



BGGN 213

Foundations of Bioinformatics

Barry Grant
UC San Diego

<http://thegrantlab.org/bggn213>

HELLO
my name is

BARRY

bjgrant@ucsd.edu

HELLO
HIS my name is

NATHAN

ndpalmer@ucsd.edu

Introduce Yourself!

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

Today's Menu

Course Logistics

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

Learning Objectives

What you need to learn to succeed in this course.

Course Structure

Major lecture topics and specific learning goals.

Introduction to Bioinformatics

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

Bioinformatics Database

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

<http://thegrantlab.org/bggn213/>

The screenshot shows a web browser window displaying the course website. The URL in the address bar is bioboot.github.io/bggn213_S18/. The page title is "Bioinformatics (BGGN 213, Spring 2018)". On the left, there is a sidebar with the UC San Diego logo and the course name "BGGN 213". Below the course name, a description states: "A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD". A list of navigation links includes: Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. At the bottom of the sidebar are social media icons for Twitter, GitHub, Email, and RSS. The main content area features sections for Course Director (Prof. Barry J. Grant, email: bjgrant@ucsd.edu), Instructional Assistant (Yuansheng Zhou, email: yuz461@ucsd.edu), and Course Syllabus (Spring 2018 (PDF)). To the right of the syllabus link is a cartoon DNA helix icon with a magnifying glass focusing on the sequence "101 110". The "Overview" section defines Bioinformatics as the application of computational and analytical methods to biological problems, describing it as a rapidly maturing field driving data collection, analysis, and interpretation. It also states that the course is designed for bioscience graduate students.

bioboot.github.io/bggn213_S18/

Home Gmail Gcal Bitbucket GitHub News Disqus BGGN-213 BIMM-143 BIMM-194 Atmosphere Blink GDocs Galaxy

UC San Diego

BGGN 213

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

Overview

Lectures

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

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

Bioinformatics (BGGN 213, Spring 2018)

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

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

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

Overview

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

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

<http://thegrantlab.org/bggn213/>

UC San Diego

BGNN 213

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

Overview

Lectures

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

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

bioboot.github.io/bggn213_S18/

**Bioinformatics
(BGNN 213, Spring 2018)**

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

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

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

Overview

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

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

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

The screenshot shows a web browser window with the URL bioboot.github.io/bggm213_f17/goals/. The page is titled "Learning Goals". On the left, there is a sidebar for the course BGGN 213, which includes links for Overview, Lectures, Computer Setup, Learning Goals (which is highlighted with a red box), Assignments & Grading, Ethics Code, and Screen Cast Videos. Below the sidebar is a footer with social media icons for Twitter, GitHub, Email, and RSS.

Learning Goals

At the end of this course students will:

- Understand the increasing necessity for computation in modern life sciences research.
- Be able to use and evaluate online bioinformatics resources including major biomolecular and genomic databases, search and analysis tools, genome browsers, structure viewers, and select quality control and analysis tools to solve problems in the biological sciences.
- Be able to use the UNIX command line and the R environment to analyze bioinformatics data at scale.
- Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.
- Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genomics, Transcriptomics and Structural bioinformatics.

In short, students will develop a solid foundational knowledge of bioinformatics and be able to evaluate new biomolecular and genomic information using existing bioinformatic tools and resources.

At the end of this course students will:

- Understand the increasing necessity for computation in modern life sciences research.
- Be able to use and evaluate online bioinformatics resources and analysis tools to solve problems in the biological sciences.
- Be able to use the R environment to analyze bioinformatics data at scale.
- Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genome informatics, Transcriptomics and Structural informatics.

In short, you will develop a solid foundational knowledge of **bioinformatics** and be able to evaluate new biomolecular and genomic information using **existing bioinformatic tools and resources.**

Specific Learning Goals....

What I want you to know by course end!

The screenshot shows a web browser window with the URL bioboot.github.io/bggn213_f17/goals/. The page title is "UCSanDiego BGGN 213". The sidebar on the left lists course navigation links: Overview, Lectures, Computer Setup, Learning Goals (which is highlighted with a red box), Assignments & Grading, Ethics Code, and Screen Cast Videos. The main content area is titled "Specific Learning Goals". It contains a text block about teaching goals and a table listing four learning objectives with their corresponding lecture numbers.

		Lecture(s):
1	Appreciate and describe in general terms the role of computation in hypothesis-driven discovery processes within the life sciences.	1, 2, 20
2	Be able to query, search, compare and contrast the data contained in major bioinformatics databases and describe how these databases intersect (GenBank, GENE, UniProt, PFAM, OMIM, PDB, UCSC, ENSEMBLE).	2, 12, 13
3	Describe how nucleotide and protein sequence and structure data are represented (FASTA, FASTQ, GenBank, UniProt, PDB).	3, 10
4	Be able to describe how dynamic programming works for pairwise sequence alignment and appreciate the differences	4, 5

Course Structure

Derived from specific learning goals

The screenshot shows a web browser window with the URL bioboot.github.io/bggn213_S18/lectures/. The page is titled "Lectures".

Lectures

All Lectures are Wed/Fri 1:00-4:00 pm in Warren Lecture Hall 2015 (WLH 2015) ([Map](#)). Clicking on the class topics below will take you to corresponding lecture notes, homework assignments, pre-class video screen-casts and required reading material.

#	Date	Topics for Spring 2018
1	Wed, 04/04	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	Fri, 04/06	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

Course Structure

Derived from specific learning goals

bioboot.github.io/bggn213_S18/lectures/

Lectures

All Lectures are Wed/Fri 1:00-4:00 pm in Warren Lecture Hall 2015 (WLH 2015) ([Map](#)). Clicking on the class topics below will take you to corresponding lecture notes, homework assignments, pre-class video screen-casts and required reading material.

#	Date	Topics for Spring 2018
1	Wed, 04/04	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	Fri, 04/06	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

Class Details

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

The screenshot shows a web browser window with the URL bioboot.github.io/bggn213_f17/lectures/#1. The page content is as follows:

UCSanDiego
BGGN 213
A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

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

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

Homework

Goals, Class material, Screencasts & Homework

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

- URL:** bioboot.github.io/bggn213_f17/lectures/#1
- Page Title:** UC San Diego BGGN 213
- Page Content:**
 - Homework:**
 - [Questions](#),
 - Readings:
 - PDF1: [What is bioinformatics? An introduction and overview](#),
 - PDF2: [Advancements and Challenges in Computational Biology](#),
 - Other: [For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights](#) New York Times, 2014.
 - Screen Casts:**
 - Welcome to "Foundations of Bioinformatics" (BGGN-21...)A video thumbnail for a screen cast titled "Welcome to Foundations of Bioinformatics". The thumbnail features a man with short brown hair, Barry Grant, standing in front of a dark background with colorful, glowing spheres (representing data points) floating around him. The text "BGGN 213 Foundations of Bioinformatics" is overlaid on the left side of the thumbnail, and the URL "http://thegrantlab.org/bggn213" is at the bottom.

Homework

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

UCSanDiego

BGGN 213

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

Overview

Lectures

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

Screen Cast Videos

Homework:

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

Screen Casts:

Welcome to “Foundations of Bioinformatics” (BGGN-21...)

BGGN 213

Foundations of Bioinformatics

Barry Grant

UC San Diego

<http://thegrantlab.org/bggn213>

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

Homework

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

The screenshot shows a Google Forms survey window. At the top, the URL is `docs.google.com/forms/d/e/1FAIpQLSeN3pg-AaRg5la3PxZuqSj`. Below the URL, there are navigation icons for Home, Gmail, Gcal, Bitbucket, GitHub, News, and Disqus. The main title of the form is "BGGN213 Lecture 1 Homework (F17)". A purple header bar contains the title. The form asks for a UCSD username/email address, which is described as the part before "@ucsd.edu". There is a text input field labeled "Your answer". A question follows: "Which of the following operating systems is most frequently used for bioinformatics tool development?" with four options: Windows, iOS, Unix, and Perl. The first three options have radio buttons, while Perl has a checkbox.

BGGN213 Lecture 1 Homework (F17)

Please answer the following questions

* Required

Your UCSD username/email address *

The first part of your UCSD email address before the '@[ucsd.edu](#)' part

Your answer

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

Windows

iOS

Unix

Perl

Homework

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

docs.google.com/forms/d/e/1FAIpQLSeN3pg-AaRg5la3PxZuqSj

Home Gmail Gcal Bitbucket GitHub News Disqus

BGGN213 Lecture 1 Homework

Please answer the following questions

* Required

Your name/email address *

Part of your UCSD email address before the '@ucsd.edu' part

Your answer

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

- Windows
- iOS
- Unix
- Perl

Which of the following databases contains primarily protein

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

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

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

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

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

BGGN-213 Learning Goals....

Advanced UNIX and R based learning goals

UCSanDiego

BGGN 213

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

Overview

Lectures

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

Screen Cast Videos

bioboot.github.io/bggn213_f17/goals/

Home Gmail Gcal Bitbucket GitHub News Disqus BGGN-213

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

BGGN-213 Learning Goals....

Delve deeper into “real-world” bioinformatics

UCSanDiego

BGGN 213

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

Overview

Lectures

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

Screen Cast Videos

bioboot.github.io/bggn213_f17/goals/

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

These support a major learning objective

At the end of this course students will:

- Understand the increasing necessity for computation in modern life sciences research.
- Be able to use and evaluate online bioinformatics resources and analysis tools to solve problems in the biological sciences.
- Be able to use the R environment to analyze bioinformatics data at scale.
- Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genome informatics, Transcriptomics and Structural informatics.

Why use R?

Productivity

Flexibility

Designed for data analysis

IEEE 2016 Top Programming Languages

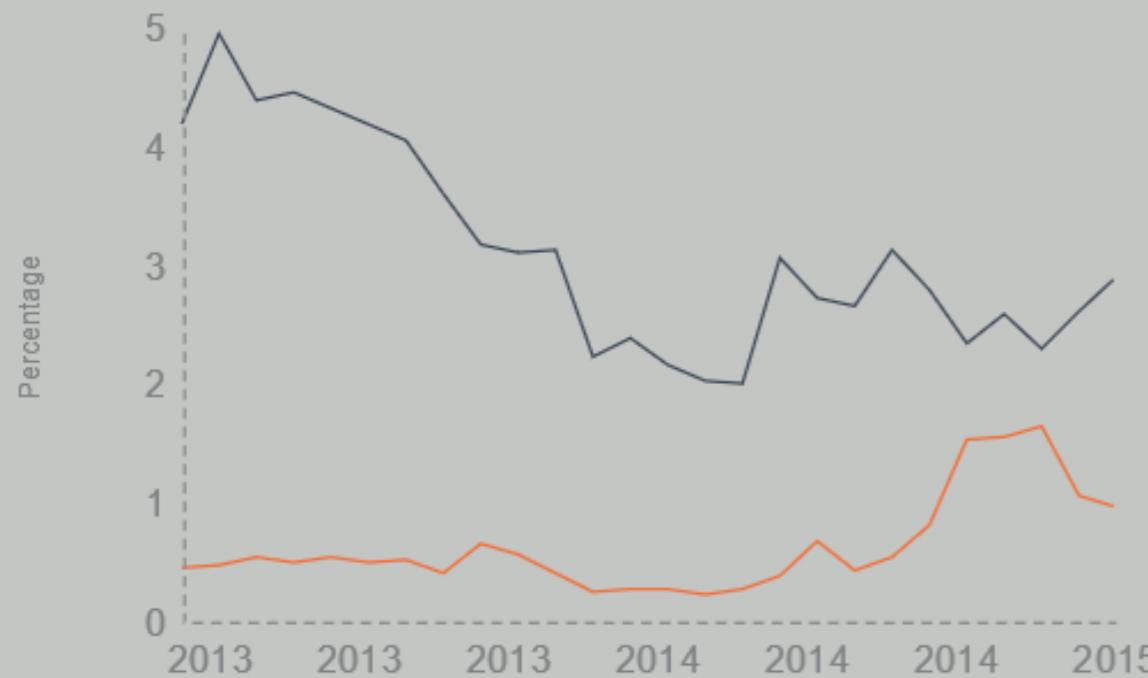
Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

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

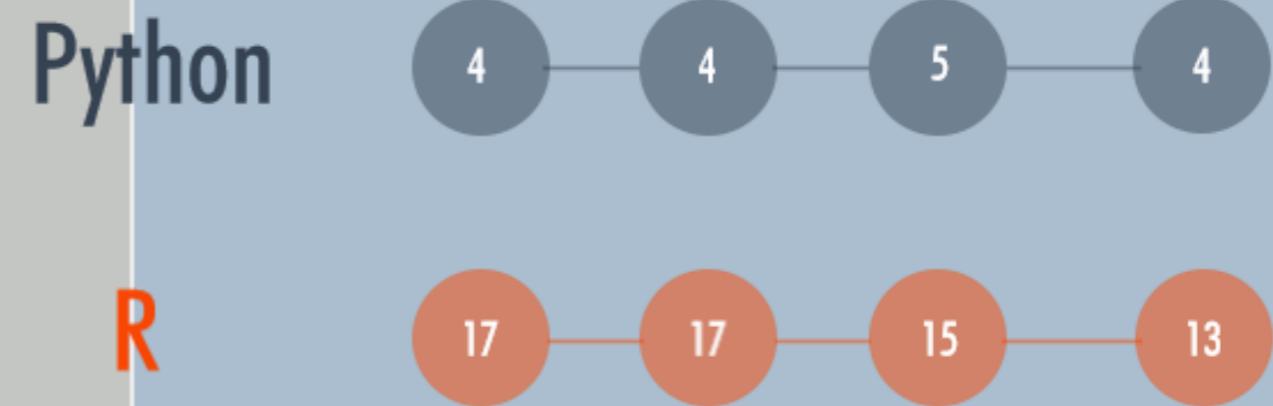
R and Python: The Numbers

Popularity Rankings

R and Pythons popularity between 2013 and February 2015 (Tiobe Index)



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



Jobs And Salary?

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



\$ 115,531



Python

\$ 94,139

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

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

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

The screenshot shows the DataCamp homepage with a red circle highlighting the user profile icon in the top right corner. The profile icon has a red notification badge with the number '3'.

Your Latest Activity

Introduction to Spark in R using dplyr

You are doing awesome barryus! So far you've earned 250 XP!

The last chapter you were working on was [Light My Fire: Starting To Use Spark With dplyr Syntax](#).

DAILY PRACTICE

Learning data science requires practice **every day**. Build your data science fluency with DataCamp practice mode.

Notifications:

- You have a new assignment: Conditionals and Con... 16 days ago
- You have a new assignment: Working with the RSt... 16 days ago
- You have a new assignment: Introduction to R 16 days ago
- bjgrant invited you to the group 'Foundations o... 16 days ago
- You have a new assignment: Orientation 9 months ago

[See all notifications](#)

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

The screenshot shows a DataCamp course page titled "What is an IDE anyway?". The page includes a brief description of RStudio, a question about what IDE stands for, and a section titled "Possible Answers" with five options. The option "Integrated Development Environment" is circled in red. Below it is a "Take Hint (-15xp)" button. At the bottom is a large red-outlined "Submit Answer" button. To the right is a screenshot of the RStudio IDE interface, showing the console output of an R session, the environment pane, and the files pane.

What is an IDE anyway? | R

Secure | https://campus.datacamp.com/courses/working-with-the-rstudio-ide-part-1/orientation?ex=2

DataCamp

Course Outline

What is an IDE anyway?

50xp

RStudio is an IDE that makes R easier to use by combining a set of tools into a single environment.

What does IDE stand for?

Possible Answers

- Intensive Design Environment
- Integrated Document Environment
- Independent Developer Ecosystem
- Integrated Development Environment

Take Hint (-15xp)

Submit Answer

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

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

Natural language support but running in an English locale

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

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

> |

Environment History

Import Dataset List Global Environment

Files Plots Packages Help Viewer

New Folder Upload Delete Rename

Home

Name	Size

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

The screenshot shows the RStudio IDE interface running within a web browser window. The browser title bar reads "What is an IDE anyway? | R". The address bar shows a secure connection to "https://campus.datacamp.com/courses/working-with-the-rstudio-ide-part-1/orientation?ex=2". The DataCamp logo is in the top left, and a course outline navigation bar is at the top right.

The main content area displays a course exercise titled "What is an IDE anyway?". A message says "Exercise Completed" with a blue button containing "50xp" (experience points). This button is circled in red. Below it, a message says "Nice job! Move onto the next video to start learning more about the RStudio IDE!".

A sidebar on the left lists "Possible Answers" under "PRESS ENTER TO CONTINUE". One answer is highlighted with a red circle around the "Continue" button. Other options include "Intensive Design Document" and "Integrated Document".

A callout box in the bottom left corner encourages becoming a power user, mentioning keyboard shortcuts like "Ctrl + Shift + Enter". It also links to "See all keyboard shortcuts" and "Take Hint (-15xp)".

The RStudio interface includes a top menu bar with File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help, and a session dropdown for "barryus". The "Console" tab is active, showing the R startup message:

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

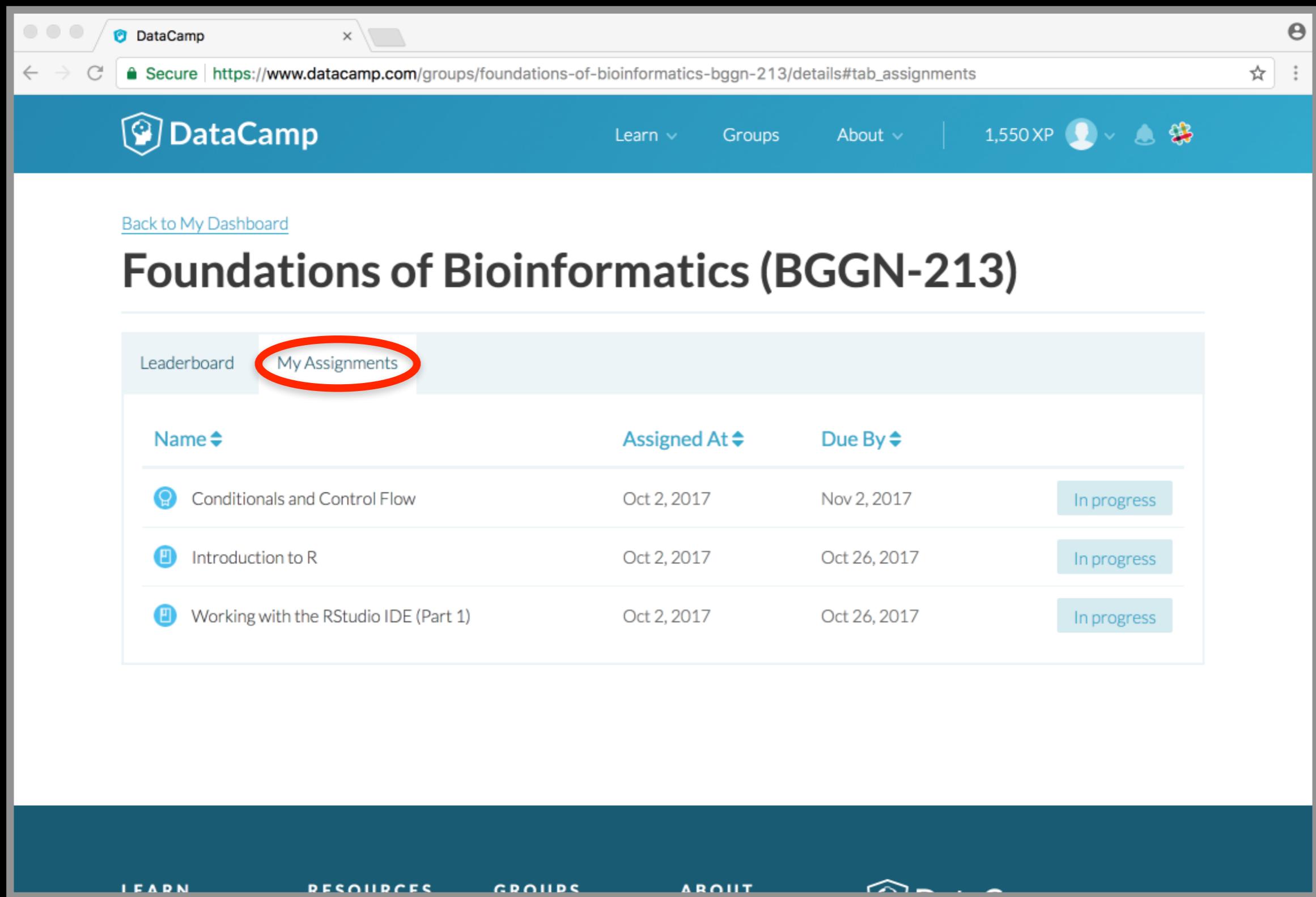
The "Environment" tab in the top right shows an empty global environment. The "Files" tab in the bottom right shows a single folder named "Home".

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

The screenshot shows a web browser window for DataCamp. The URL in the address bar is <https://www.datacamp.com/groups-foundations-of-bioinformatics-bggn-213/details>. The DataCamp logo is in the top left, and the user's profile information (1,550 XP) is in the top right. A red circle highlights the 'Groups' button in the top navigation bar. Below the header, there is a 'Back to My Dashboard' link and the title 'Foundations of Bioinformatics (BGGN-213)'. A navigation bar below the title includes 'Leaderboard' and 'My Assignments' buttons. Underneath is a section for tracking progress over time: '30 Days' (selected), '90 Days', and 'Last Year'. A table then lists the top 8 members of the group, showing their rank, profile picture, name, XP, Courses completed, and Chapters completed.

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

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



The screenshot shows a web browser window for DataCamp. The URL in the address bar is https://www.datacamp.com/groups-foundations-of-bioinformatics-bggm-213/details#tab_assignments. The page title is "Foundations of Bioinformatics (BGGN-213)". At the top, there are navigation links for "Learn", "Groups", and "About", along with a user profile icon showing "1,550 XP". Below the title, there are two tabs: "Leaderboard" and "My Assignments", with "My Assignments" circled in red. The main content area displays a table of assignments:

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

At the bottom of the page, there are links for "LEARN", "RESOURCES", "GROUPS", and "ABOUT".

Today's Menu

Course Logistics

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

Learning Objectives

What you need to learn to succeed in this course.

Course Structure

Major lecture topics and specific learning goals.

Introduction to Bioinformatics

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

Computer Setup

Ensuring your laptop is all set for future sections of this course.

OUTLINE

Overview of bioinformatics

- The what, why and how of bioinformatics?
- Major bioinformatics research areas.
- Skepticism and common problems with bioinformatics.

Online databases and associated tools

- Primary, secondary and composite databases.
 - Nucleotide sequence databases (GenBank & RefSeq).
 - Protein sequence database (UniProt).
 - Composite databases (PFAM & OMIM).

Database usage vignette

- How-to productively navigate major databases.

Q. What is Bioinformatics?

“Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.”

... Bioinformatics is a hybrid of biology and computer science

Q. What is Bioinformatics?

“Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.”

- ... Bioinformatics is a hybrid of biology and computer science
- ... **Bioinformatics is computer aided biology!**

Q. What is Bioinformatics?

“Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.”

- ... Bioinformatics is a hybrid of biology and computer science
- ... **Bioinformatics is computer aided biology!**

Computer based management and analysis of biological and biomedical data with useful applications in many disciplines, particularly genomics, proteomics, metabolomics, etc...

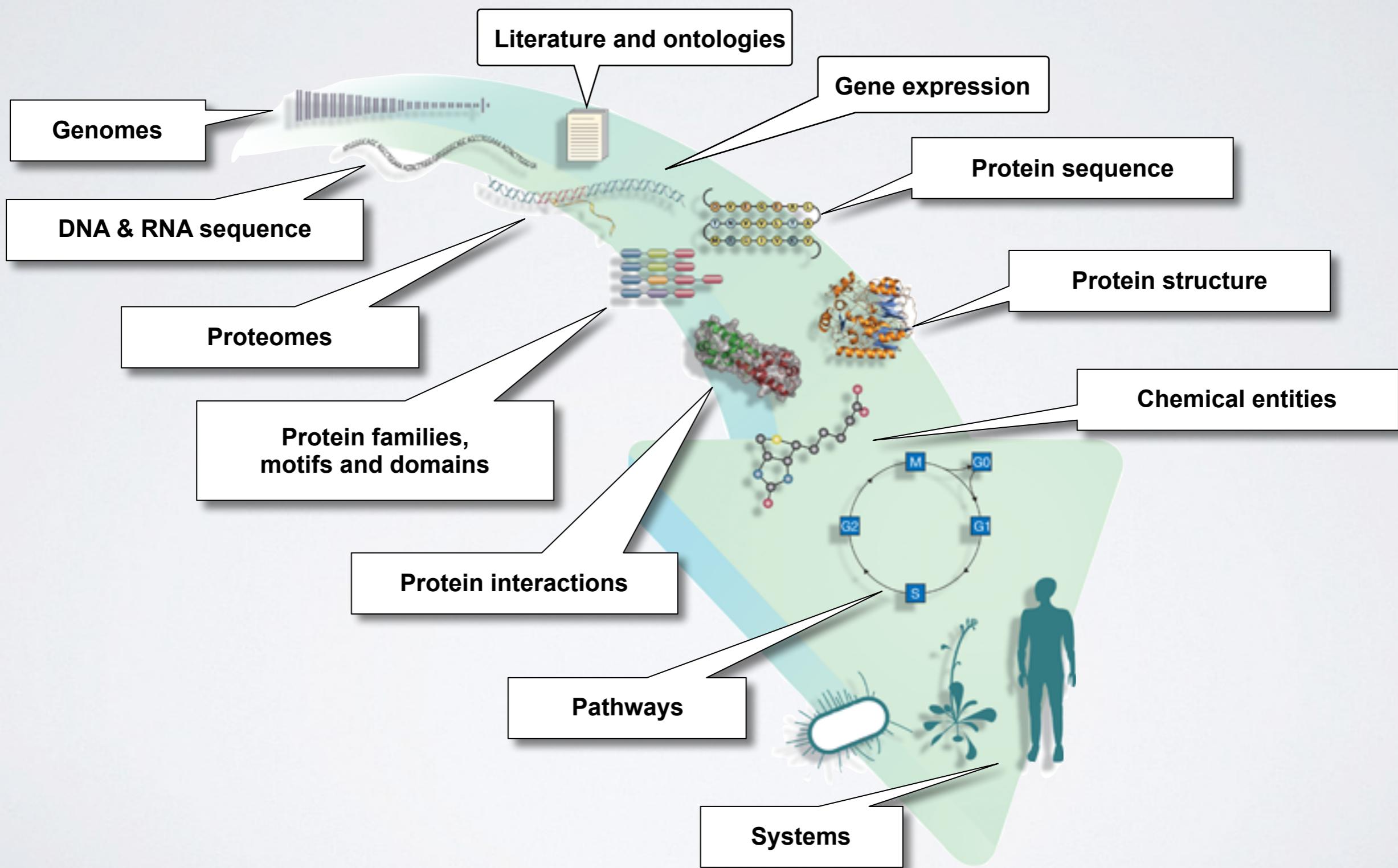
MORE DEFINITIONS

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

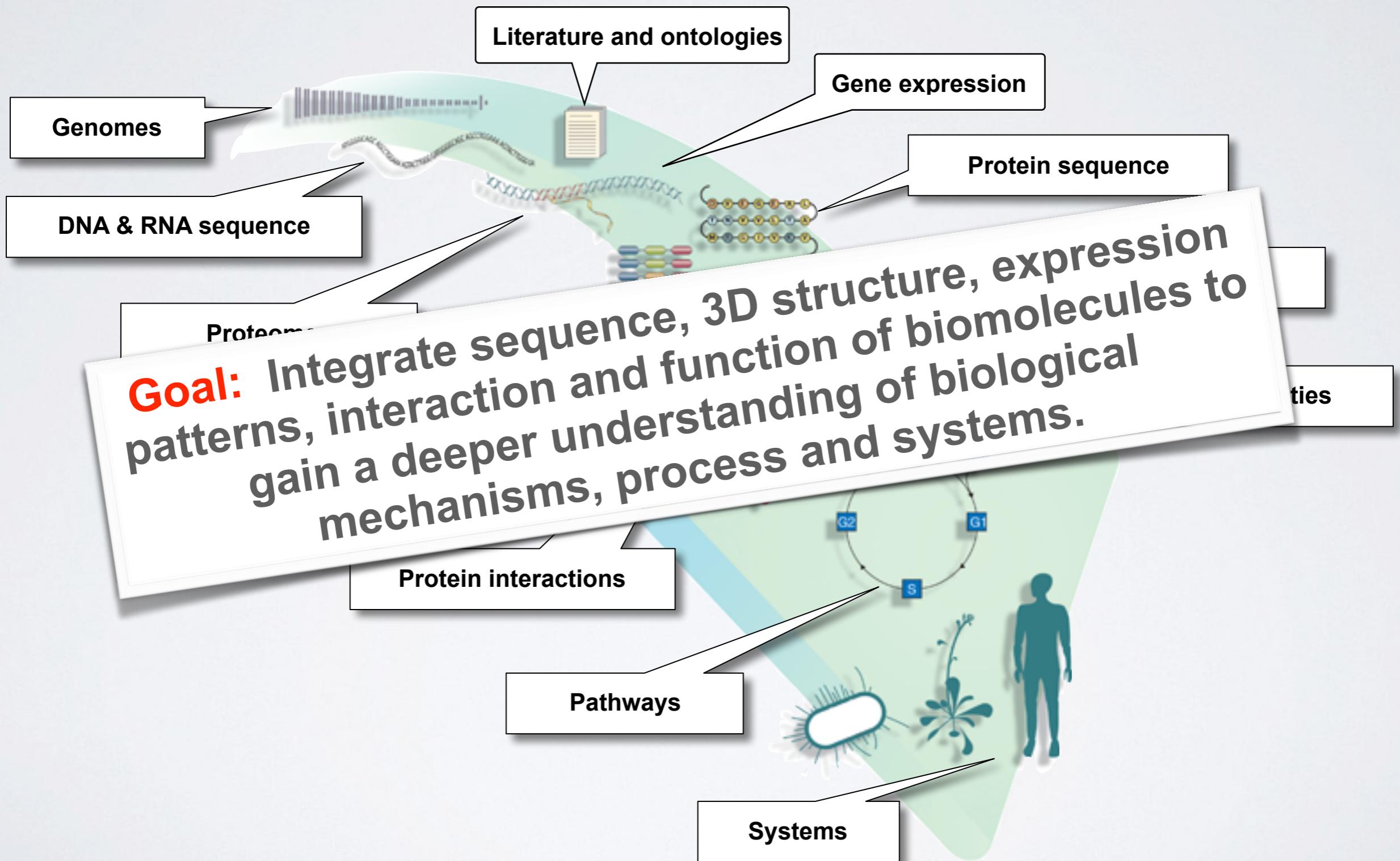
MORE DEFINITIONS

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

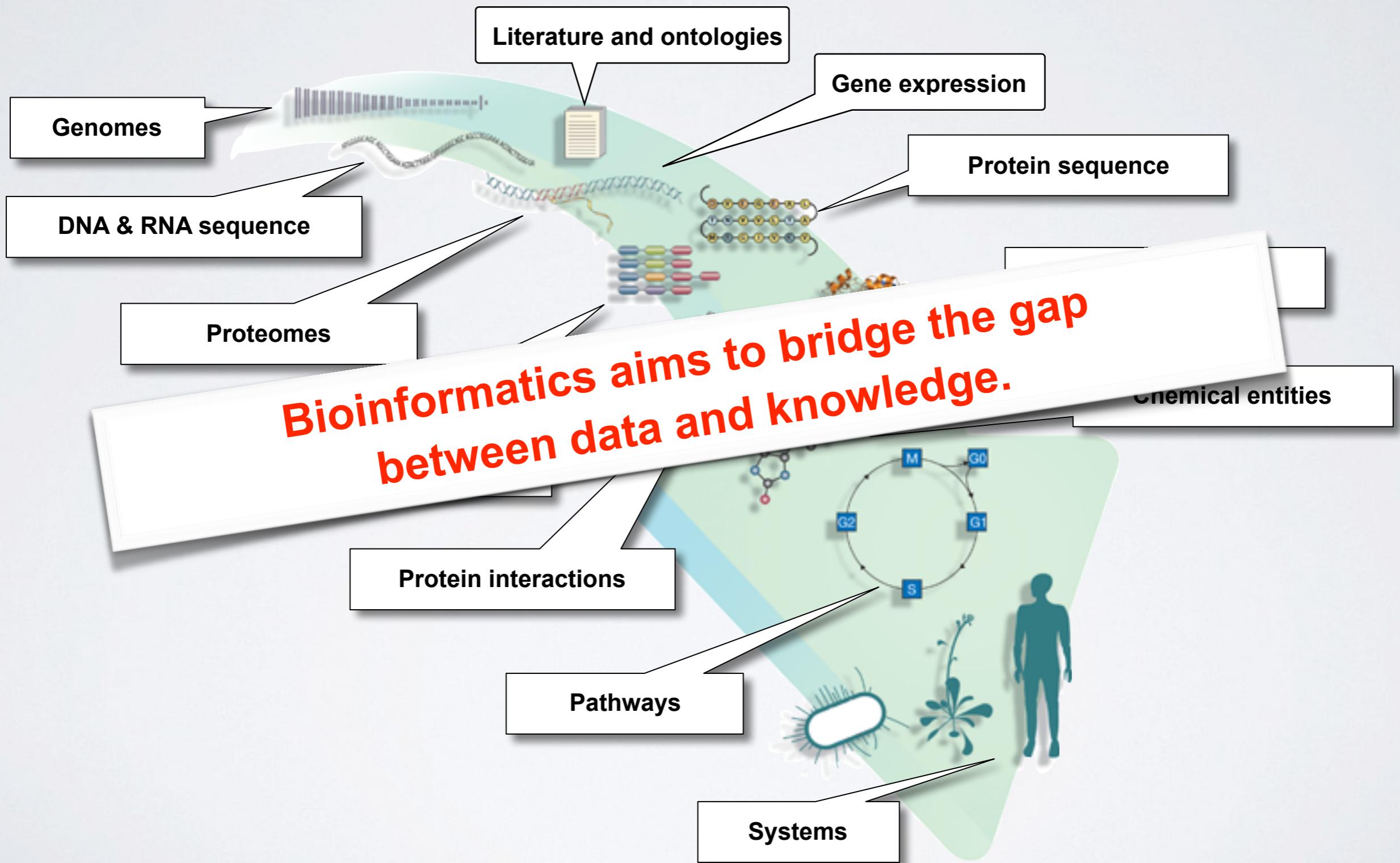
Major types of Bioinformatics Data



Major types of Bioinformatics Data



Major types of Bioinformatics Data



BIOINFORMATICS RESEARCH AREAS

Include but are not limited to:

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

With applications to Biology, Medicine, Agriculture and Industry

Where did bioinformatics come from?

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

Recap: The key dogmas of molecular biology

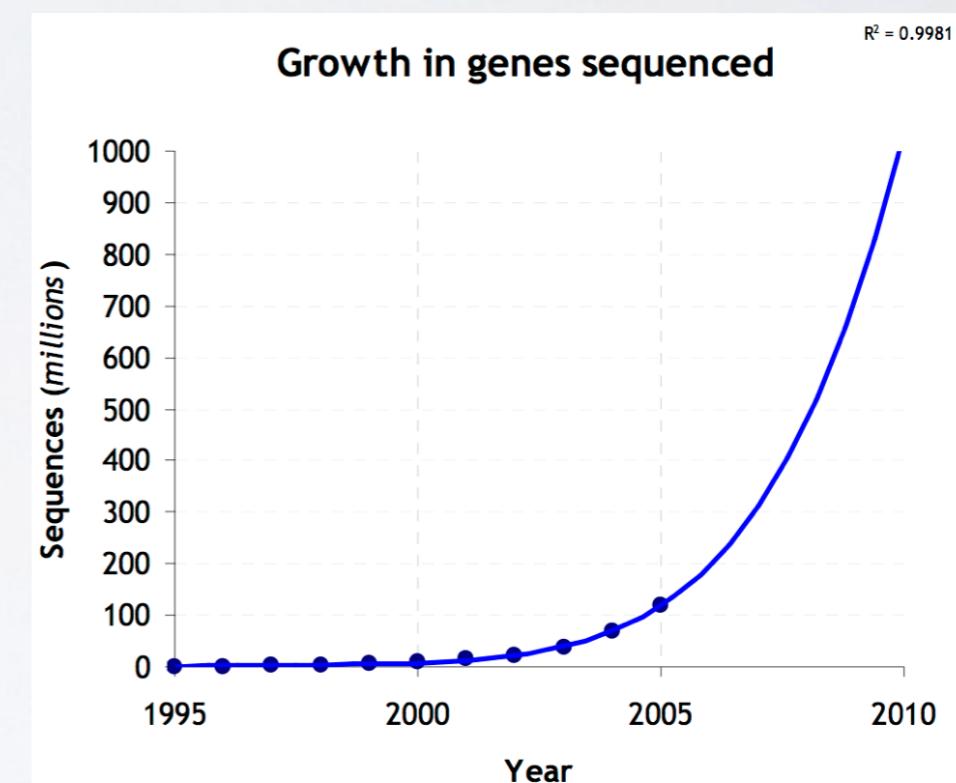
- *DNA sequence determines protein sequence.*
- *Protein sequence determines protein structure.*
- *Protein structure determines protein function.*
- *Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.*

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

Why do we need Bioinformatics?

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

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

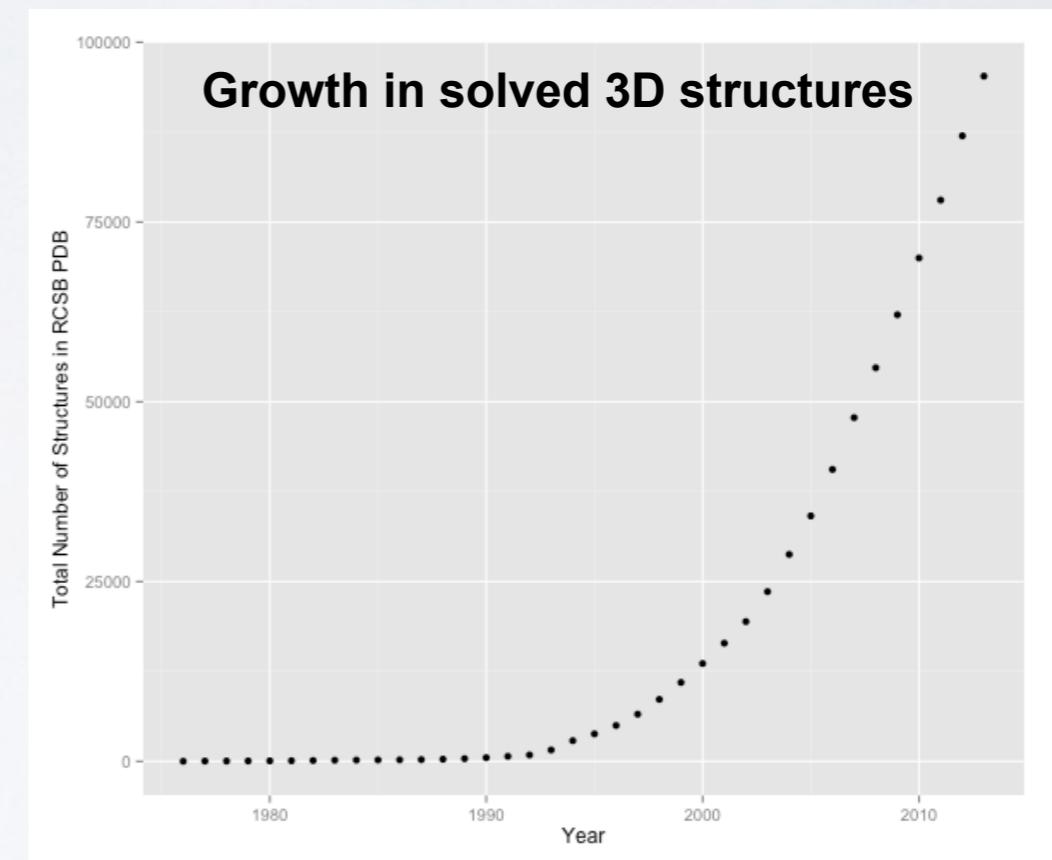


E.G. data from sequencing, structural genomics, proteomics, new high throughput assays, etc...

Why do we need Bioinformatics?

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

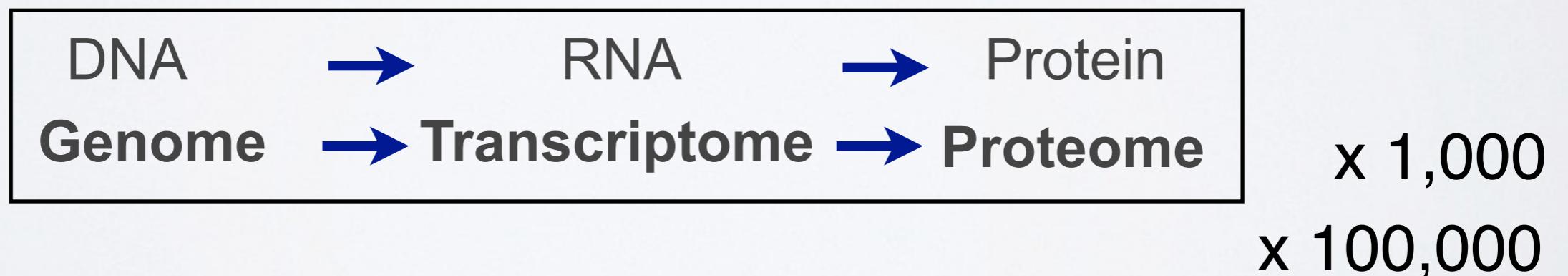
- Bioinformatics provides methods for the efficient:
 - **storage**
 - **annotation**
 - **search and retrieval**
 - **data integration**
 - **data mining and analysis**



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

How do we do Bioinformatics?

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



How do we *actually* do Bioinformatics?

Pre-packaged tools and databases

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

Advanced tool application & development

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

How do we *actually* do Bioinformatics?

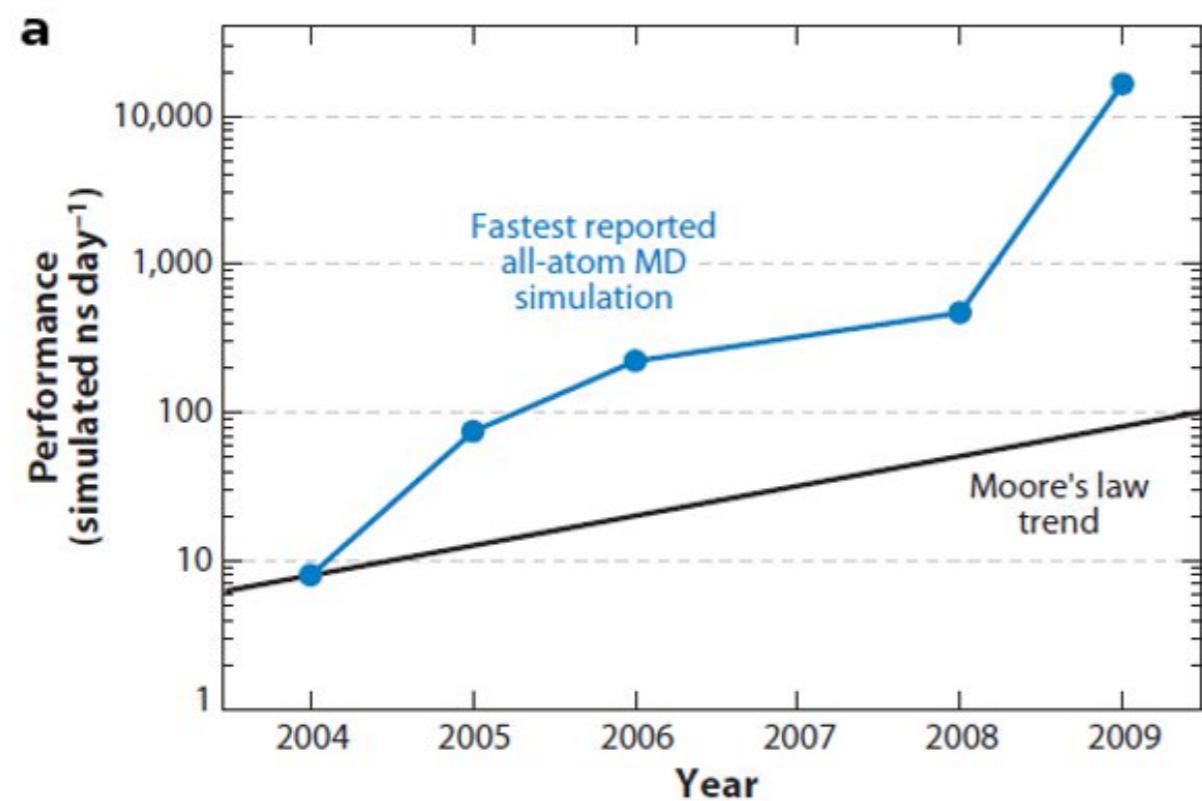
Pre-packaged tools and databases

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

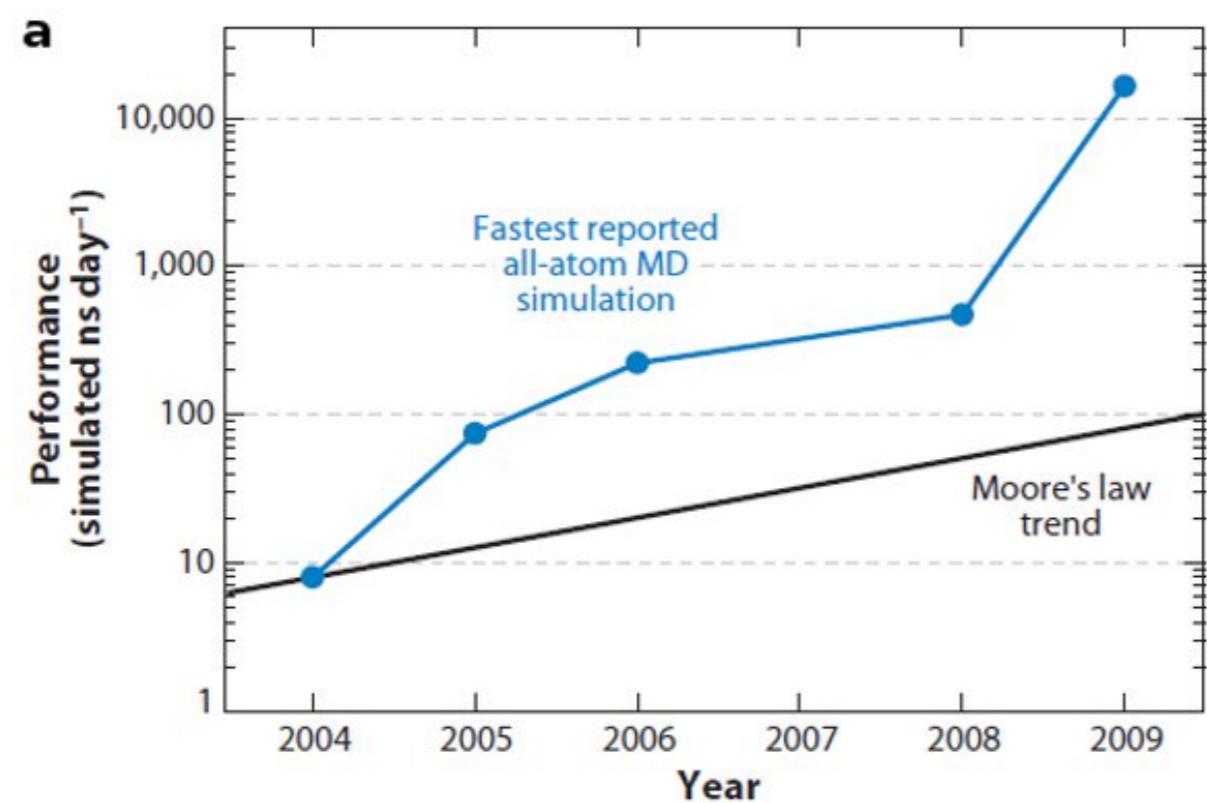
Advanced tool application & development

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

SIDE-NOTE: SUPERCOMPUTERS ANDGPUS



SIDE-NOTE: SUPERCOMPUTERS AND GPUS



HOW COMPUTERS HAVE CHANGED

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

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



NSF Extreme Science and Engineering Discovery Environment (XSEDE)

The screenshot shows a web browser window with the URL www.xsede.org/community-engagement/educator-pro. The page features a dark blue header with the XSEDE logo and navigation links for Home, Gmail, Gcal, Bitbucket, GitHub, News, Disqus, About, For Users, Ecosystem, Community Engagement (which is currently selected), News, XUP, and a search icon. The main content area has a background image of a star-filled galaxy. A large white title 'Curriculum and Educator Programs' is centered above a black sidebar. The sidebar contains the text 'XSEDE pursues innovation and collaboration in computational science education.' and a section titled 'Campus Visits' with descriptive text. To the right of the sidebar is a 'Key Points' section with a bulleted list and a 'Related Links' section with several blue hyperlinks.

Curriculum and Educator Programs

XSEDE pursues innovation and collaboration in computational science education.

Campus Visits

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

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

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

Key Points

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

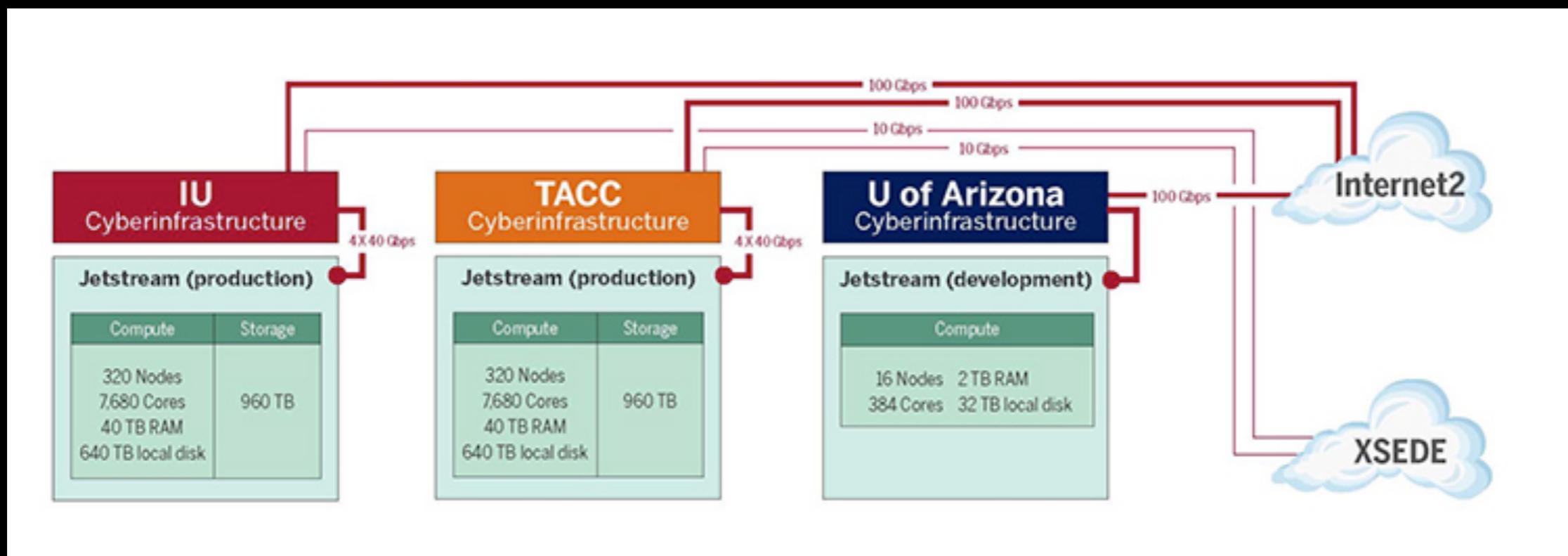
XSEDE campus visits emphasize the need for computational science education and offer guidance concerning course content

Related Links

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

What is *Jetstream*?

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



Jetstream tutorials

Developed *user friendly* labs for Jetstream basics

The screenshot shows a web browser window with two main pages visible.

Left Side (Course Navigation):

- UC San Diego logo
- BGGN 213
- A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD
- Overview
- Lectures (highlighted with a red border)
- Computer Setup
- Learning Goals
- Assignments & Grading
- Ethics Code
- Screen Cast Videos

Right Side (Tutorial Content):

Starting a Jetstream Computer Instance!

Here we describe the process of starting up and managing a [Jetstream](#) service virtual machine instance.

Note: Jetstream is a cloud-based on-demand virtual machine system funded by the National Science Foundation. It will provide us with computers (what we call “virtual machine instances”) that look and feel just like a regular Linux workstation but with thousands of times the computing power!

What we’re going to do here is walk through starting up and running computer (an “instance”) on the Jetstream service.

Below we walk through the process of starting up and accessing one of these instances. To begin with, just think of it like requesting and logging-in to a brand new remote computer. We have provided screenshots of the whole process that you can click on to see a larger version. The important areas to fill in are circled in red.

Note Some of the details may vary – for example, if you have your own XSEDE account, you may want to log in with that – and the name of the operating system or “image” may also vary from “Ubuntu 16.04” depending

Jetstream tutorials

Developed *user friendly* labs for Jetstream basics

The image shows a web browser window with two tabs open. The top tab is titled "bioboot.github.io/bggn213_f17/jetstream/boot/" and shows the UC San Diego BGNN 213 course website. The bottom tab is also titled "bioboot.github.io/bggn213_f17/jetstream/boot/" and shows the Jetstream application login page.

Request to log in to the Jetstream Portal

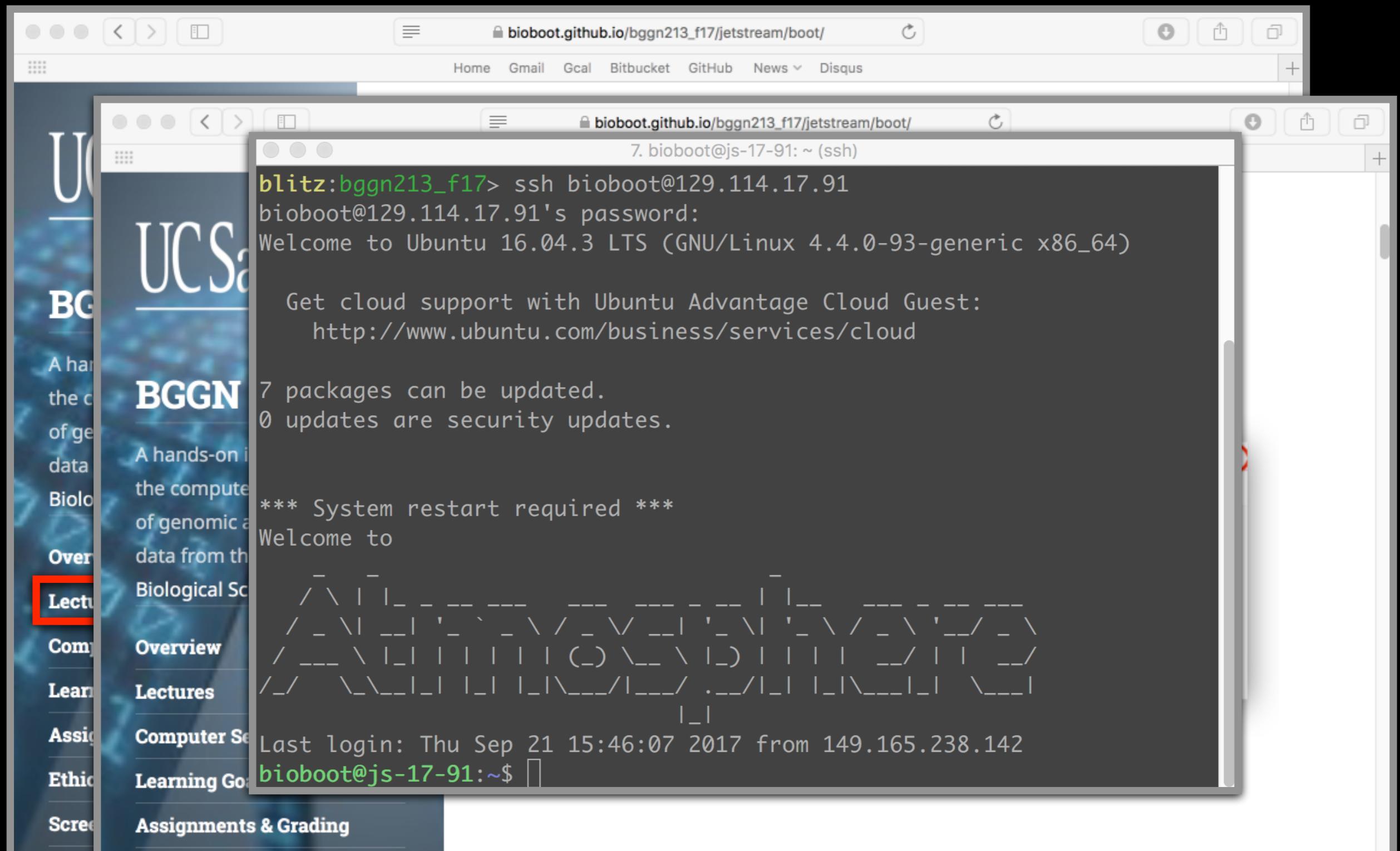
First, go to the Jetstream application at:
<https://use.jetstream-cloud.org/application>.

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

The Jetstream application login page features a red "Login" button in the top right corner, which is circled in red. The page includes a search bar, a "Image Search" section showing 93 images, and a "Featured Images" section displaying two thumbnail cards: "R with Intel compilers (CentOS 7)" and "R with GCC (CentOS 7)".

Jetstream tutorials

Developed *user friendly* labs for Jetstream basics



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; Treasure Trove or Trivial Pursuit*” states:

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

Key-Point: Avoid the miss-use of ‘black boxes’!

Common problems with Bioinformatics

Confusing multitude of tools available

- ▶ Each with many options and settable parameters

Most tools and databases are written by and for nerds

- ▶ Same is true of documentation - if any exists!

Most are developed independently

Notable exceptions are found at the:

- **EBI** (European Bioinformatics Institute) and
- **NCBI** (National Center for Biotechnology Information)

General Parameters

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

Scoring Parameters

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

Filters and Masking

Filter	<input type="checkbox"/> Low complexity regions ?
Mask	<input type="checkbox"/> Mask for lookup table only <input type="checkbox"/> Mask lower case letters ?

PSI/PHI/DELTA BLAST

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

Even Blast has many settable parameters

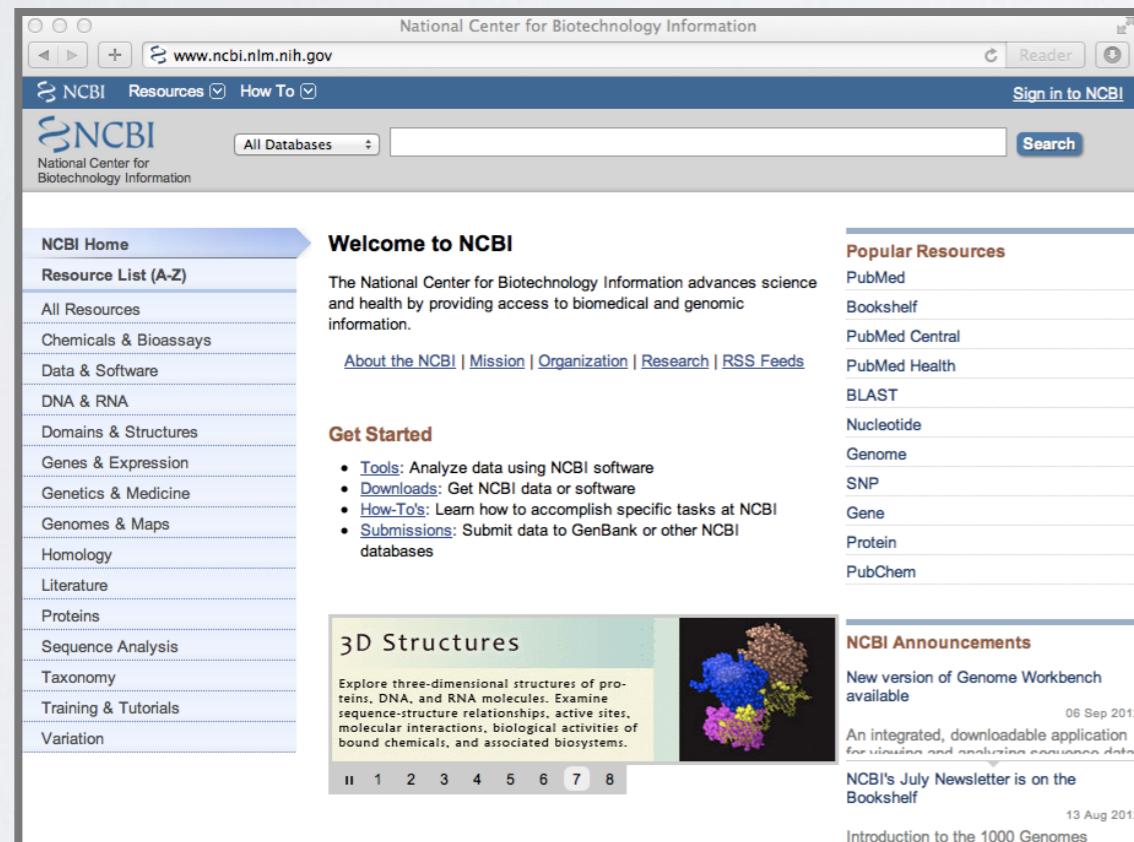
STEP 3 - Set your PROGRAM FASTA

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

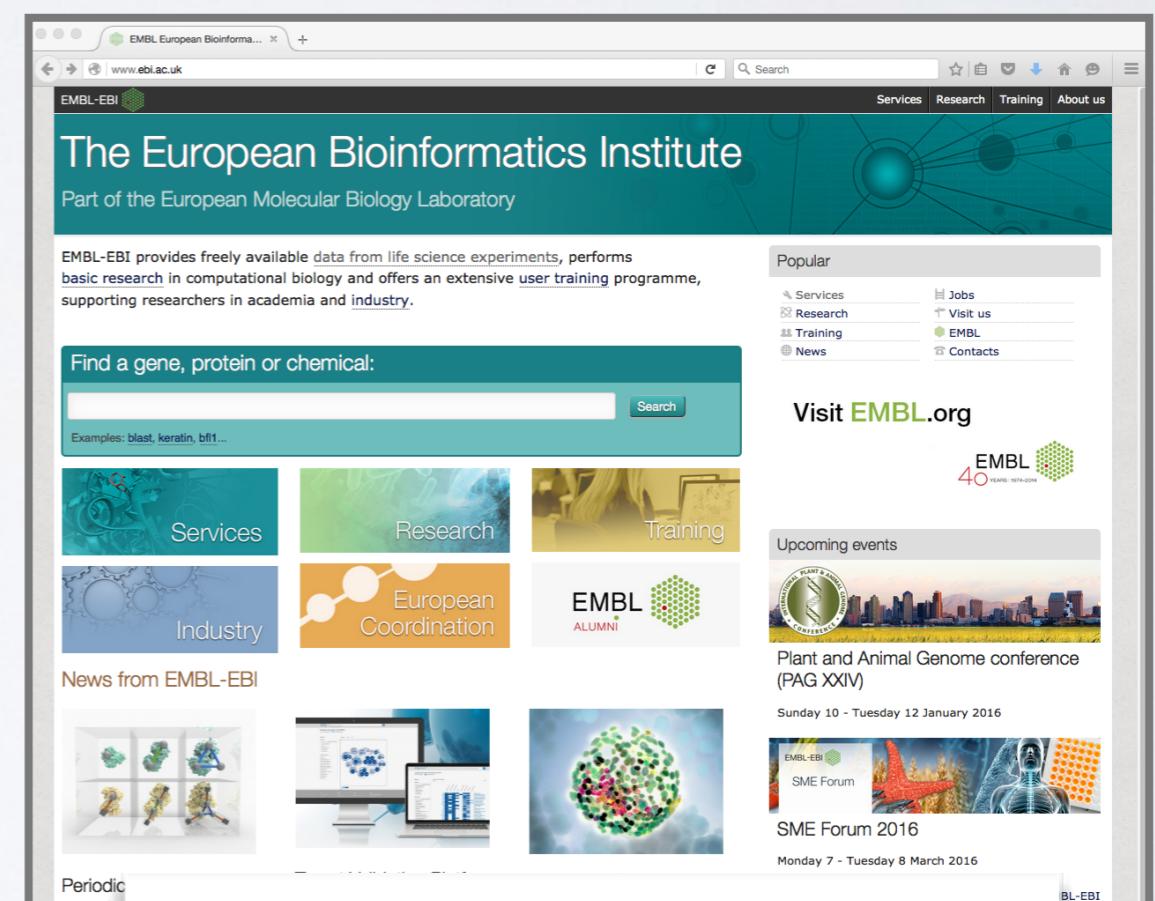
Related tools with different terminology

Key Online Bioinformatics Resources: NCBI & EBI

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



The screenshot shows the homepage of the National Center for Biotechnology Information (NCBI). The top navigation bar includes links for "NCBI Resources" and "How To". The main content area features a "Welcome to NCBI" section with a brief introduction and links to "About the NCBI", "Mission", "Organization", "Research", and "RSS Feeds". Below this is a "Get Started" section with links to "Tools", "Downloads", "How-To's", and "Submissions". A "3D Structures" section displays a molecular model. On the right, there's a sidebar titled "Popular Resources" listing links to PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. At the bottom, there are sections for "NCBI Announcements" (mentioning a new version of Genome Workbench) and "News from NCBI" (mentioning the July Newsletter and the 1000 Genomes Project).



The screenshot shows the homepage of the European Bioinformatics Institute (EMBL-EBI). The top navigation bar includes links for "Services", "Research", "Training", and "About us". The main content area features a large search bar with the placeholder "Find a gene, protein or chemical:" and a "Search" button. Below the search bar are several colored boxes representing different services: "Services" (blue), "Research" (green), "Training" (yellow), "Industry" (blue), "European Coordination" (orange), and "EMBL ALUMNI" (grey). To the right, there's a "Popular" section with links to "Services", "Research", "Training", and "News". Below this is a "Visit EMBL.org" section featuring the EMBL logo and a "40 years" anniversary graphic. Further down are sections for "Upcoming events" (listing the "Plant and Animal Genome conference (PAG XXIV)" and "SME Forum 2016"), and "News from EMBL-EBI" (showing images of molecular structures and data analysis tools).

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

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

National Center for Biotechnology Information (NCBI)

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



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

National Center for Biotechnology Information

NCBI Resources How To Sign in to NCBI

All Databases Search

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

Welcome to NCBI

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

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

Get Started

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

3D Structures

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

Popular Resources

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

NCBI Announcements

New version of Genome Workbench available 06 Sep

An integrated, downloadable applicati

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

National Center for Biotechnology Information

NCBI Resources How To Sign in to NCBI

All Databases

Search

Popular Resources

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

Welcome to NCBI

The National Center for Biotechnology Information provides access to unique information, tools and resources in molecular biology, genetics and health by providing access to its databases and information.

About the NCBI | Mission | Our History

Get Started

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

3D Structures

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

New version of Genome Workbench available

06 Sep

An integrated, downloadable application

A screenshot of the NCBI homepage. The sidebar on the left lists various resources like NCBI Home, Resource List (A-Z), and 3D Structures. The main content area features a 'Welcome to NCBI' section and a 'Get Started' list. On the right, there's a 'Popular Resources' sidebar with links to PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. A red bracket on the right side of the sidebar groups the links for BLAST, SNP, Gene, Protein, and PubChem, with red arrows pointing from the bracket to each of these five links.

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

National Center for Biotechnology Information

NCBI Resources How To Sign in to NCBI

All Databases Search

NCBI Home Resource List (A-Z)

Welcome to NCBI
The National Center for Biotechnology Information advances science

Popular Resources PubMed

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

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

Homology Literature Proteins Sequence Analysis Taxonomy Training & Tutorials Variation

databases

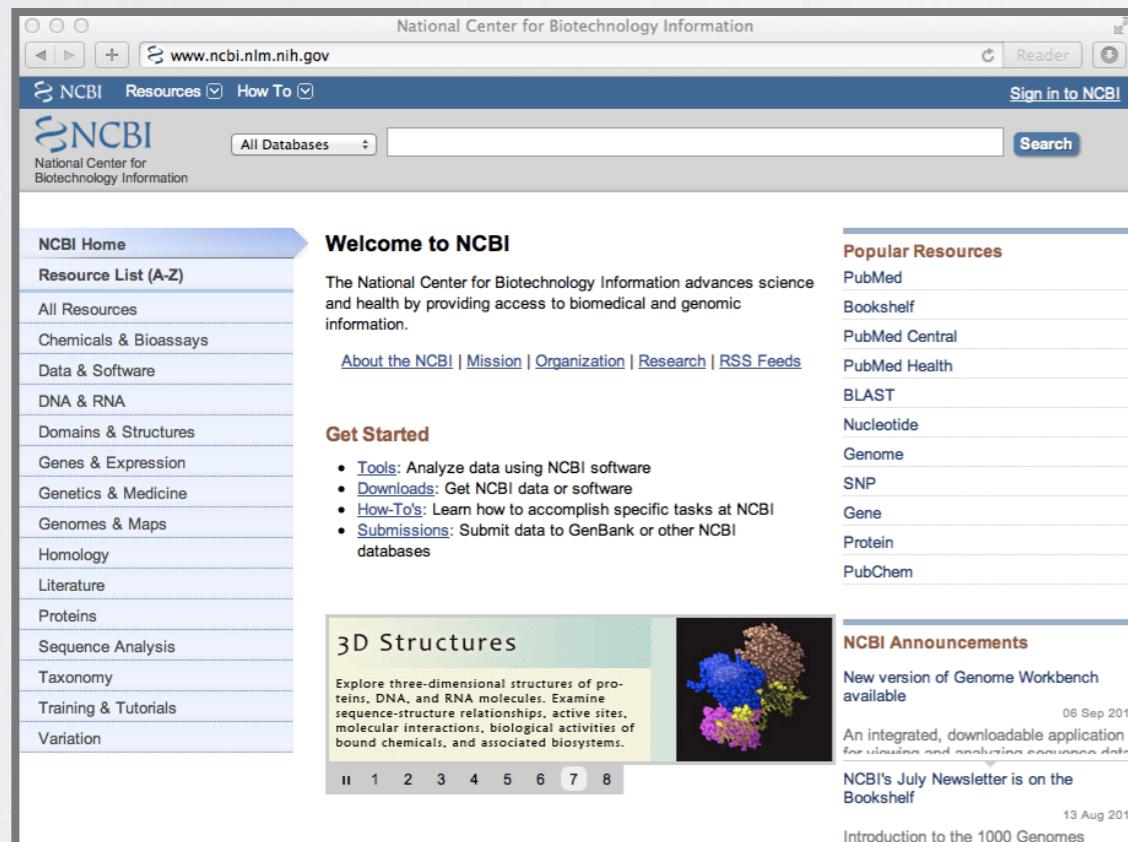
3D Structures
Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of bound chemicals and associated biosystems

Protein PubChem

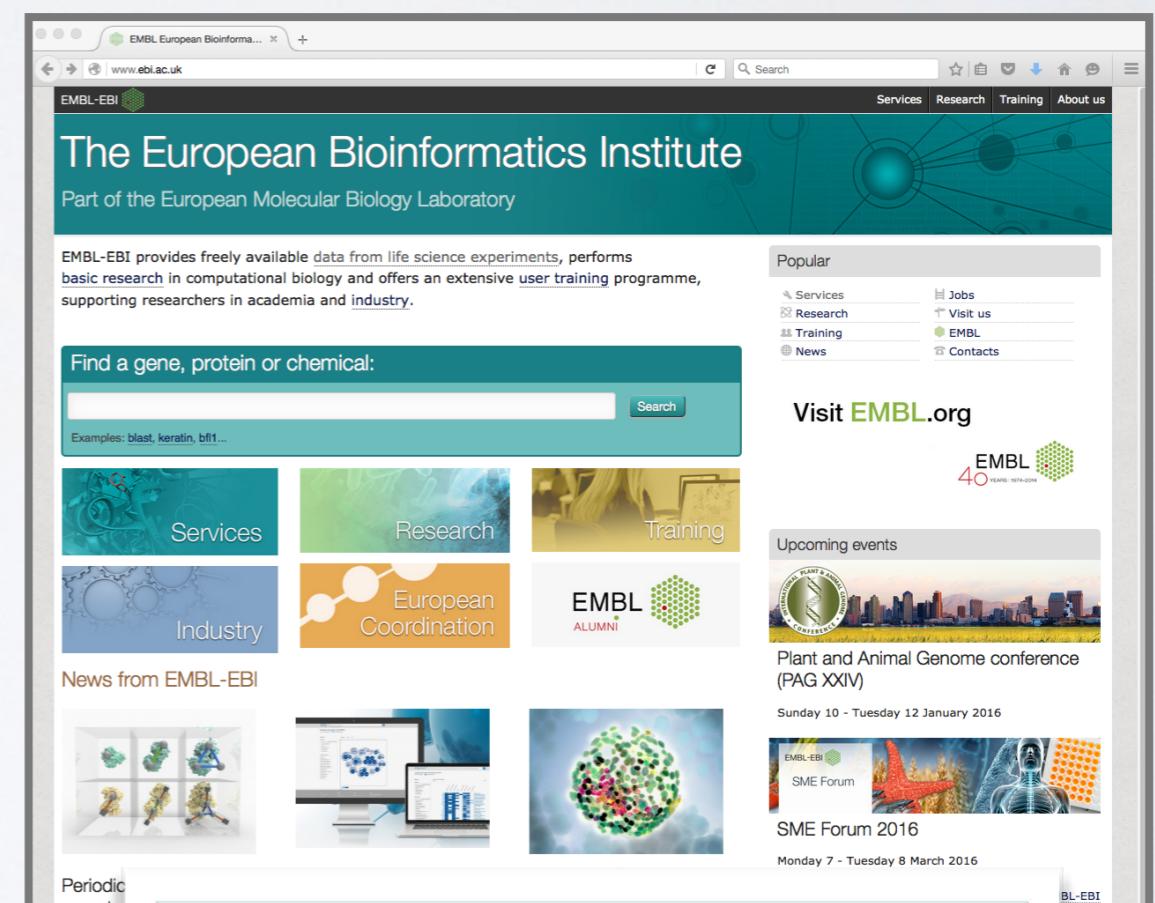
NCBI Announcements
New version of Genome Workbench available 06 Sep
An integrated, downloadable applicati

Key Online Bioinformatics Resources: NCBI & EBI

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



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



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

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

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

European Bioinformatics Institute (EBI)

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



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

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

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

The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

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

Find a gene, protein or chemical:

Examples: blast, keratin, bfl1...

Search

Services

Research

Training

EMBL ALUMNI

Upcoming events

Plant and Animal Genome conference (PAG XXIV)

Sunday 10 - Tuesday 12 January 2016

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

The screenshot shows the EBI Services website (www.ebi.ac.uk/services) with a teal header and a banner featuring a molecular structure. The main content area displays nine service categories in a grid:

- 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

On the right, there's a "Popular" sidebar with links to Ensembl, UniProt, PDBe, ArrayExpress, ChEMBL, BLAST, Europe PMC, Reactome, Train online, and Support. Below the sidebar is a "Service news" section featuring a monarch butterfly on a DNA helix.

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

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

Services < EMBL-EBI

www.ebi.ac.uk/services

EMBL-EBI

Services | Research | Training | About us

Services

Overview | A to Z | Data submission | Support

Bioinformatics services

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

DNA & RNA
genes, genomes & variation

Gene expression
RNA, protein & metabolite expression

Proteins
sequences, families & motifs

Structures
Molecular & cellular structures

Systems
reactions, interactions & pathways

Chemical biology
chemogenomics & metabolomics

Ontologies
taxonomies & controlled vocabularies

Literature
Scientific publications & patents

Cross domain
cross-domain tools & resources

Popular

Ensembl

UniProt

PDB

ArrayExpress

ChEMBL

Training

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

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

Proteins

Popular services



UniProt: The Universal Protein Resource

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



InterPro

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



PRIDE: The Proteomics Identifications Database

An archive of protein expression data determined by mass spectrometry.



Pfam

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



Clustal Omega

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



HMMER - protein homology search

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



InterProScan 5

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

Quick links

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

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

The screenshot shows the homepage of the EMBL European Bioinformatics Institute (EBI) at www.ebi.ac.uk. The page features a dark teal header with the EMBL-EBI logo and navigation links for Services, Research, Training, and About us. Below the header is a large banner with the text "The European Bioinformatics Institute" and "Part of the European Molecular Biology Laboratory". A search bar is located above a main content area. The content area includes a "Find a gene, protein or chemical:" search box with examples like "blast, keratin, bfl1...". To the right of the search box is a "Popular" sidebar with links to Services, Research, Training, News, Jobs, Visit us, EMBL, and Contacts. Below the search box are several colored boxes: Services (blue), Research (green), Training (yellow, highlighted with a red border), European Coordination (orange), Industry (grey-blue), and EMBL ALUMNI (white). A "News from EMBL-EBI" section is also present. On the right side, there's a "Visit EMBL.org" section with the EMBL 40th anniversary logo, an "Upcoming events" section featuring the "Plant and Animal Genome conference (PAG XXIV)" with a city skyline background, and a decorative footer banner.

EMBL European Bioinforma...

www.ebi.ac.uk

Search

Services | Research | Training | About us

The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

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

Find a gene, protein or chemical:

Search

Examples: blast, keratin, bfl1...

Services

Research

Training

European Coordination

Industry

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

EMBL-EBI

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

The screenshot shows a web browser displaying the EBI Training online course page. The URL in the address bar is www.ebi.ac.uk/training/online/course/using-sequence-similarity-searching-tools-embl-ebi. The page title is "Using sequence similarity searching tools at EMBL-EBI: webinar". The main content area shows a thumbnail of the webinar video, which features a blue background with white text and a portrait of a man. Below the video thumbnail, there is a play button and a timestamp of "0:00 / 37:42". To the right of the video, there is a sidebar with a "Popular" section containing links to "Train online", "Find us", and "Funding". Another sidebar on the right is titled "Find us at..." and lists links to "Open days and career days", "Conference exhibitions", "EMBL courses and events", "Genome campus events", and "Science for schools". The top navigation bar includes links for "Services", "Research", "Training" (which is highlighted), and "About us". The header also features the EMBL-EBI logo and a search bar.

Using sequence similarity searching tools at EMBL-EBI: webinar

Using sequence similarity searching tools at EMBL-EBI: webinar

Using sequence similarity search tools at EMBL-EBI

Finding homologous sequences with BLAST, FASTA, PSI-Search etc.

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

EMBL-EBI

0:00 / 37:42

This webinar focuses on how to use tools like **BLAST** and PSI-Search to find homologous sequences in EMBL-EBI databases, including tips on which tool and database to use, input formats, how to change parameters and how to interpret the results pages.

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

A screenshot of a web browser displaying the EBI Train online website. The address bar shows the URL www.ebi.ac.uk/training/online/. The page header includes the EMBL-EBI logo and links for Train online, Find, Help, and Feedback. A red 'Beta' badge is visible in the top right corner. The main navigation menu at the top has options for Databases, Tools, Research, Training, Industry, About Us, and Help. A secondary navigation bar on the left is titled 'Navigation' and includes a link to 'Train online Home'. The main content area features a large heading 'Train online'.

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, BiolImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlilesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, KloTho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5 Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, MycDB, NDB, NRSub, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-MODEL Repository, SWISS-PROT, TeIDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc ..!!!!

Bioinformatics Databases

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

Course Logistics

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

Learning Objectives

What you need to learn to succeed in this course.

Course Structure

Major lecture topics and specific learning goals.

Introduction to Bioinformatics

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

Bioinformatics Database

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

Your Turn!

https://bioboot.github.io/bggn213_S18/lectures/#1

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.
- The goals of the hands-on session is to introduce a range of core bioinformatics databases and associated online services whilst actively investigating the molecular basis of several common human disease.

Material:

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

Homework:

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

BGGN-213: FOUNDATIONS OF BIOINFORMATICS (Lecture 1)

Bioinformatics Databases and Key Online Resources

https://bioboot.github.io/bggn213_S18/lectures/#1

Dr. Barry Grant

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

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

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

Section 1

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

```
>example1
ATGGTGCATCTGACTCCTGTGGAGAAGTCTGCCGTTACTGCCCTGTGGGCAAGGTGAACGTGGATGAAG
TTGGTGGTGAGGCCCTGGCAGGCTGCTGGTGTACCCCTGGACCCAGAGGTTCTTGAGTCCTTGG
GGATCTGTCCACTCCTGATGCAGTTATGGCAACCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTAGTGTGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTGCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACCTCAGGCTCCTGGCAACGTGCTGGTCTGTGTGCTGGCCA
TCACCTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCCCACAAGTATCACTAAGCTCGCTTCTGCTGTCCAATTT
```

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:

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

End times:

[2:35 pm]

2. GENE database @ **NCBI**

[2:55 pm]

— BREAK —

— 3:10 pm —

3. UniProt & Muscle @ **EBI**

[3:30 pm]

4. PFAM, PDB & NGL

[4:00 pm]

— BREAK —

— 4:10 pm —

5. Extension exercises

[4:40 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.
- 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

<http://thegrantlab.org/bggn213/>

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

THANK YOU