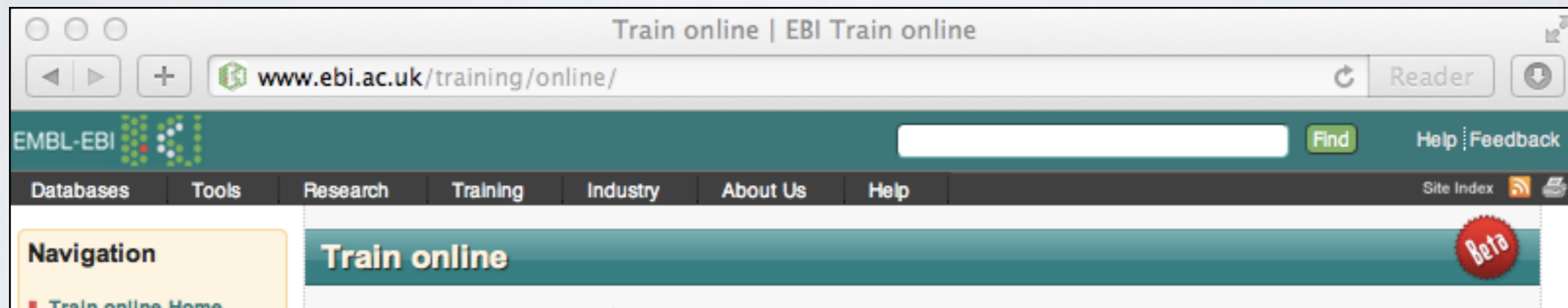
- Train online
- Find us
- Funding

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

The screenshot shows a web browser window with the address bar displaying `bioboot.github.io/bimm143_W18/lectures/#1`. 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** (highlighted with a red box)
- 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](#) [↗](#).
- Setup your [laptop computer](#) for this course.

**Material:**

- Lecture Slides: [Large PDF](#) [↗](#), [Small PDF](#) [↗](#),
- Lab: [Hands-on section worksheet](#) [↗](#) (highlighted with a red box)
- Feedback: [Muddy Point Assessment](#) [↗](#).
- 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

```
ATGGTGCATCTGACTCCTGTGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG
TTGGTGGTGAAGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGG
GGATCTGTCCACTCCTGATGCAGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCA
TCACCTTGGCAAAGAATTCACCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCCCAAGTATCACTAAGCTCGCTTCTTGCTGTCCAATTT
```

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

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

THANK YOU

The text "THANK YOU" is displayed in a large, bold, sans-serif font. Each letter is a different color: 'T' is green, 'H' is blue, 'A' is black, 'N' is pink, 'K' is blue, 'Y' is green, 'O' is black, and 'U' is black. Below each letter is a small number from 1 to 8, corresponding to the letter's position. The letters are slightly offset from each other, creating a layered effect.