



BIMM 143

Introduction to Bioinformatics

Barry Grant
UC San Diego

<http://thegrantlab.org/bimm143>

HELLO
my name is

BARRY

bjgrant@ucsd.edu

HELLO
my name is

HER -

ALENA

amartsul@ucsd.edu

HELLO
my name is

HER -

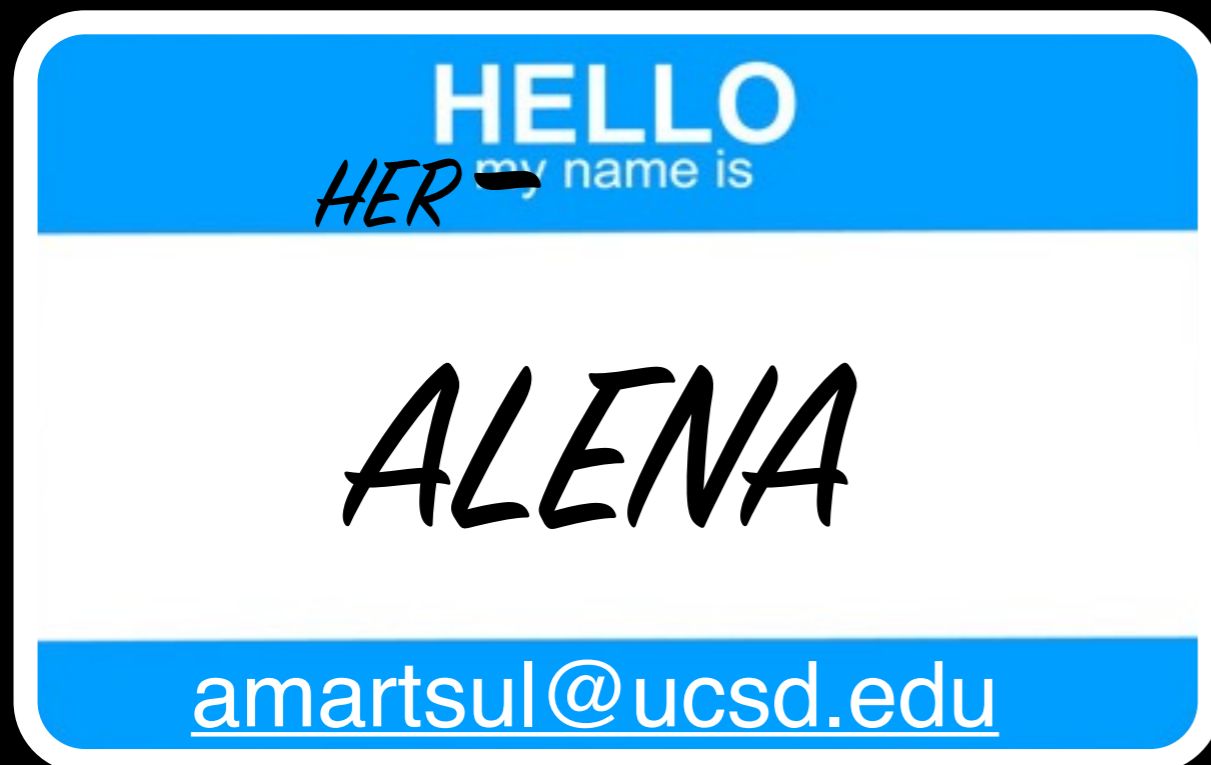
KELLY

kflander@ucsd.edu



Office Hours:
[SignUp](#)

Location:
TATA, #2501



Introduce Yourself!

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

Today's Menu

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

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

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

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

Bioinformatics (BIMM 143, Fall 2018)

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.

http://thegrantlab.org/bimm143/

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

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

Bioinformatics (BIMM 143, Fall 2018)

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

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

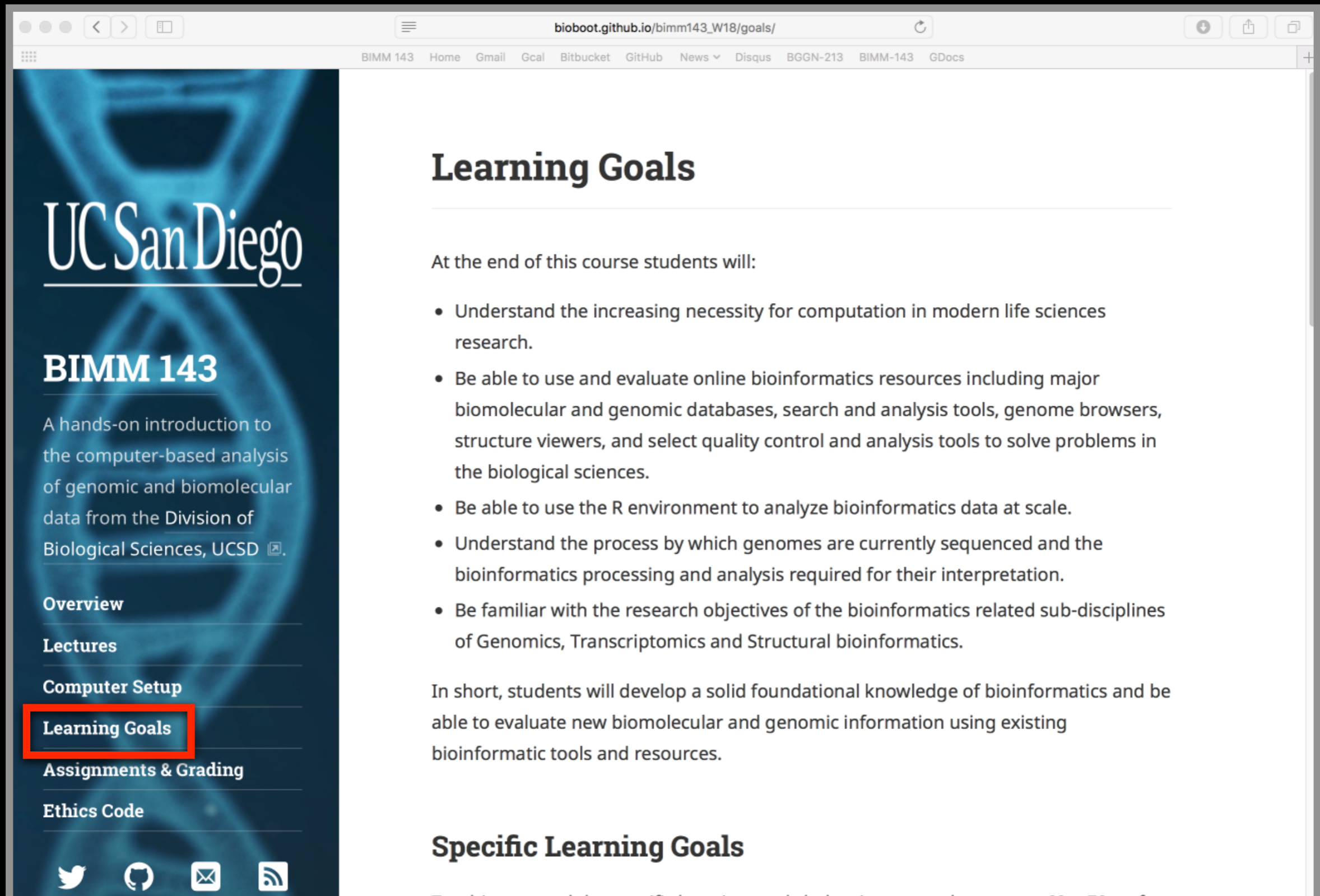
Course Syllabus
[Fall 2018 \(PDF\)](#)

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.

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



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

Twitter GitHub Email RSS

bioboot.github.io/bimm143_W18/goals/

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

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!

The screenshot shows a web browser window with the URL `bioboot.github.io/bimm143_W18/goals/`. The browser's address bar and navigation buttons are visible at the top. Below the browser window, the page content is displayed. On the left side, there is a dark blue sidebar with the UC San Diego logo and a list of navigation links: **BIMM 143**, Overview, Lectures, Computer Setup, **Learning Goals** (highlighted with a red box), Assignments & Grading, and Ethics Code. The main content area has the heading **Specific Learning Goals** and a paragraph explaining that 60%-70% of class time is dedicated to these goals. Below this is a table with five rows, each representing a learning goal and the lectures it covers. A red arrow on the right side of the page points downwards.

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 browser's address bar and navigation buttons are visible at the top. Below the browser window, the course structure page is displayed. On the left side, there is a navigation menu for the course, with the 'Lectures' item highlighted in a red box. The main content area features a 'Lectures' section with a heading and a paragraph of text. Below this text is a table with three columns: '#', 'Date', and 'Topics for Spring 2018'. The table lists two lectures, with the first lecture on 04/03 and the second on 04/05. The table is partially cut off at the bottom.

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

Sidebar (Left):

- UC San Diego logo
- 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

Main Content Area (Right):

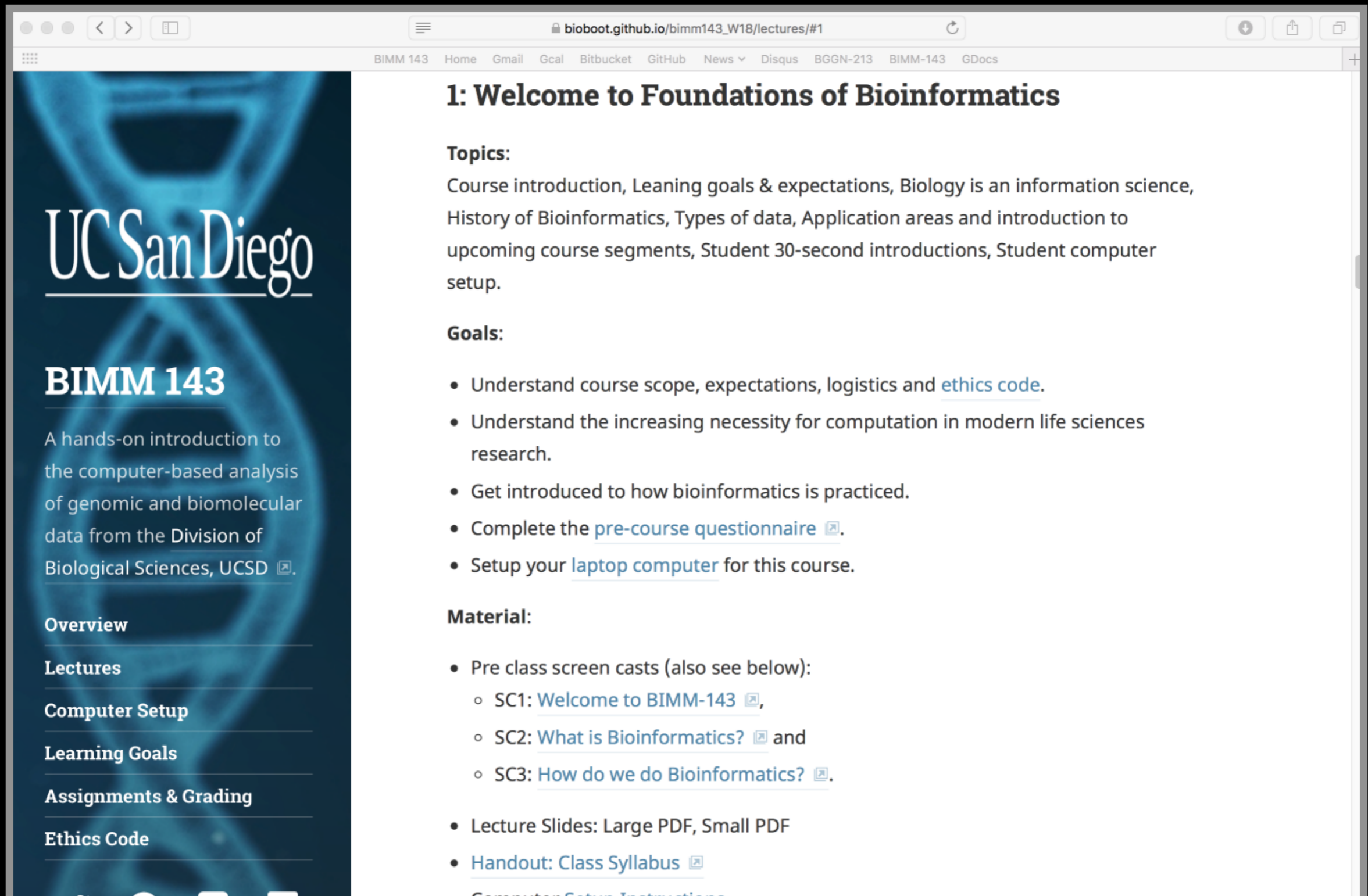
Lectures

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 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 browser's address bar and navigation buttons are visible at the top. Below the browser window, the page content is displayed. On the left side, there is a dark blue sidebar with the UC San Diego logo and the course title **BIMM 143**. The sidebar also contains a list of navigation links: Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. The main content area on the right is white and features the heading **1: Welcome to Foundations of Bioinformatics**. Below this heading, there are sections for **Topics:**, **Goals:**, and **Material:**. The **Topics:** section describes the course introduction, learning goals, and history of bioinformatics. The **Goals:** section lists five bullet points, including understanding course scope, the necessity for computation, and completing a pre-course questionnaire. The **Material:** section lists pre-class screen casts (SC1, SC2, SC3), lecture slides, and a handout.

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

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:

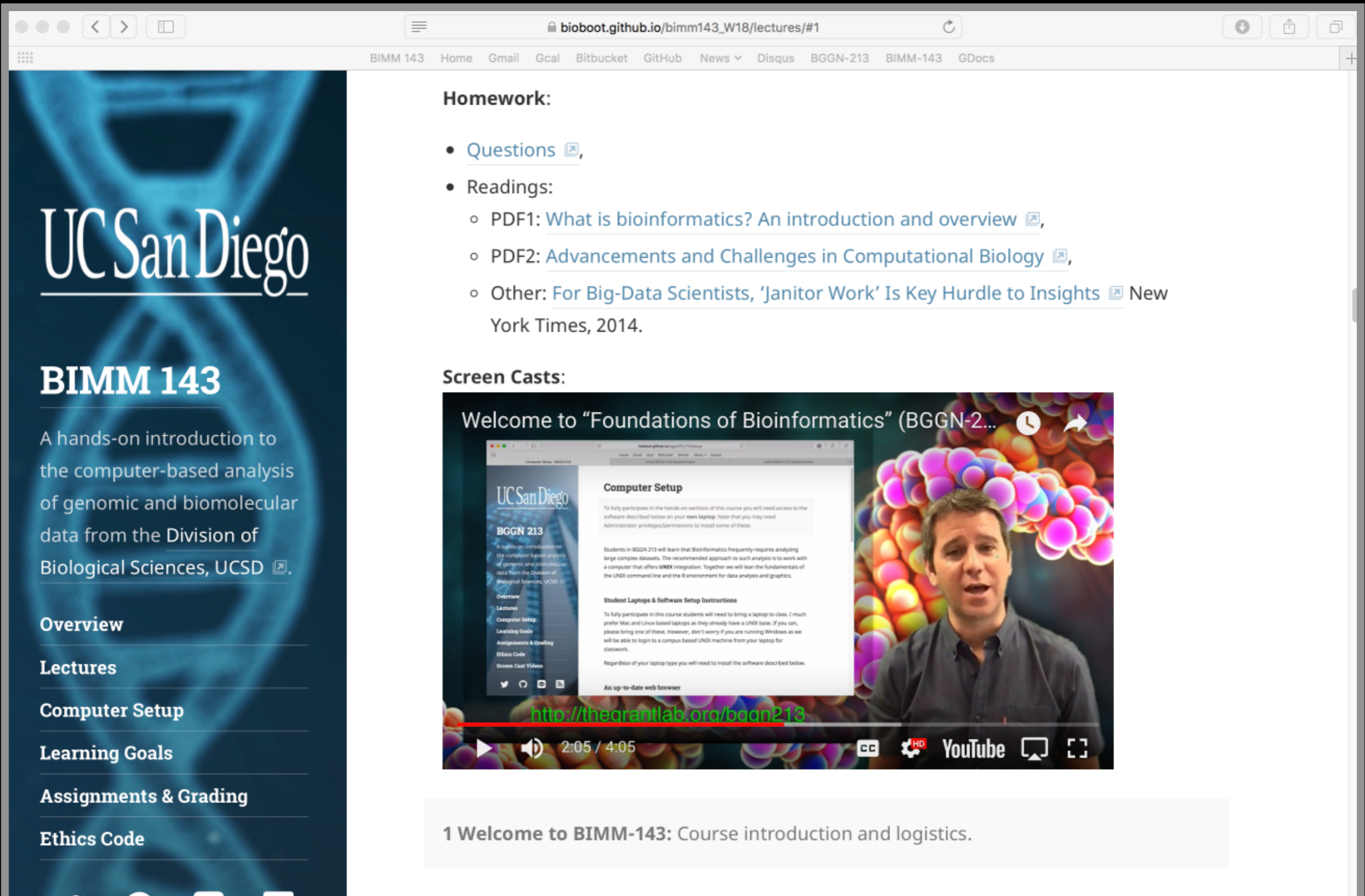
- Understand course scope, expectations, logistics and [ethics code](#).
- Understand the increasing necessity for computation in modern life sciences research.
- Get introduced to how bioinformatics is practiced.
- Complete the [pre-course questionnaire](#) [↗](#).
- Setup your [laptop computer](#)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

Homework

Goals, Class material, Screencasts & Homework



The screenshot shows a web browser window with the URL bioboot.github.io/bimm143_W18/lectures/#1. The browser tabs include BIMM 143, Home, Gmail, Gcal, 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://theorantlab.org/bggan213>

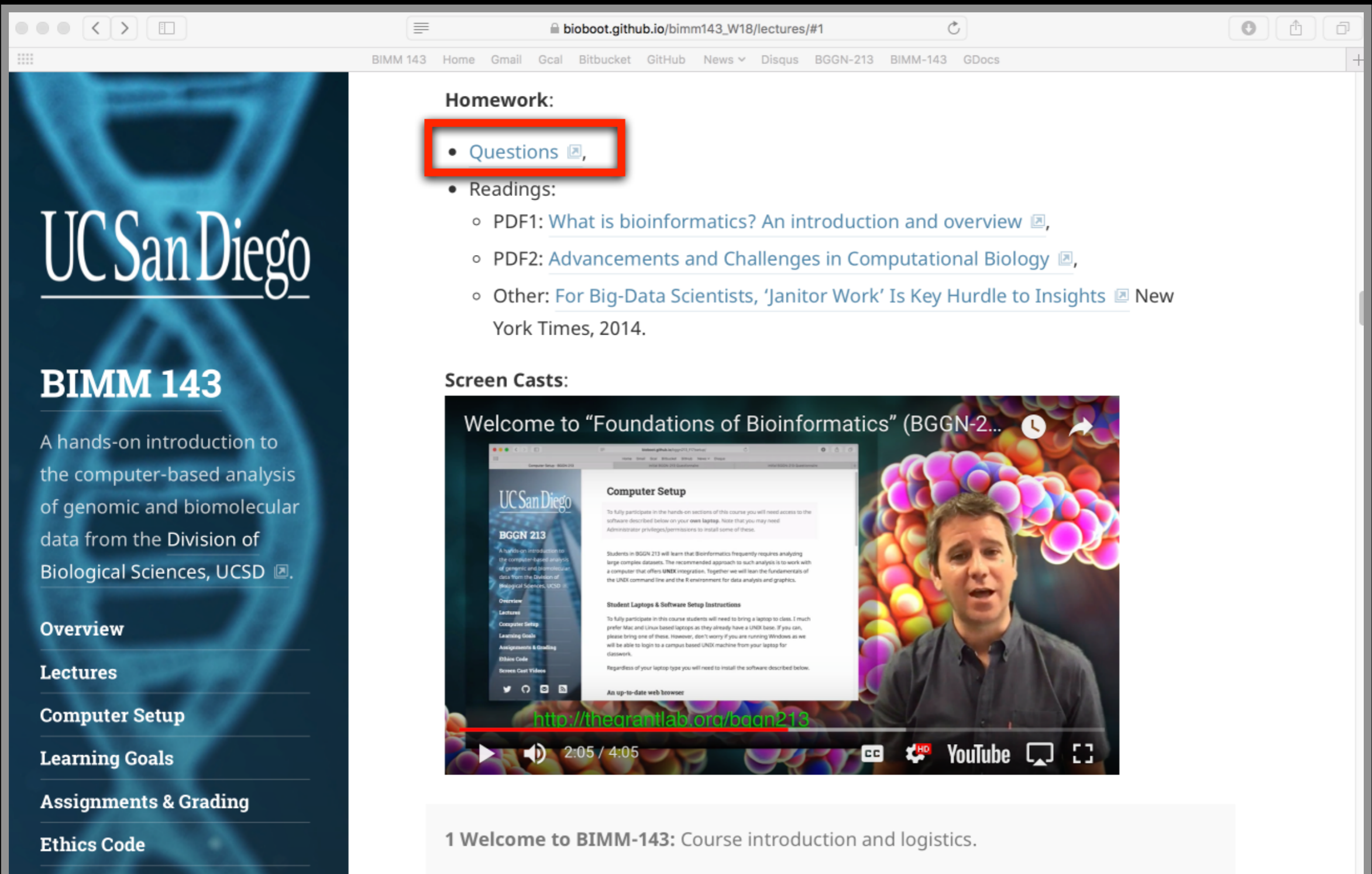
2:05 / 4:05

YouTube

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 is as follows:

- UC San Diego** logo and **BIMM 143** title.
- Text: "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, Assignments & Grading, Ethics Code.
- 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:**
 - Video player: "Welcome to 'Foundations of Bioinformatics' (BGGN-2...)"
 - Video content: A man speaking in front of a molecular model background. A browser window in the background shows the "Computer Setup" page for BGGN 213.
 - Video URL: <http://theorantlab.org/bqon213>
 - Video player controls: 2:05 / 4:05, YouTube logo, and other controls.
- Footer: 1 Welcome to BIMM-143: Course introduction and logistics.

Homework

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

BIMM143 Lecture 1 Homework (W19)

Please answer the following questions including your main [@ucsd.edu](mailto: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 (35% of course grade)

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

BIMM143 Lecture 1 Homework

Please answer the following questions. Please provide your email address and UCSD PID number so you can receive your grade.

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

Projects

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

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

Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News [+](#)

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](#) [↗](#).

Twitter GitHub Email RSS

Projects

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

The image shows two overlapping browser windows. The background window displays the UC San Diego BIMM 143 course page, which includes the university logo and a navigation menu with links for Overview, Lectures, Computer Setup, Learning Goals, and Ethics. The foreground window shows a specific lecture page titled "Designing a personalized cancer vaccine".

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

Designing a personalized cancer vaccine

BIMM-143 Lecture 18:
Barry Grant < <http://thegrantlab.org> >
Date: 2018-03-07 (15:24:21 PST on Wed, Mar 07)

Notes: To identify somatic mutations in a tumor, DNA from the tumor is sequenced and compared to DNA from normal tissue in the same individual using *variant calling algorithms*.

Comparison of tumor sequences to those from normal tissue (rather than 'the human genome') is important to ensure that the detected differences are not germline mutations.

To identify which of the somatic mutations leads to the production of aberrant proteins, the location of the mutation in the genome is inspected to identify non-

Projects (20% of course grade)

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

The image displays three overlapping browser windows from the website `bioboot.github.io`. The top window shows a navigation menu with links for Home, Gmail, Gcal, GitHub, BIMM143, BGGN213, Atmosphere, BIMM194, Blink, and News. The middle window shows the course page for BIMM 143, featuring the UC San Diego logo and a sidebar menu with links for Overview, Lectures, Computer Setup, and Learning. The bottom window shows a lecture page titled "10: (Project:) Find a Gene Assignment Part 1".

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.

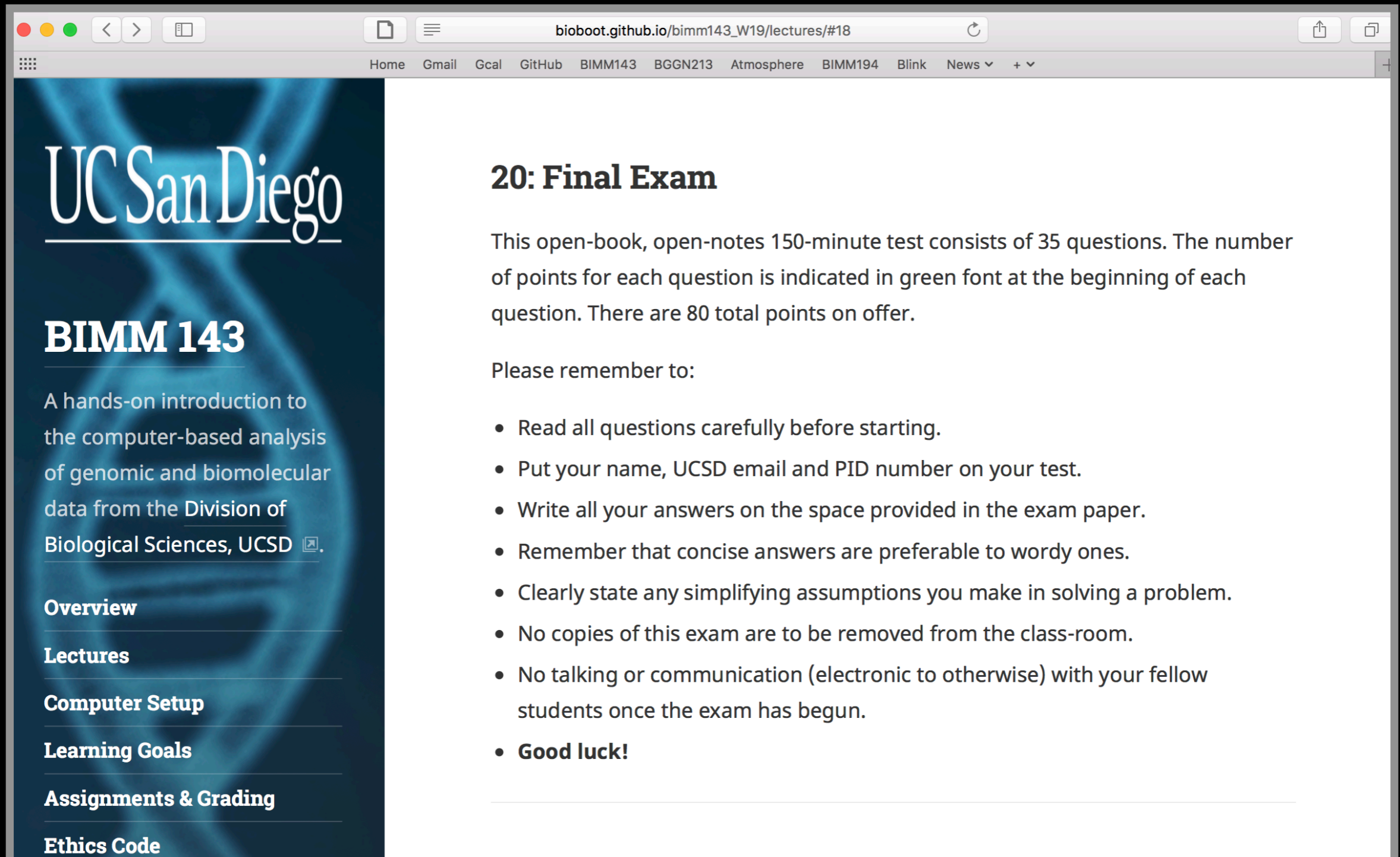
Your responses to questions Q1-Q4 are due at the beginning of class **Thursday Nov 15th** (11/15/18).

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.

Final Exam

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



The screenshot shows a web browser window with the address bar displaying `bioboot.github.io/bimm143_W19/lectures/#18`. The browser's tab bar includes links for Home, Gmail, Gcal, GitHub, BIMM143, BGGN213, Atmosphere, BIMM194, Blink, News, and a plus sign. The main content area is divided into a left sidebar and a main text area. The sidebar, on a dark blue background, features the UC San Diego logo and a list of navigation links: **BIMM 143**, Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. The main text area has a white background and contains the following content:

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!**

Bonus:

Bioinformatics & Genomics in industry

bioboot.github.io/bimm143_W19/lectures/#21

Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News +





Lectures · BIMM 143 eGrade-BIMM143_W19 - Google Sheets

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

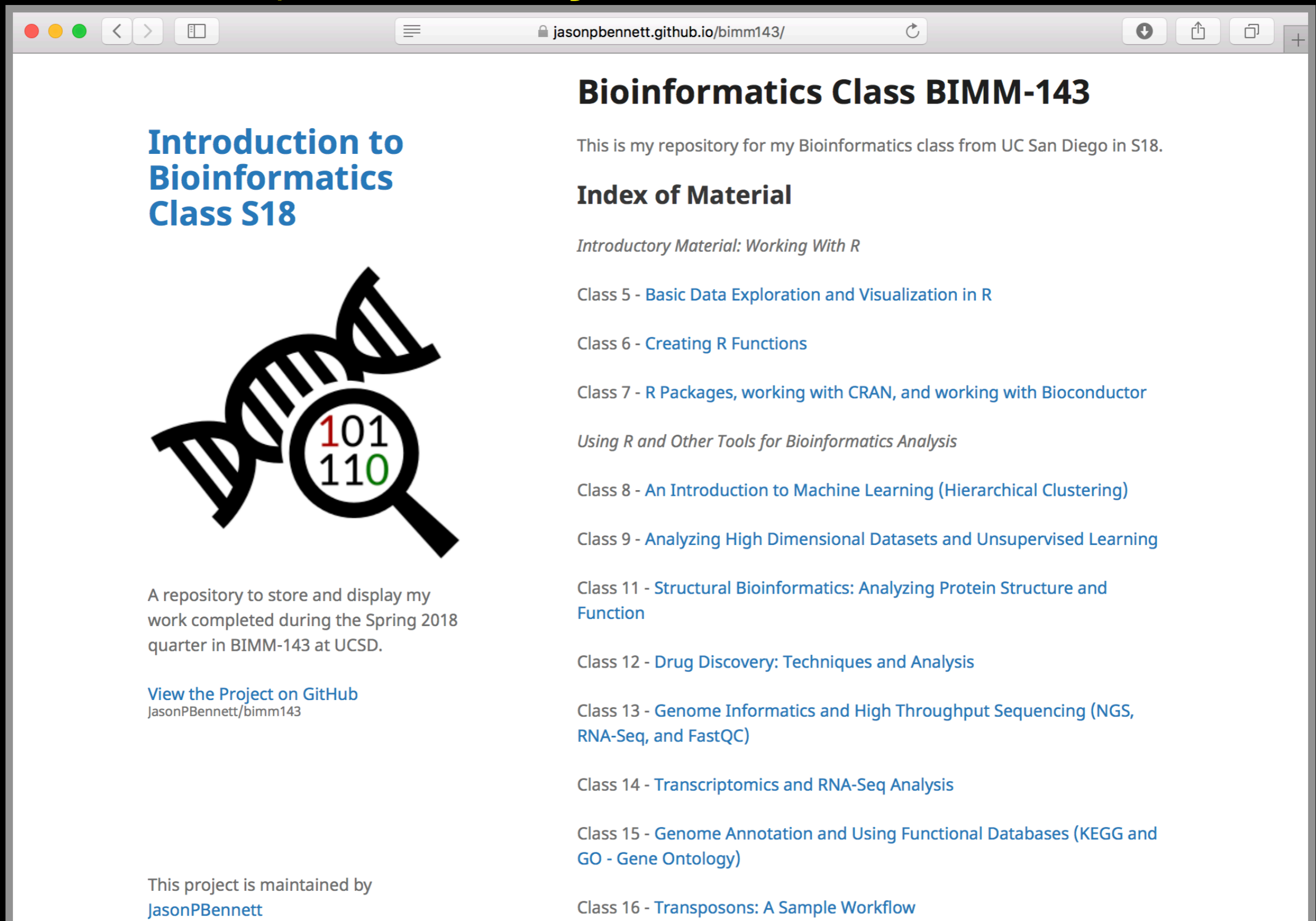
21: Bonus: Bioinformatics & Genomics in industry

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 address bar containing 'jasonpbennett.github.io/bimm143/'. The page is titled 'Bioinformatics Class BIMM-143' and features a navigation menu on the right side. The main content area on the left includes a title 'Introduction to Bioinformatics Class S18', a DNA double helix icon with a magnifying glass over the binary code '101110', and a description of the repository. A link to 'View the Project on GitHub' is provided, along with the repository name 'JasonPBennett/bimm143'. At the bottom, it states 'This project is maintained by JasonPBennett'.

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

This project is maintained by
[JasonPBennett](#)

Bioinformatics Class BIMM-143

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

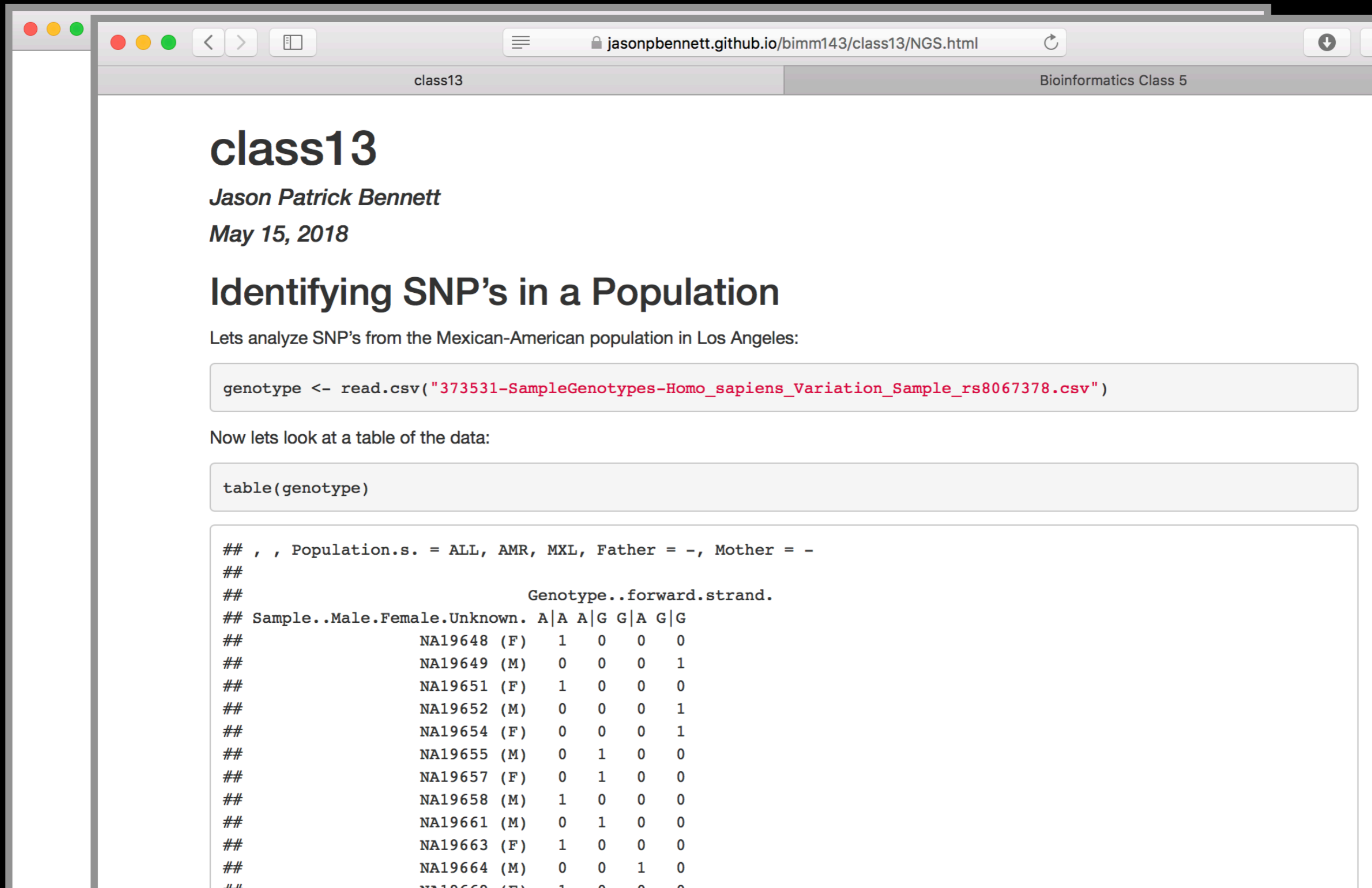
Index of Material

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](#)

Bonus:

Online portfolio of **your** bioinformatics work!



The screenshot shows a web browser window with the address bar containing the URL `https://jasonpbennett.github.io/bimm143/class13/NGS.html`. The page title is "class13" and the browser tab is labeled "Bioinformatics Class 5".

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")
```

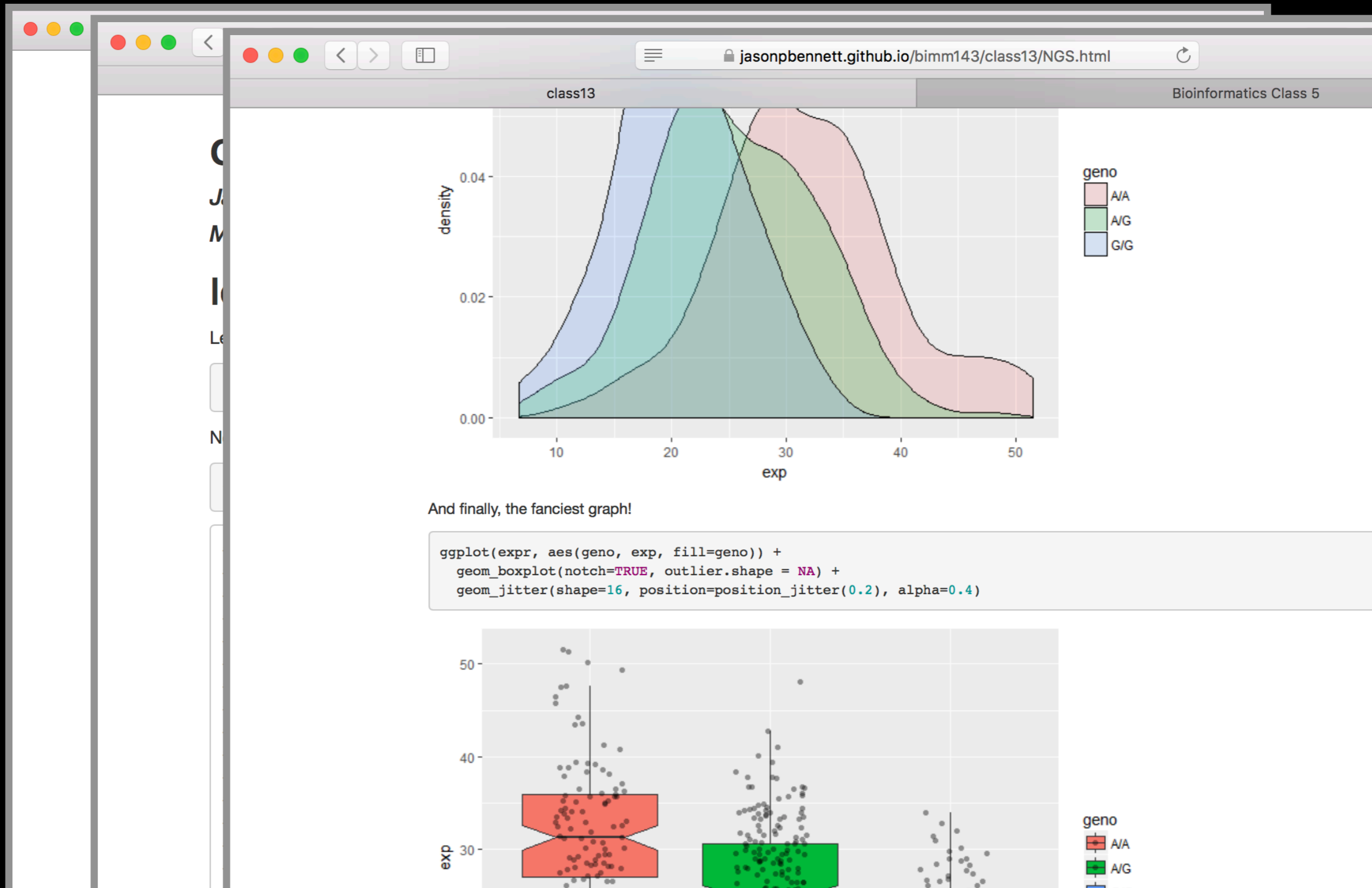
Now lets look at a table of the data:

```
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  
##                               NA19666 (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

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

The screenshot shows a web browser window displaying the BIMM-143 Learning Goals page. The page is titled "BIMM 143" and is part of the UC San Diego course. The left sidebar contains navigation links: Overview, Lectures, Computer Setup, Learning Goals (highlighted with a red box), Assignments & Grading, and Ethics Code. The main content area is a table of learning goals, with a green box highlighting goals 12 through 17. A red arrow points to the bottom right corner of the page.

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.



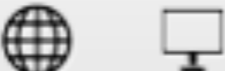





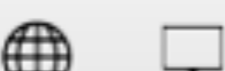

Why use R?

Productivity

Flexibility

Genomic data analysis

IEEE 2016 Top Programming Languages

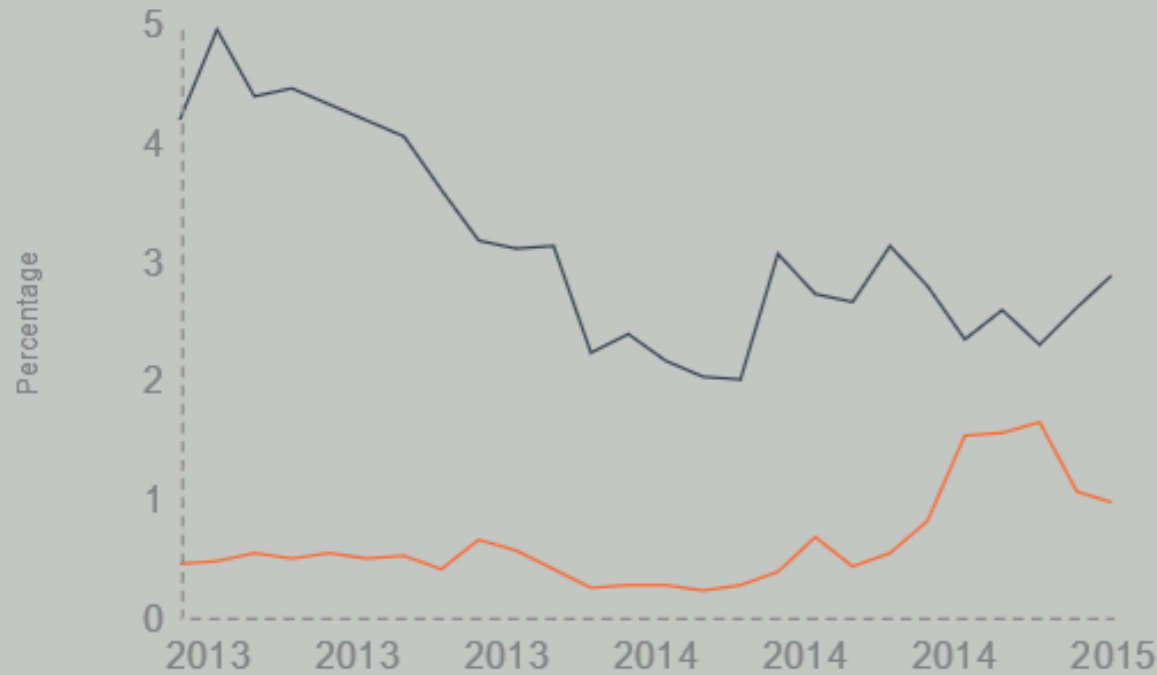
Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

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

R and Python: The Numbers

Popularity Rankings

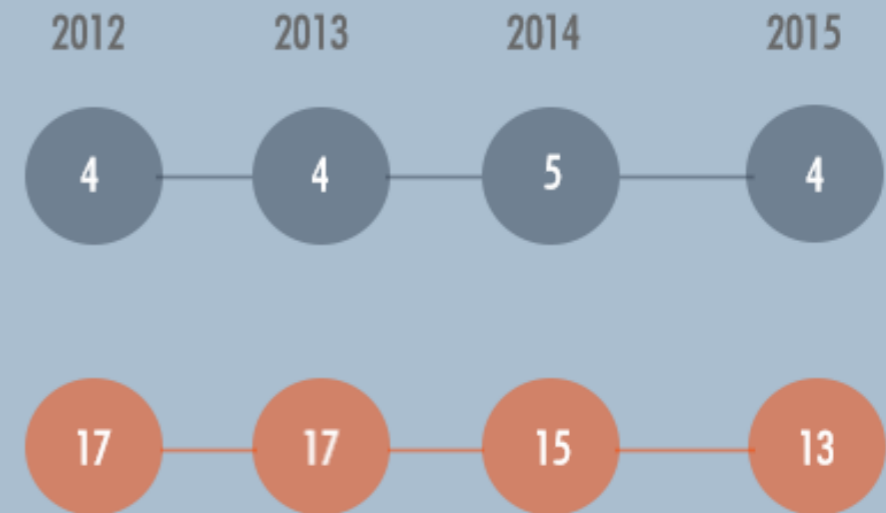
R and Python's popularity between 2013 and February 2015 (Tiobe Index)



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 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...

The screenshot shows a web browser window with the URL etherpad.net/p/bimm143_f18. The browser's address bar and tabs are visible. The Etherpad interface includes a top navigation bar with links to Home, Gmail, Gcal, GitHub, BIMM143, BGGN213, Atmosphere, BIMM194, Blink, and News. Below this is a toolbar with various editing tools such as bold (B), italic (I), underline (U), strikethrough (ABC), bulleted list, numbered list, indent, outdent, undo, redo, eye (visibility), style dropdown, table, link, unlink, settings, and code editor. The main editing area contains a list of responses to a question: "Q1. Did you enjoy this course in relation to others you have experienced at UCSD?". The responses are: "Hell Yeah!", "Yes", "it was too lit", "Yes!", "Yes!", "yes", "yes!", "I do too!", "One of the best", "The best", "yes", "Ye", "Yes", "yes", "Yes", "Yes", "yes!", "yes, one of the most useful classes I've had", and "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. :)". A chat window is visible in the bottom right corner with the text "Chat 0".

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...

7
8 **Q1. Did yo**
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 be
18 The best
19 yes
20 Ye
21 Yes
22 yes
23 Yes
24 Yes
25 yes!
26 yes, one of th
27 no but im jus
28

etherpad.net/p/bimm143_S18
Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News +
bimm143 S18 | etherpad.net Pad eGrade-BIMM143_W19 - Google Sheets

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

7
8 **Q1. Did you**
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 be
18 The best
19 yes
20 Ye
21 Yes
22 yes
23 Yes
24 Yes
25 yes!
26 yes, one of th
27 no but im jus
28

etherpad.net/p/bimm143_S18
Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink News +
bimm143 S18 | etherpad.net Pad eGrade-BIMM143_W19 - Google Sheets

8 **Q1. Did you**
9 Yes
10 - Yes.
11 Yes
12 Yes
13 yes, quite.
14 yes
15 - I enjoyed this
16 This is the best
17 Yes
18 Yes this course
19 Yes even as a b
20 Yes this course
21 This is one of th
22 Yes
23 Yes. I very muc
24 yes
25 Yes!
26 I enjoyed this c
27 This is one of th
28 Yes
29 yes, it was a ve

etherpad.net/p/bgggn213_S18
Home Gmail Gcal GitHub BIMM143 BGGN213 Atmosphere BIMM194 Blink New
bggn213 S18 | etherpad.net Pad eG

7
8 **Q1. Did you enjoy this course in relation to others you have experience**
9 Yes, very much
10 Yes, absolutely!
11 Yes
12 Yes, I like the focus on applying R to real world biological datasets
13 Yes
14 yes
15 Yes
16 It was a lot harder than I was expecting
17 yes
18 Yes!
19 yes
20 Yes!
21 yes
22 Yes, I learned lots of things that are very useful in reserach but hard to learn ourselves
23 Yes this class was awesome!
24 Yes, this course was amazingly put together in a logical way and was extremely thorough.
25

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 leaning goals.
Introduction to Bioinformatics	Introducing the <i>what, why</i> and <i>how</i> 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

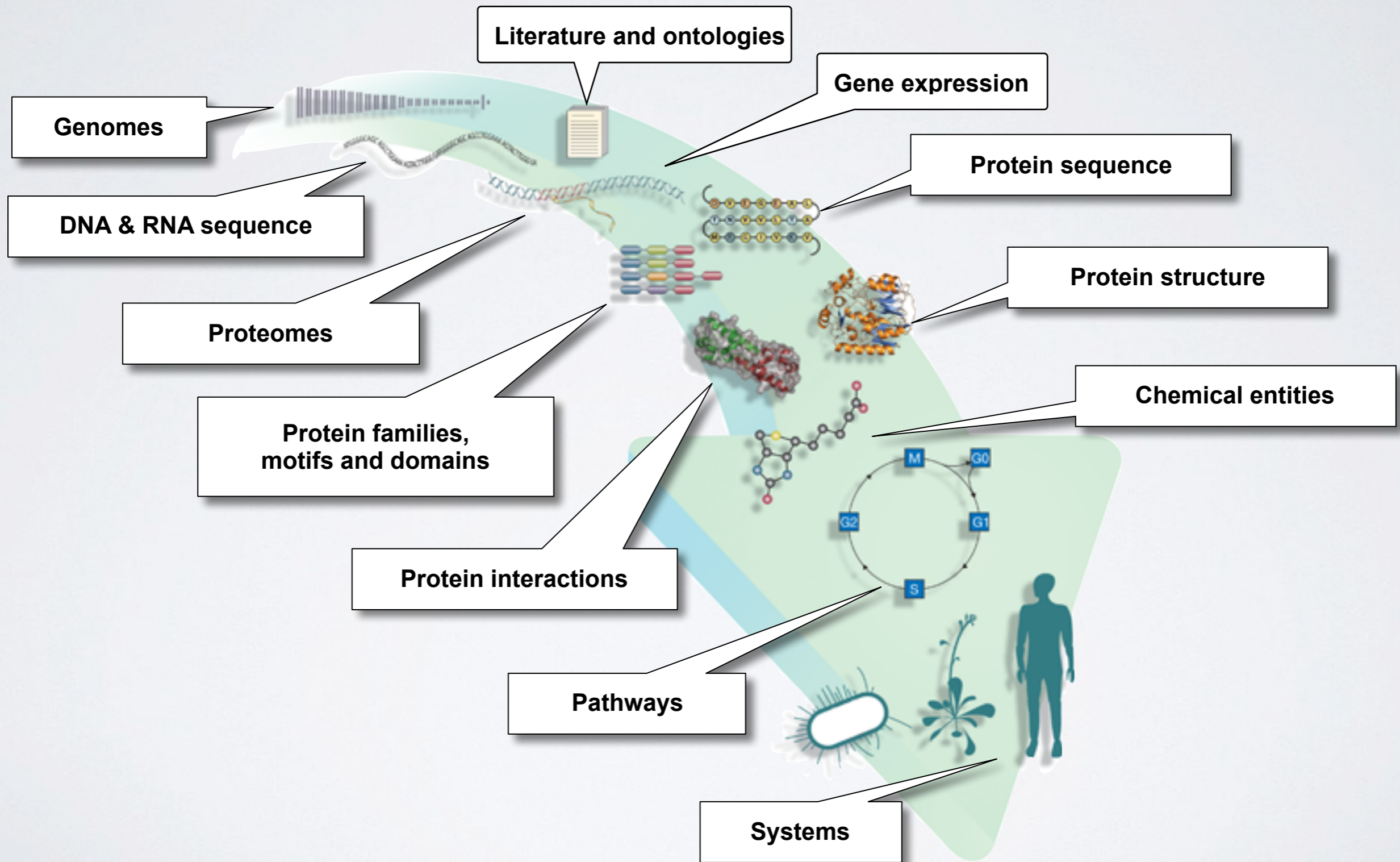
- ▶ “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “informatics” techniques (derived from disciplines such as applied maths, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a **large-scale**.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

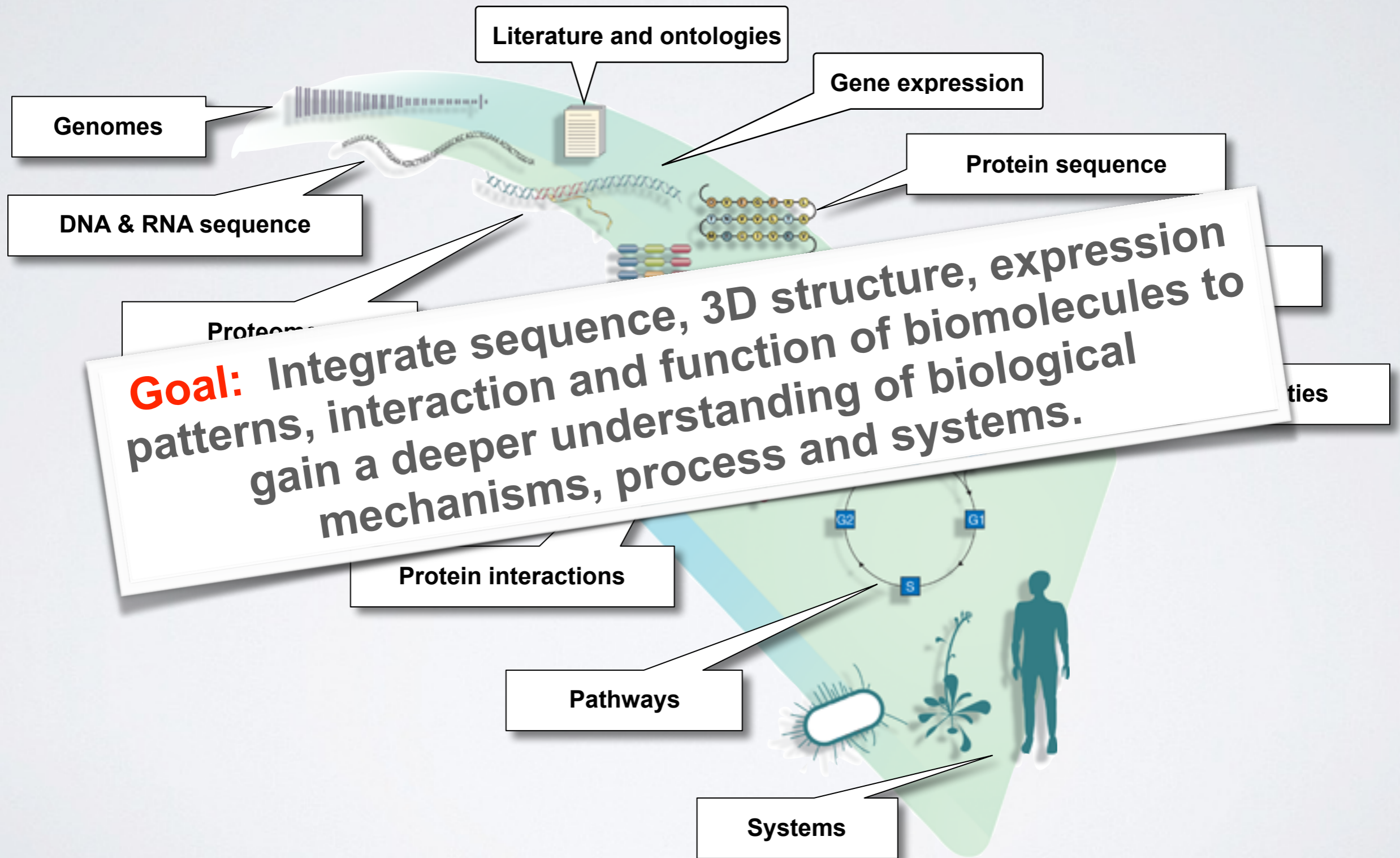
- ▶ “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “informatics” techniques (derived from disciplines such as applied mathematics, computer science, and statistics) to **understand and analyze** the information associated with these systems, on a **large-scale**.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

Key Point: Bioinformatics is Computer Aided Biology

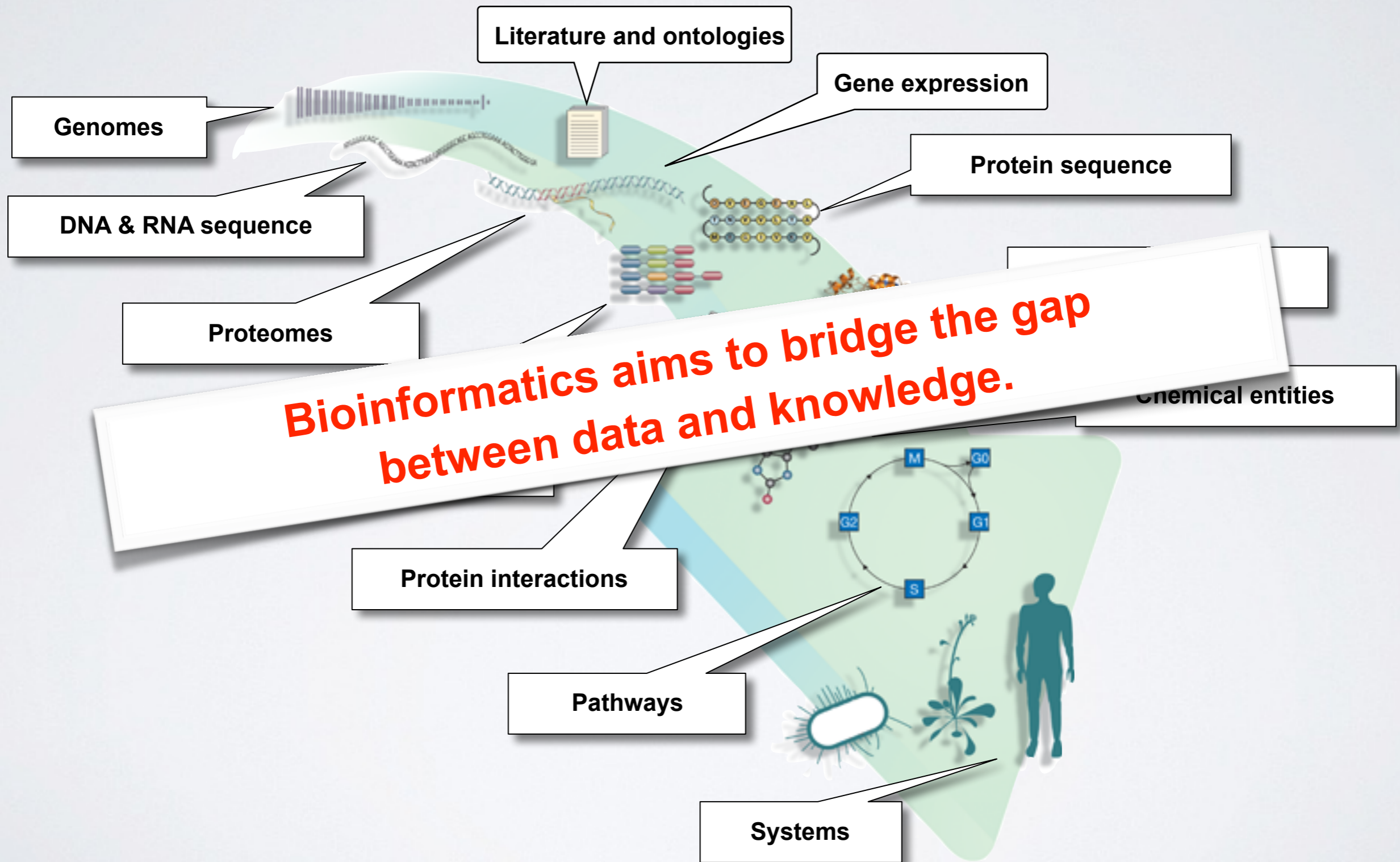
Major types of Bioinformatics Data



Major types of Bioinformatics Data



Major types of Bioinformatics Data



BIOINFORMATICS RESEARCH AREAS

Include but are not limited to:

- Organization, classification, dissemination and analysis of biological and biomedical data (particularly '-omics' data).
- Biological sequence analysis and phylogenetics.
- Genome organization and evolution.
- Regulation of gene expression and epigenetics.
- Biological pathways and networks in healthy & disease states.
- Protein structure prediction from sequence.
- Modeling and prediction of the biophysical properties of biomolecules for binding prediction and drug design.
- Design of biomolecular structure and function.

With applications to Biology, Medicine, Agriculture and Industry

Where did bioinformatics come from?

Bioinformatics arose as molecular biology began to be transformed by the emergence of molecular sequence and structural data

Recap: The key dogmas of molecular biology

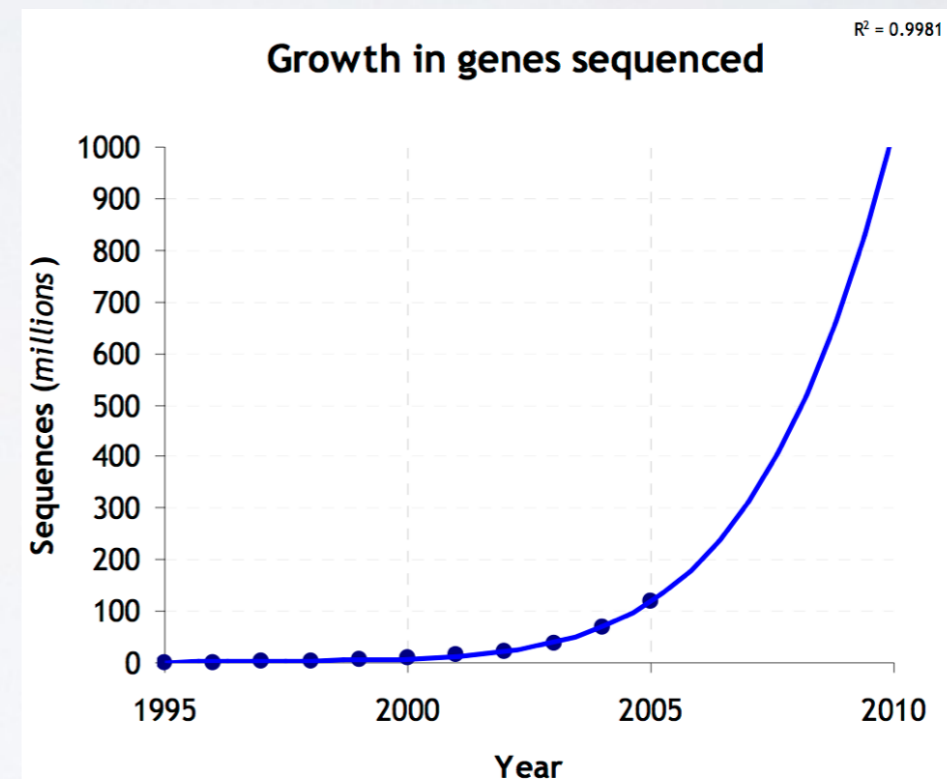
- *DNA sequence determines protein sequence.*
- *Protein sequence determines protein structure.*
- *Protein structure determines protein function.*
- *Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.*

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

Why do we need Bioinformatics?

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

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

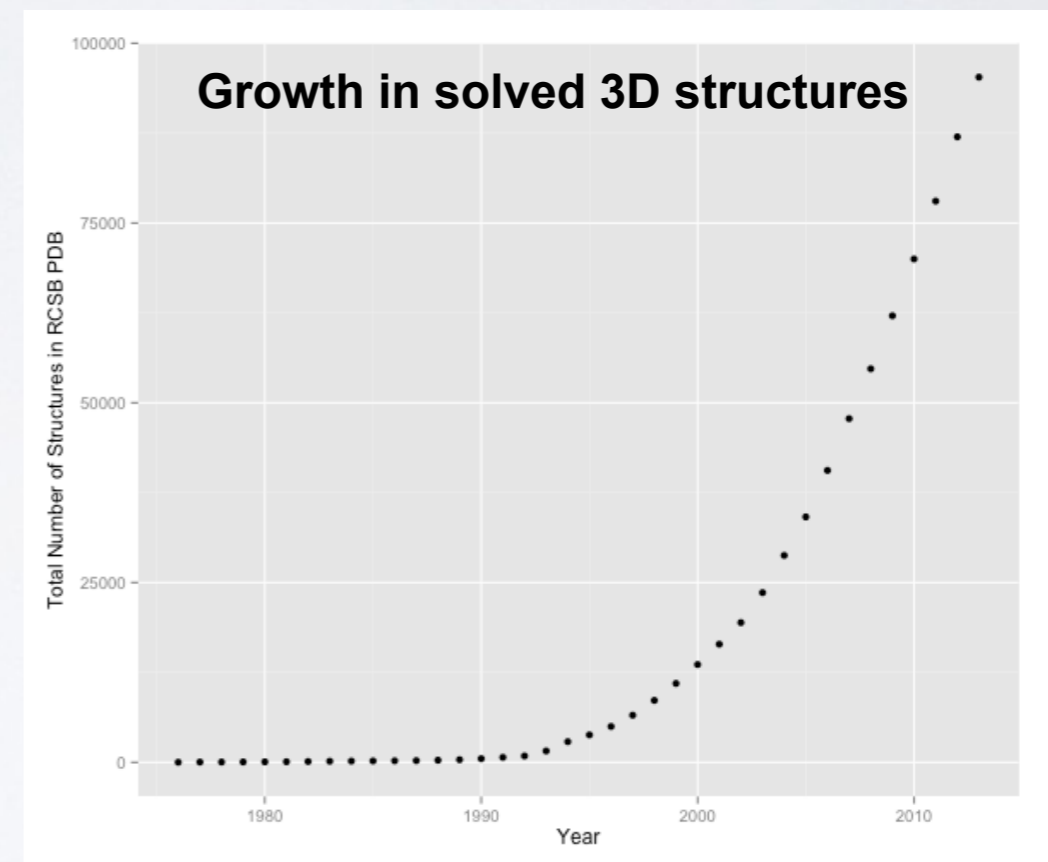


E.G. data from sequencing, structural genomics, proteomics, new high throughput assays, *etc...*

Why do we need Bioinformatics?

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

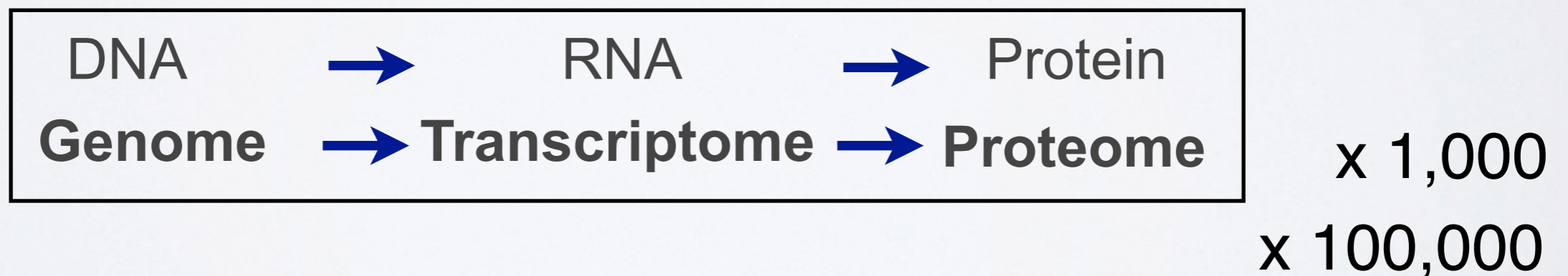
- Bioinformatics provides methods for the efficient:
 - ▶ **storage**
 - ▶ **annotation**
 - ▶ **search and retrieval**
 - ▶ **data integration**
 - ▶ **data mining and analysis**



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

How do we do Bioinformatics?

- A “*bioinformatics approach*” involves the application of **computer algorithms**, **computer models** and **computer databases** with the broad goal of understanding the action of both individual genes, transcripts, proteins and large collections of these entities.



How do we *actually* do Bioinformatics?

Pre-packaged tools and databases

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

Advanced tool application & development

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

How do we *actually* do Bioinformatics?

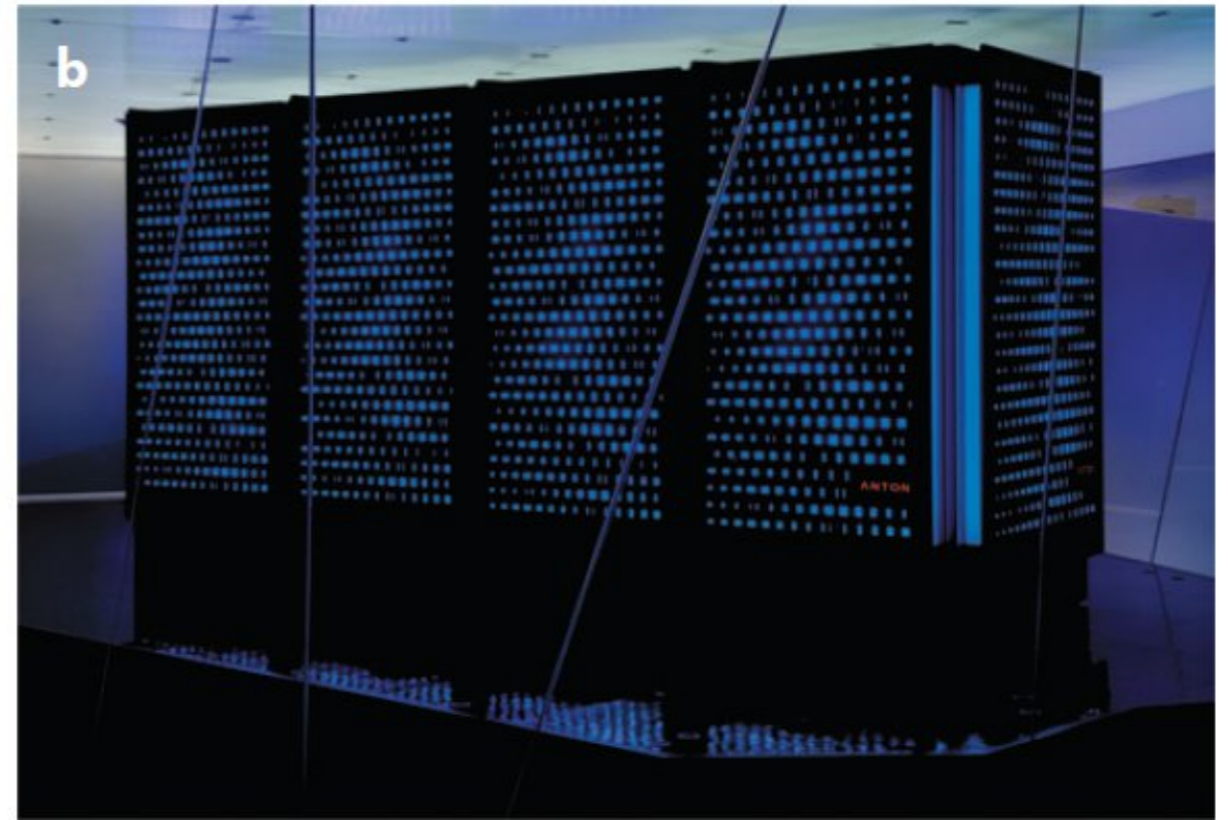
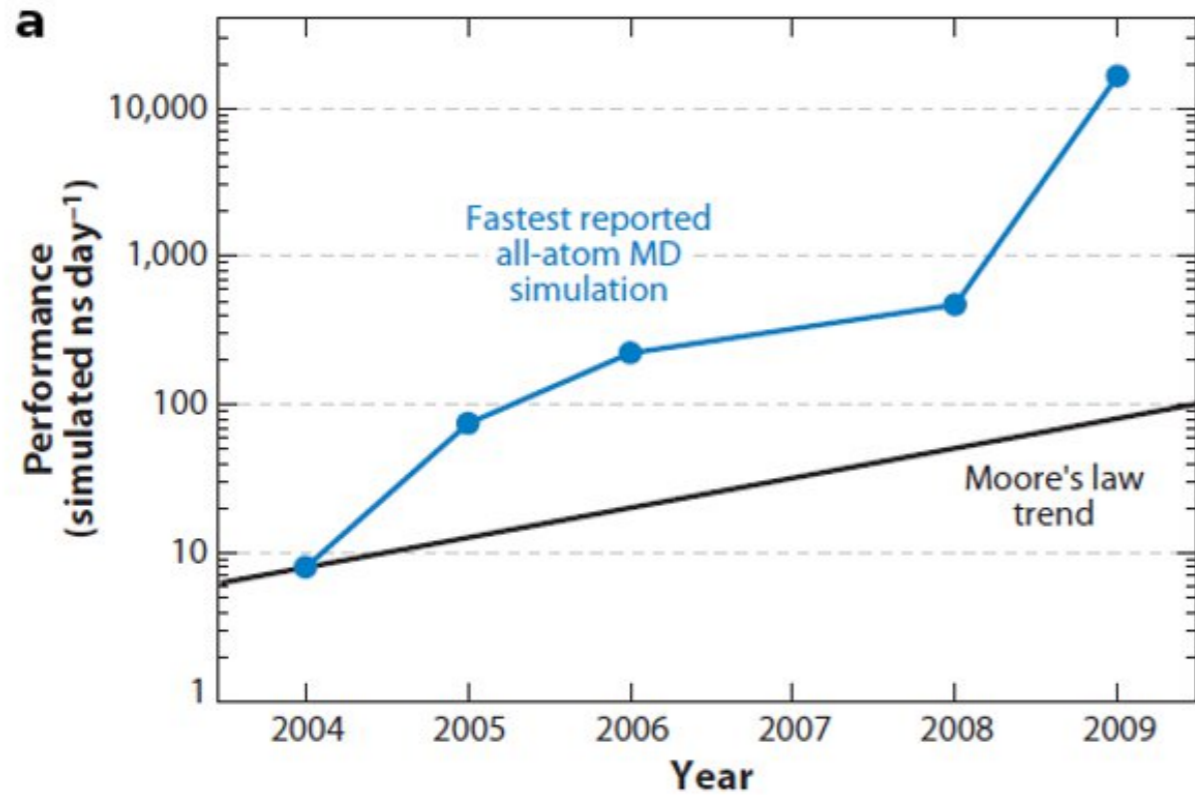
Pre-packaged tools and databases

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

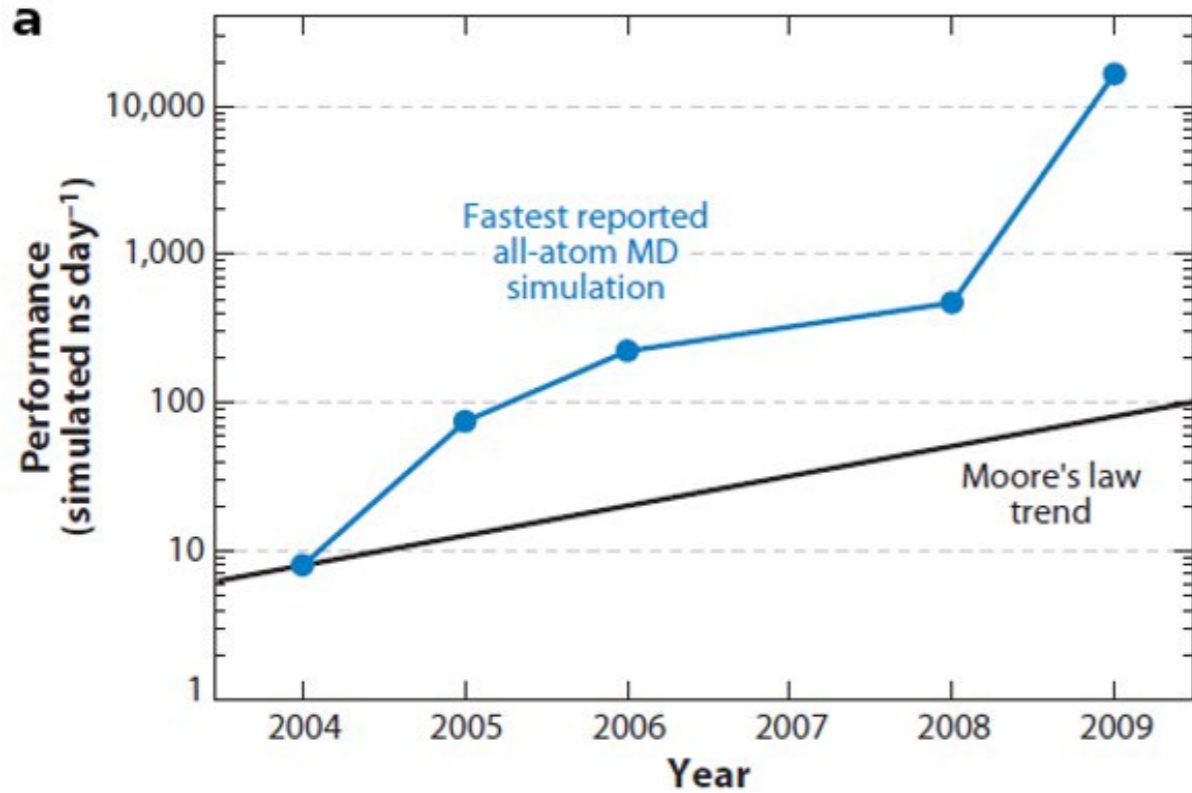
Advanced tool application & development

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

SIDE-NOTE: SUPERCOMPUTERS AND GPUS



SIDE-NOTE: SUPERCOMPUTERS AND GPUS



HOW COMPUTERS HAVE CHANGED

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

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



Skepticism & Bioinformatics

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

- Do they make sense?
- Is it what we expected?
- Do we have adequate controls, and how did they come out?
- Modeling is modeling, but biology is different...

What does this model actually contribute?

- Avoid the miss-use of 'black boxes'

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
 Select the maximum number of aligned sequences to display [?](#)

Short queries Automatically adjust parameters for short input sequences [?](#)

Expect threshold [?](#)

Word size [?](#)

Max matches in a query range [?](#)

Scoring Parameters

Matrix [?](#)

Gap Costs [?](#)

Compositional adjustments [?](#)

Filters and Masking

Filter Low complexity regions [?](#)

Mask Mask for lookup table only [?](#)
 Mask lower case letters [?](#)

PSI/PHI/DELTA BLAST

Upload PSSM no file selected
Optional

PSI-BLAST Threshold [?](#)

Pseudocount [?](#)

Even Blast has many settable parameters

Related tools with different terminology

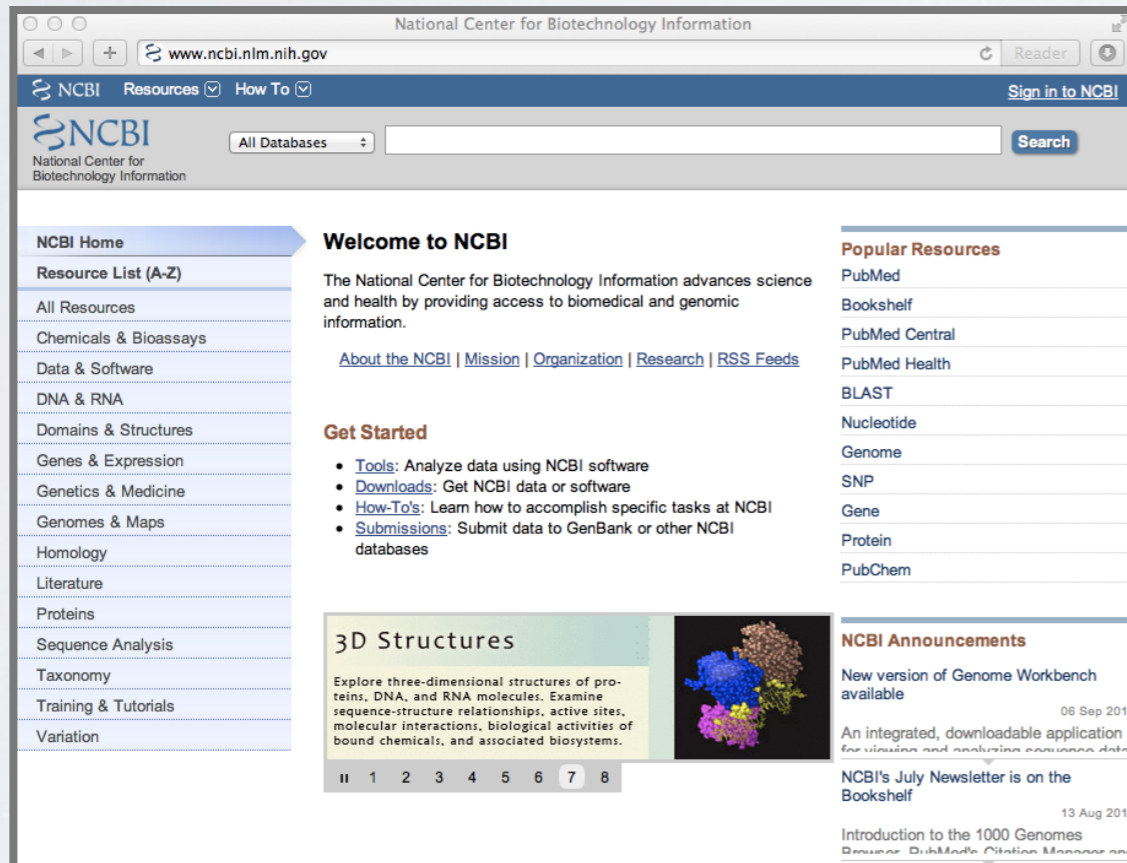
STEP 3 - Set your parameters

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"/>					

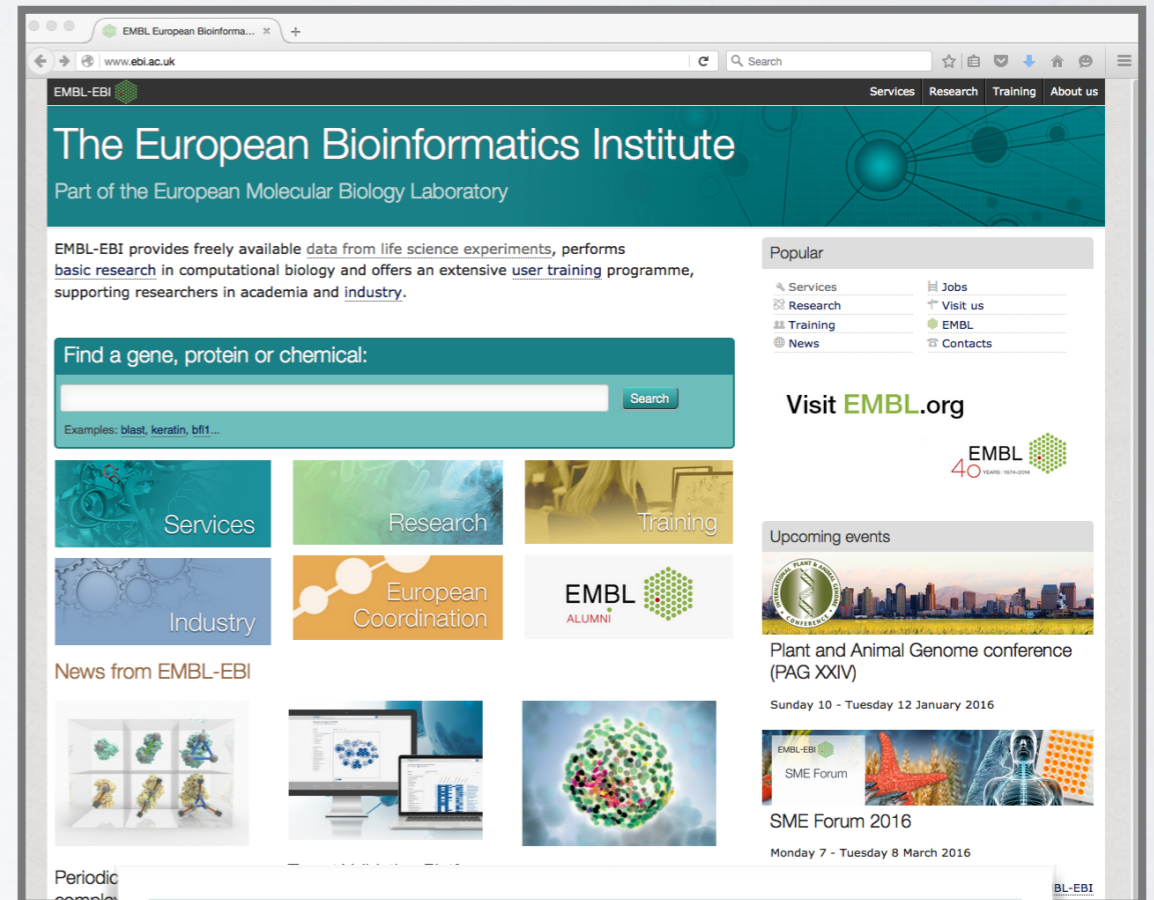
Key Online Bioinformatics Resources: NCBI & EBI

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



The screenshot shows the NCBI website homepage. The browser address bar displays '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 an 'NCBI Announcements' section.

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



The screenshot shows the EBI website homepage. The browser address bar displays 'www.ebi.ac.uk'. The page features a navigation menu on the right with categories like 'Services', 'Research', 'Training', and 'About us'. The main content area includes a 'The European Bioinformatics Institute' header, a search bar, and a 'Find a gene, protein or chemical:' section. There are also sections for 'Services', 'Research', 'Training', 'Industry', 'European Coordination', 'EMBL ALUMNI', 'News from EMBL-EBI', and 'Upcoming 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.

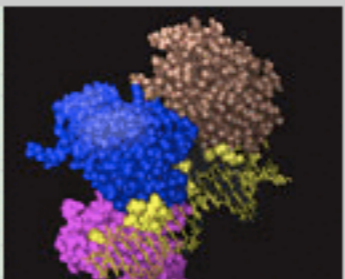
[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

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 'All Resources'. The main content area includes a 'Welcome to NCBI' message and a 'Get Started' section with links to 'Tools', 'Downloads', 'How-To's', and 'Submissions'. A 'Popular Resources' box is overlaid on the right side of the page, listing various services: PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. Red arrows point to PubMed, BLAST, and SNP, 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 and health information.

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

Get Started

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

3D Structures

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

Resources

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, and 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 navigation menu with "NCBI Home" and "Resource List (A-Z)", and a "Popular Resources" section with a link to "PubMed".

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

This screenshot shows a section of the NCBI website with a sidebar on the left containing links to "Homology", "Literature", "Proteins", "Sequence Analysis", "Taxonomy", "Training & Tutorials", and "Variation". The main content area is titled "databases" and features a "3D Structures" section with a description: "Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals, and associated biosystems." To the right of this text is a 3D molecular model. Further right, there are links for "Protein" and "PubChem". At the bottom right, there is an "NCBI Announcements" section with 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

The screenshot shows the NCBI website homepage. At the top, there is a navigation bar with 'NCBI Resources' and 'How To' menus, and a 'Sign in to NCBI' button. Below this is a search bar with a dropdown menu set to 'All Databases'. The main content area is divided into several sections: 'NCBI Home' with a 'Resource List (A-Z)' sidebar; 'Welcome to NCBI' with a brief description of the center's mission; 'Popular Resources' listing services like PubMed, Bookshelf, and BLAST; 'Get Started' with links for tools, downloads, and submissions; '3D Structures' featuring a molecular model; and 'NCBI Announcements' with recent news items.

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

The screenshot shows the EBI website homepage. The header features the EBI logo and navigation links for 'Services', 'Research', 'Training', and 'About us'. The main heading is 'The European Bioinformatics Institute', part of the European Molecular Biology Laboratory. A central search bar is labeled 'Find a gene, protein or chemical:'. Below the search bar are several featured tiles for 'Services', 'Research', 'Training', 'Industry', and 'European Coordination'. A 'News from EMBL-EBI' section displays various articles. On the right side, there is a 'Popular' section with links to 'Services', 'Research', 'Training', and 'News', along with 'Jobs' and 'Visit us' links. The bottom right corner features 'Upcoming events' including the 'Plant and Animal Genome conference (PAG XXIV)' and the 'SME Forum 2016'.

<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'. A large teal banner contains the text 'The European Bioinformatics Institute' and 'Part of the European Molecular Biology Laboratory'. Below this, a paragraph states: 'EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.' A search bar is provided with the prompt 'Find a gene, protein or chemical:' and a 'Search' button. Below the search bar, a grid of six colored buttons is visible: 'Services' (teal, highlighted with a red border), 'Research' (green), 'Training' (yellow), 'Industry' (blue), 'European Coordination' (orange), and 'EMBL ALUMNI' (white with green logo). To the right, a 'Popular' section lists links for 'Services', 'Research', 'Training', 'News', 'Jobs', 'Visit us', 'EMBL', and 'Contacts'. Below this is a 'Visit EMBL.org' section with the EMBL 40th anniversary logo (1974-2014). An 'Upcoming events' section features a banner for the 'Plant and Animal Genome conference (PAG XXIV)' held from Sunday 10 to 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 EMBL-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 provides a grid of service categories. A "Popular" sidebar lists key databases, with Ensembl and UniProt highlighted by a red box. The "Proteins" category in the main grid is also highlighted with a red box. A "Training" banner is visible at the bottom right.

Services < EMBL-EBI

www.ebi.ac.uk/services

Services Research Training About us

Services

Overview A to Z Data submission Support

Bioinformatics services

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

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

Programmatic access

Popular

- Ensembl
- UniProt
- PDB
- ArrayExpress
- ChEMBL








Training

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

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

Proteins

Popular services

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

Quick links

- 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

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, bf1...

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

INTERNATIONAL PLANT & ANIMAL GENOME CONFERENCE

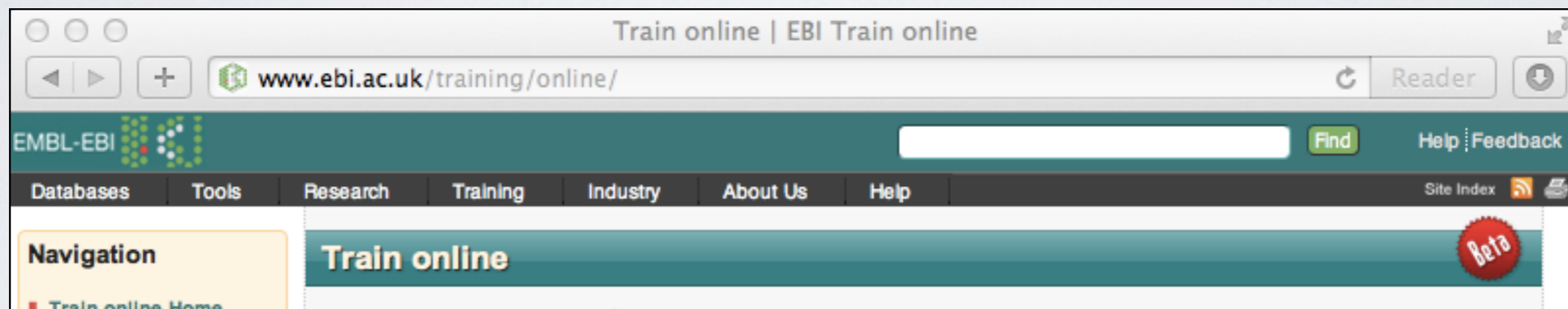
Plant and Animal Genome conference (PAG XXIV)

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-embl-ebi. The page features a navigation menu with 'Services', 'Research', 'Training', and 'About us'. The main heading is 'Train online'. Below this, there are links for 'Training', 'Train online Home', 'Course list', 'Glossary', 'Support & Feedback', and 'Log in / Register'. The breadcrumb trail is 'training » online » course-list » using-sequence-similarity-searching-tools-embl-ebi'. The 'Course content' section includes 'Using sequence similarity searching tools at EMBL-EBI: webinar' (highlighted) and 'Contributors'. A 'Print Course' link is also present. The main content area displays a video player for the webinar, with a title 'Using sequence similarity searching tools at EMBL-EBI: webinar'. The video thumbnail shows the text 'Using sequence similarity search tools at EMBL-EBI' and 'Finding homologous sequences with BLAST, FASTA, PSI-Search etc.' along with a photo of Andrew Cowley and his contact information: 'andrew.cowley@ebi.ac.uk' and 'support@ebi.ac.uk'. The video player shows a duration of 0:00 / 37:42. To the right of the video player, 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'. Below the video player, a text box states: 'This webinar focuses on how to use tools like **BLAST** and PSI-Search to find homologous sequences in EMBL-EBI databases, including tips on which tool and database to use, input formats, how to change parameters and how to interpret the results pages.'

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



Notable EBI databases include:
ENA, UniProt, Ensembl
and the tools FASTA, BLAST, InterProScan,
MUSCLE, DALI, HMMER

Find a course

Browse by subject



[Genes and Genomes](#)



[Gene Expression](#)



[Interactions, Pathways and Networks](#)

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, H, H, K, MHC, MZRGbase, IMG, Kabat, KDNA, Medline, Mendel, MEROPS, MGDB, MGI, MMTDB, Mol-R-U, MPDB, MRR, MutBase, Myc, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-MODEL Repository, SWISS-PROT, TeIDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc !!!!

There are lots of Bioinformatics Databases

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

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

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

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](#).
- 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

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
TTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGG
GGATCTGTCCACTCCTGATGCAGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCA
TCACTTTGGCAAAGAATTCACCCACAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCCACAAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTT
```

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

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

