



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

YUANSHENG

yuz461@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 with the URL bioboot.github.io/bggn213_S18/ in the address bar. The page content is as follows:

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

Social Media: Twitter, GitHub, Email, RSS

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/ in the address bar. 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 (This link is highlighted with a red box)

Assignments & Grading

Ethics Code

Screen Cast Videos

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.

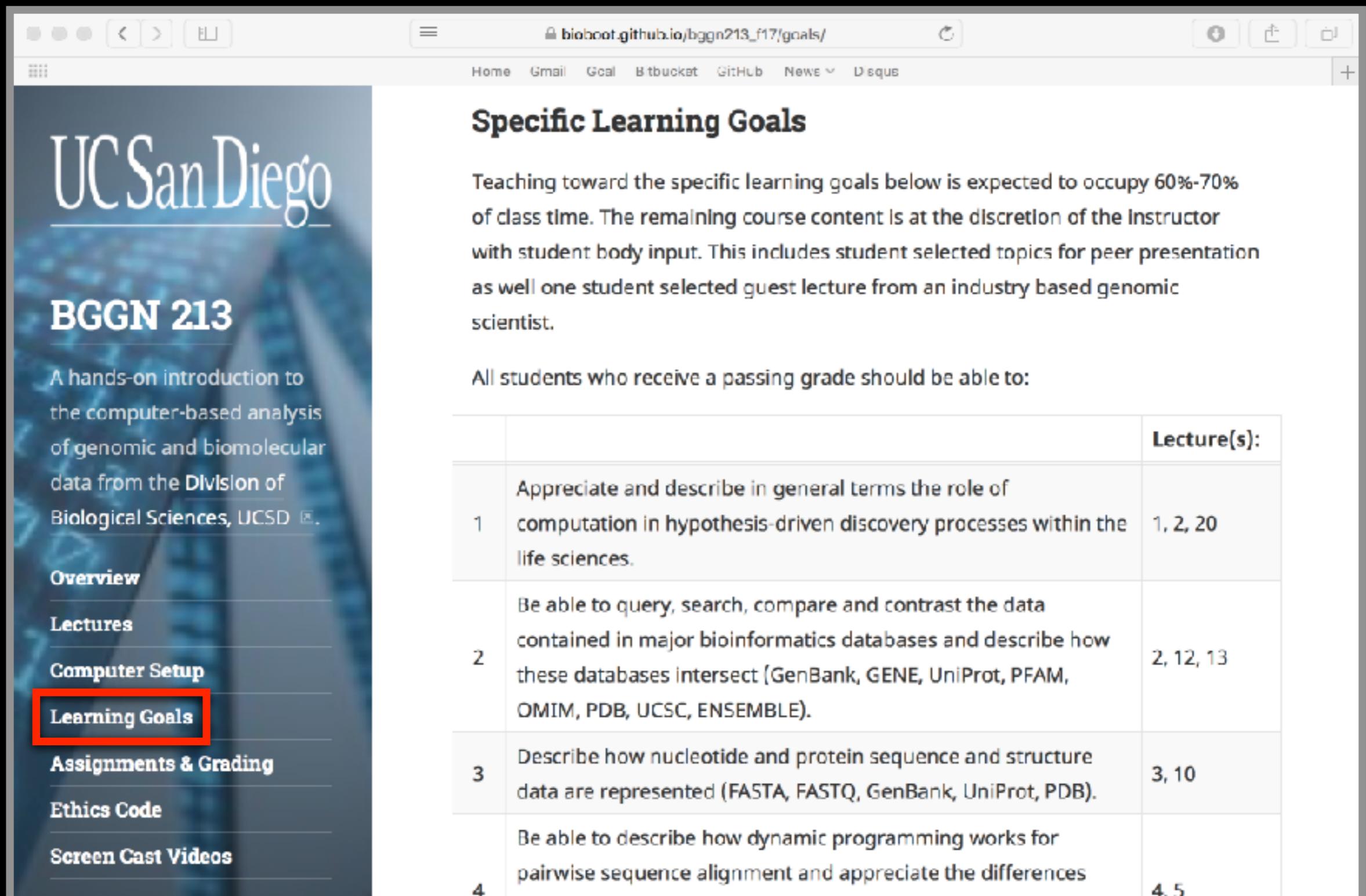
At the end of this course students will:

- 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 "Specific Learning Goals". The main content area describes the goals and lists four numbered learning objectives with their corresponding lectures.

Specific Learning Goals

Teaching toward the specific learning goals below is expected to occupy 60%-70% of class time. The remaining course content is at the discretion of the Instructor with student body input. This includes student selected topics for peer presentation as well one student selected guest lecture from an industry based genomic scientist.

All students who receive a passing grade should be able to:

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

Course Structure

Derived from specific learning goals

The screenshot displays a web page for a course structure. On the left, there is a sidebar with the UC San Diego logo and course information:

- BGGN 213**
- A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.
- Navigation links: Overview, Lectures (highlighted with a red box), Computer Setup, Learning Goals, Assignments & Grading, Ethics Code.
- Social media icons: Twitter, GitHub, Email, RSS.

The main content area is titled "Lectures". It contains a table with the following data:

#	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:

- Address Bar:** bioboot.github.io/bggn213_f17/lectures/#1
- Navigation:** Home, Gmail, Gcal, Bitbucket, GitHub, News, Disqus
- Left Sidebar (Course Information):**
 - 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.
 - Links: Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, Screen Cast Videos.
- Content Area:**
 - 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...)
 - Barry Grant
UC San Diego
<http://thegrantlab.org/bggn213>

Page Footer: Welcome to BGGN-213: Course introduction and logistics.

Homework

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

The screenshot shows a web browser window with the URL bioboot.github.io/bggn213_f17/lectures/#1. The main content is organized into sections:

- Homework:**
 - [Questions](#) (highlighted with a red box)
 - Readings:
 - PDF1: [What is bioinformatics? An introduction and overview](#)
 - PDF2: [Advancements and Challenges in Computational Biology](#)
 - Other: [For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights](#) (New York Times, 2014).
- Screen Casts:**
 - Welcome to “Foundations of Bioinformatics” (BGGN-21...)

At the bottom of the page, there is a footer with the text "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, there's a navigation bar with icons for back, forward, and search, followed by the URL 'docs.google.com/forms/d/e/1FAIpQLSeN3pg-AaRgBla3PkZuqSj' and a refresh button. Below the URL is a horizontal bar with links to 'Home', 'Gmail', 'Gcal', 'Bitbucket', 'GitHub', 'News', 'Disqus', and a plus sign for new tabs.

The main content area has a purple header bar with the title 'BGGN213 Lecture 1 Homework (F17)'. Below the header, the text 'Please answer the following questions' is displayed. A red asterisk indicates that the next field is required.

Your UCSD username/email address *
The first part of your UCSD email address before the '@ucsd.edu' part

Your answer

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

- Windows
- iOS
- Unix
- Perl

Which of the following databases contains primarily protein

Homework

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

docs.google.com/forms/d/e/1FAIpQLSeN3pg-AaRgBla3PxZuqSj

Home Gmail Gcal Bitbucket GitHub News Disqus

BGGN213 Lecture 1 Homework

Please answer the following questions

* Required

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

Homework is due before the next weeks class!

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

The screenshot shows a web browser window with the URL bioboot.github.io/bggn213_f17/goals/. The page is titled "BGGN 213". On the left, there's a sidebar with 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 displays 12 learning goals, each with a number, a description, and a list of associated numbers. Goals 6, 7, 8, and 9 are grouped together and highlighted with a green border. A red arrow points downwards from the bottom right corner of the page.

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 them.	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

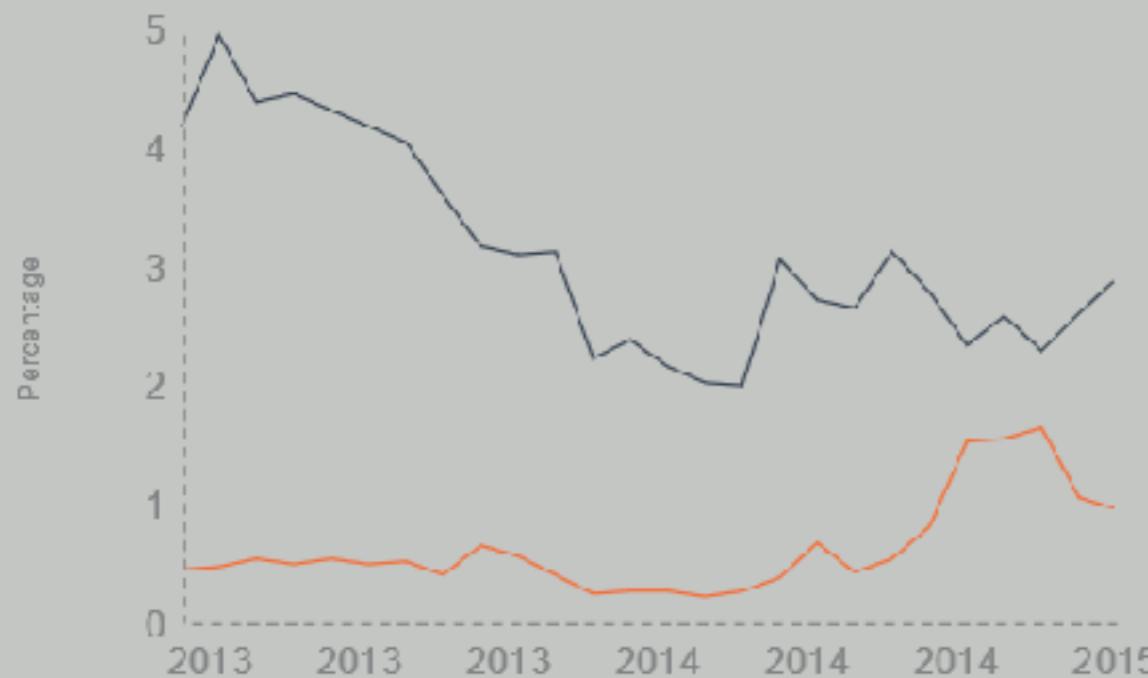
Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

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

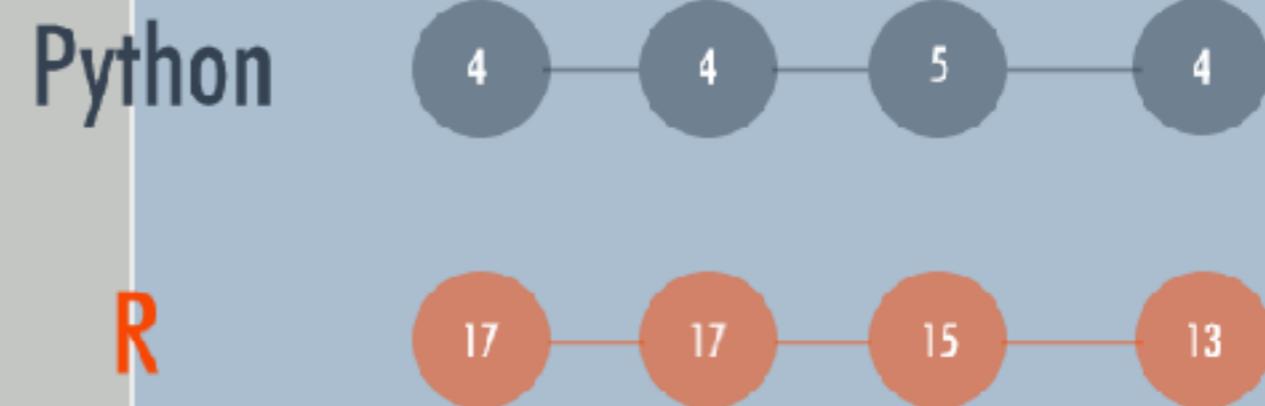
R and Python: The Numbers

Popularity Rankings

R and 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)



Jobs And Salary?

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



\$ 115,531



\$ 94,139

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

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

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

The screenshot shows the DataCamp homepage with a red circle highlighting the notifications icon in the top right corner of the header bar.

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 (3)

- 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 web browser displaying a DataCamp course page. The title of the page is "What is an IDE anyway?". On the left, there's a sidebar with a list of "Possible Answers" and a "Take Hint (-15xp)" button. One answer, "Integrated Development Environment", is circled in red. At the bottom of the sidebar is a large red circle around the "Submit Answer" button. The main content area shows an RStudio IDE interface. The console tab displays 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)  
  
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.
```

The RStudio interface includes tabs for Environment, History, Import Dataset, Global Environment, Files, Plots, Packages, Help, and Viewer. The Environment tab shows "Environment is empty".

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

The screenshot shows a web browser window for DataCamp's "What is an IDE anyway?" course. The main content area displays a completed exercise titled "Exercise Completed". The exercise summary states: "RStudio is an IDE that makes it easier to use by combining a set of tools into a single environment. Nice job! Move onto the next video to start learning more about the RStudio IDE!" A blue button labeled "50xp" indicates the completion reward. Below this, a "Possible Answers" section lists "PRESS ENTER TO" followed by a "Continue" button, which is highlighted with a red oval. To the right, the RStudio interface is visible, showing the console output of R version 3.3.1, the environment pane (empty), and the files pane (also empty). A sidebar on the left provides keyboard shortcuts for the RStudio interface.

What is an IDE anyway?

Exercise Completed

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

Nice job! Move onto the next video to start learning more about the RStudio IDE!

Possible Answers

PRESS ENTER TO Continue

Become a power user!

Submit Answer **Ctrl + Shift + Enter**

Environment See all keyboard shortcuts

Take Hint (-15xp)

Submit Answer

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

R

File Edit Code View Plots Session Build Debug Profile Tools Help barryus Sessions Project: (None) R 3.3 Environment History Import Dataset List Global Environment

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

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

Natural language support but running in an English locale

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

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

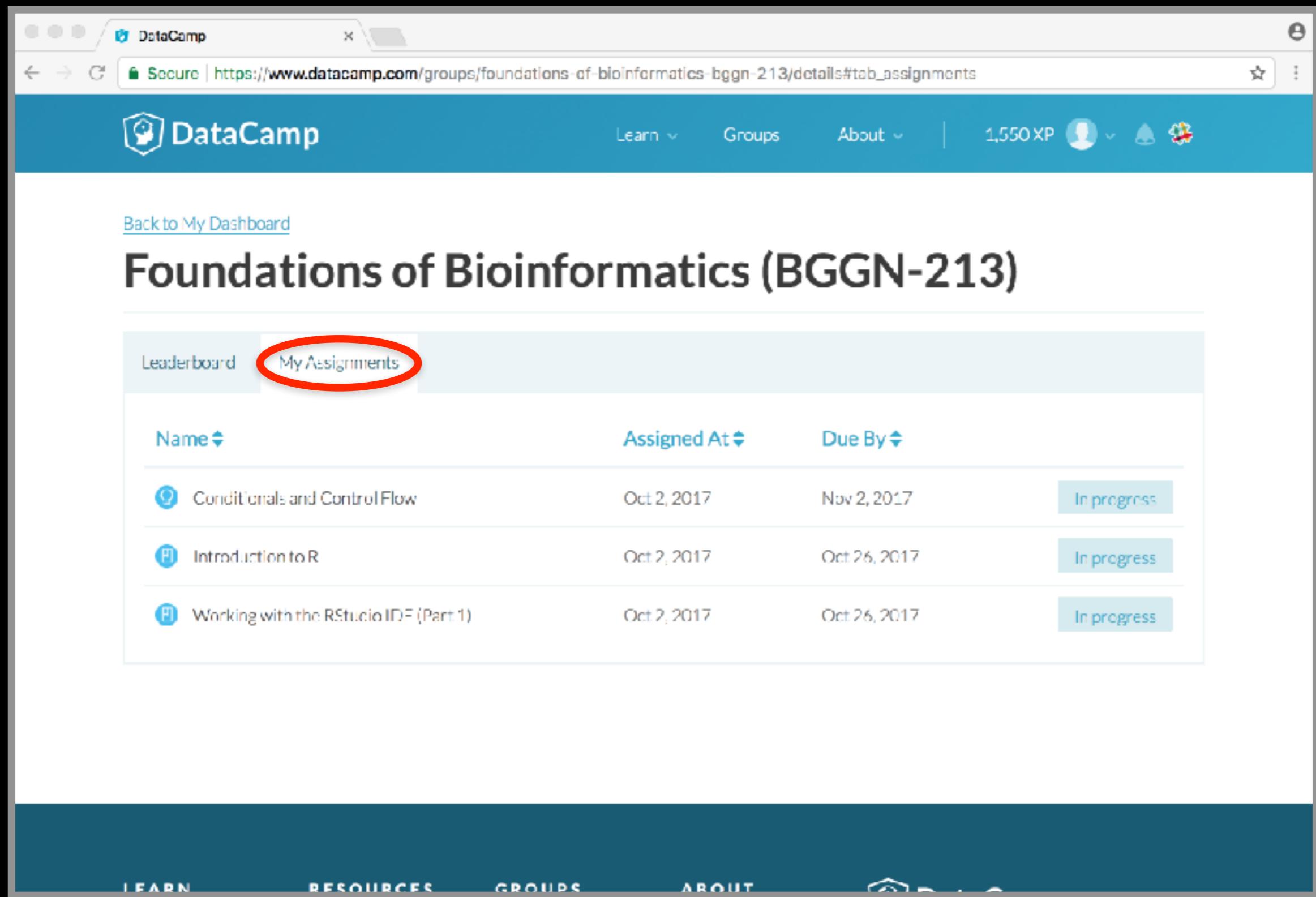
> |

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

The screenshot shows a web browser window for DataCamp. The URL in the address bar is <https://www.datacamp.com/groups/foundations-of-bioinformatics-bggn-213/details>. The page title is "Foundations of Bioinformatics (BGGN-213)". The top navigation bar includes links for "Learn", "Groups" (which is highlighted with a red circle), and "About". A user profile icon indicates 1,550 XP. Below the title, there are two tabs: "Leaderboard" (selected) and "My Assignments". Under the "Leaderboard" tab, there are three time filters: "30 Days" (selected), "90 Days", and "Last Year". A table displays the top 8 members of the group, showing their rank, 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 Pagluso	12085	2	11
5 oehernan	11055	2	10
6 Erin Schikaris	10350	2	9
7 Zachary Warburg	9110	1	8
8 Alexander Weitzel	6950	1	6

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



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#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. A table follows, listing three assignments with columns for Name, Assigned At, Due By, and status (In progress). The assignments are:

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", along with social media icons.

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...

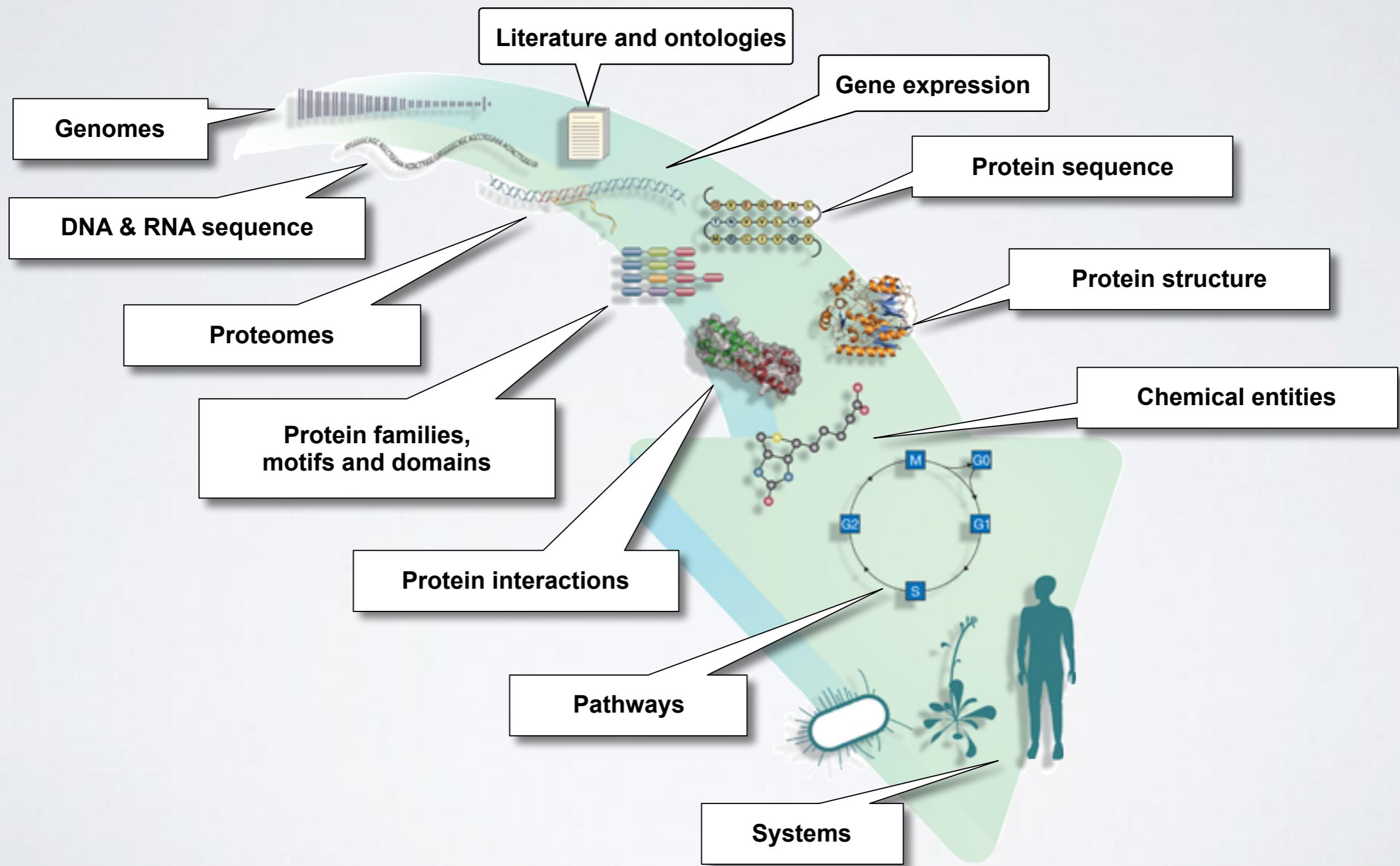
MORE DEFINITIONS

- ▶ “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>)

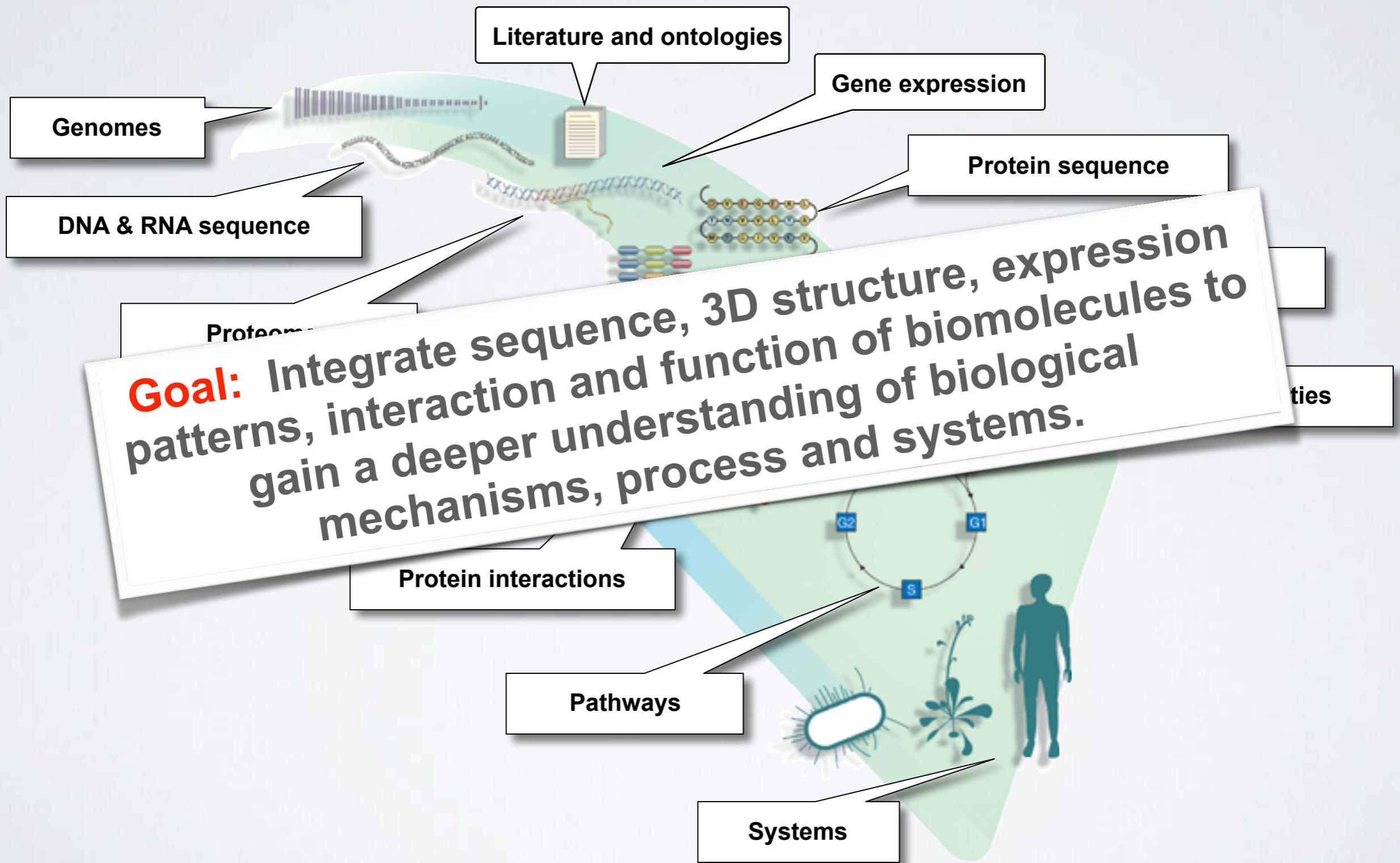
MORE DEFINITIONS

- ▶ “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*

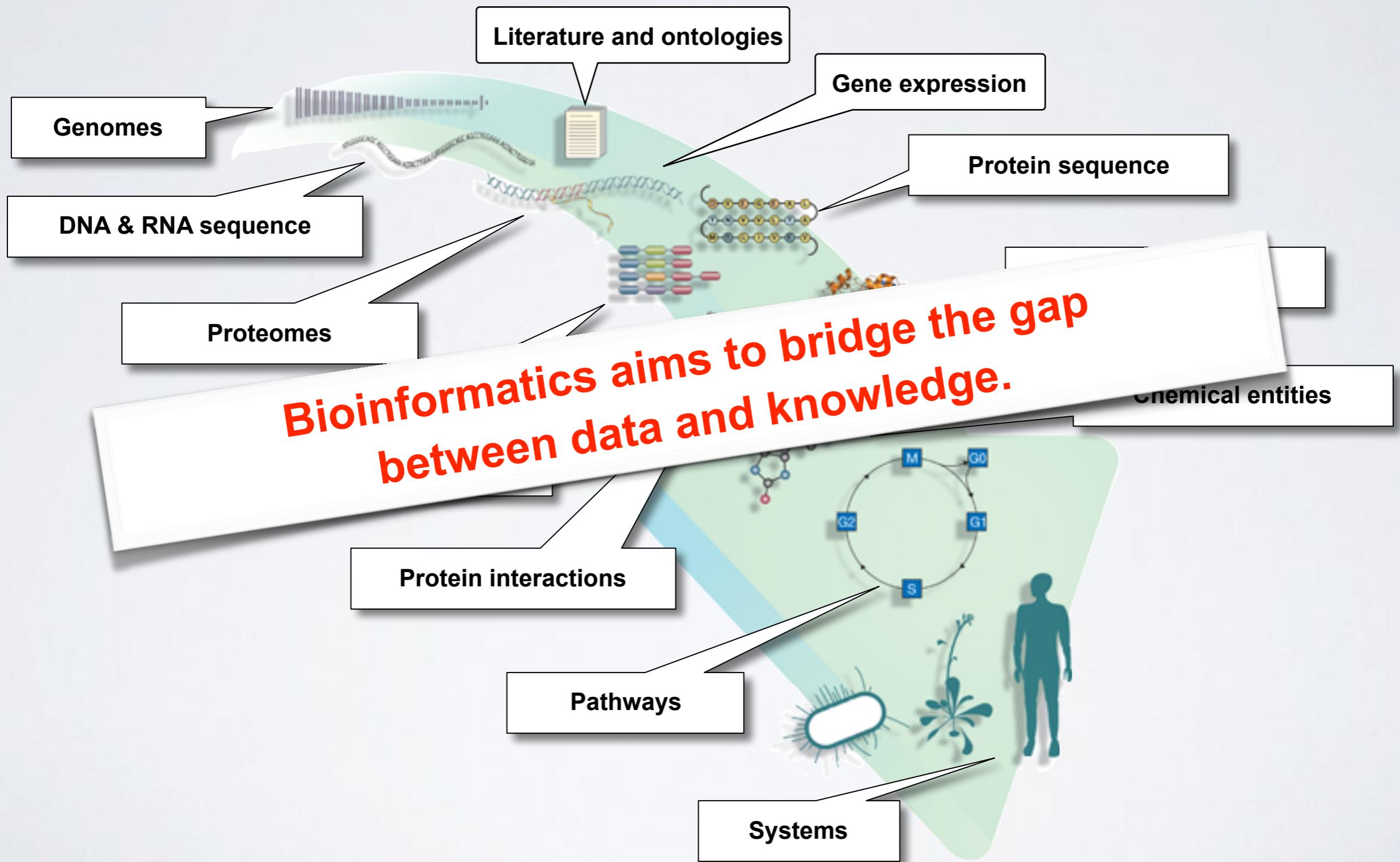
Major types of Bioinformatics Data



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

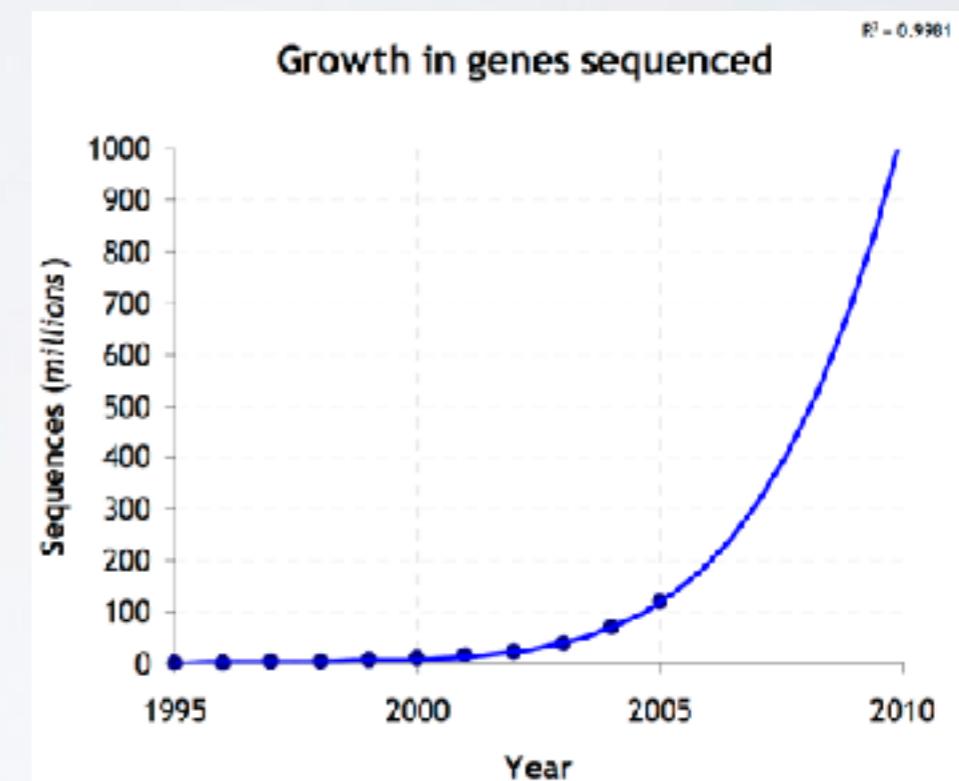
- *DNA sequence determines protein sequence.*
- *Protein sequence determines protein structure.*
- *Protein structure determines protein function.*
- *Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.*

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**

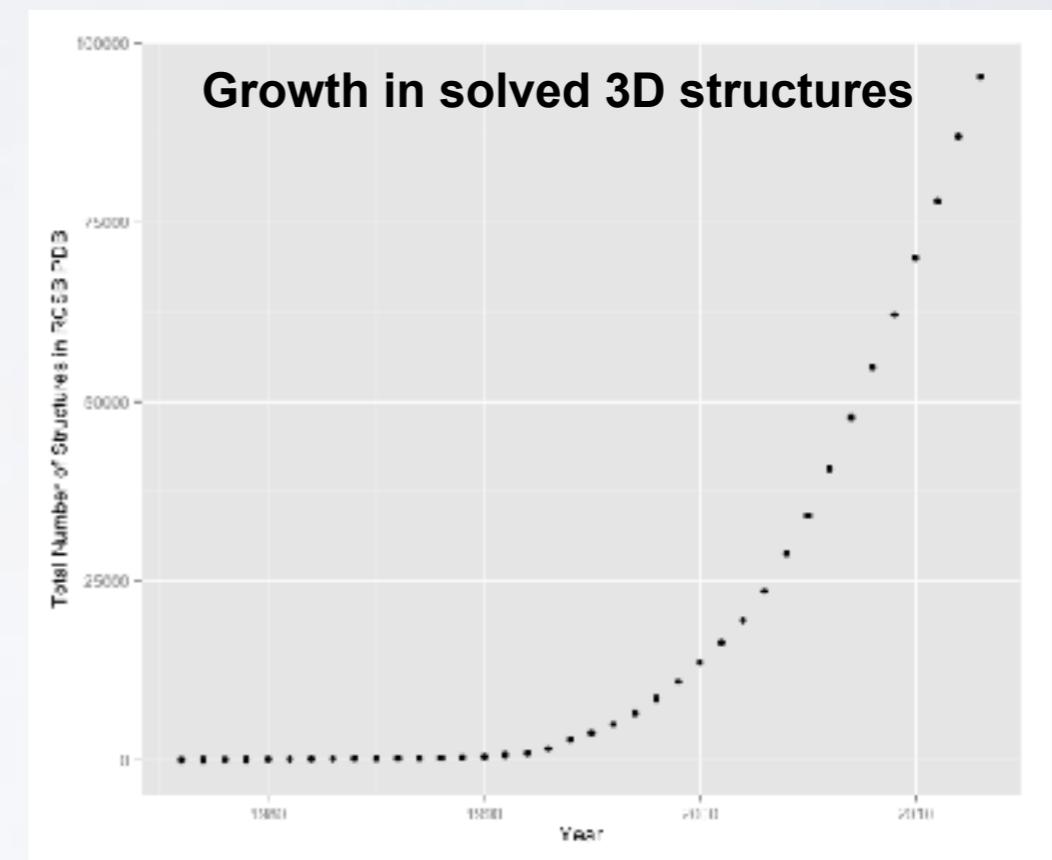


E.G. data from sequencing, structural genomics, proteomics, new high throughput assays, etc...

Why do we need Bioinformatics?

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

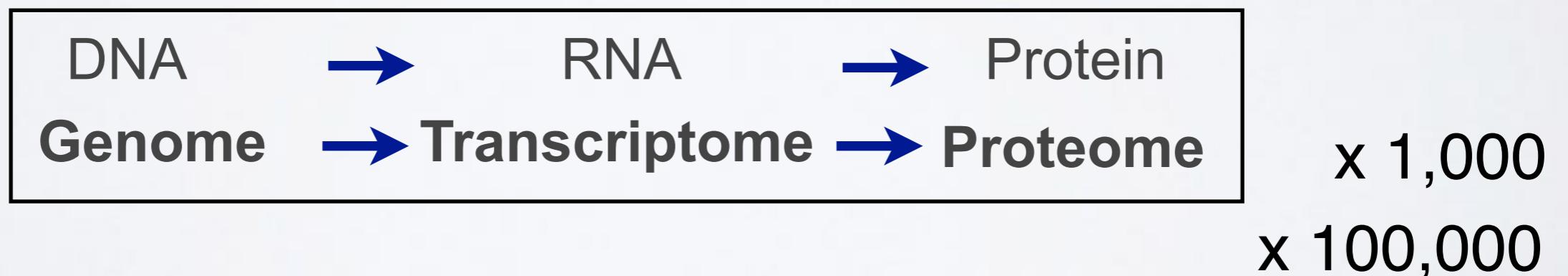
- Bioinformatics provides methods for the efficient:
 - ▶ **storage**
 - ▶ **annotation**
 - ▶ **search and retrieval**
 - ▶ **data integration**
 - ▶ **data mining and analysis**



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?

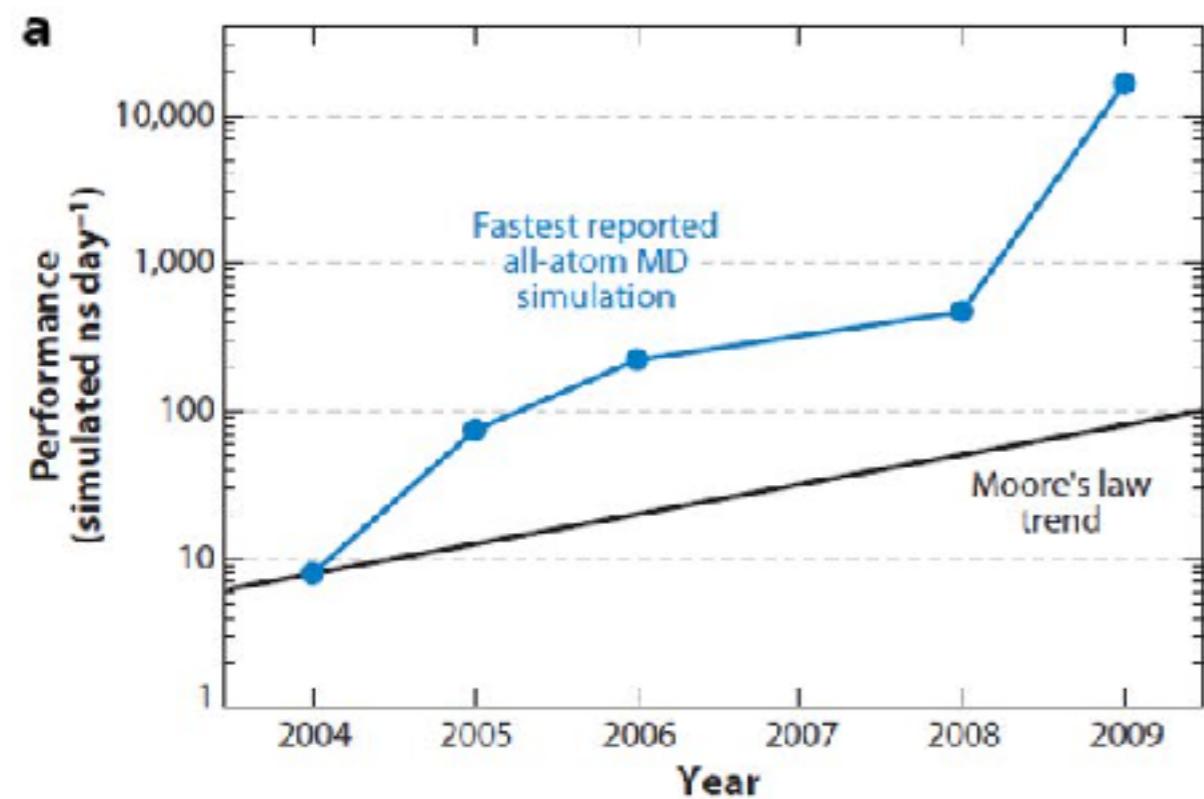
Pre-packaged tools and databases

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

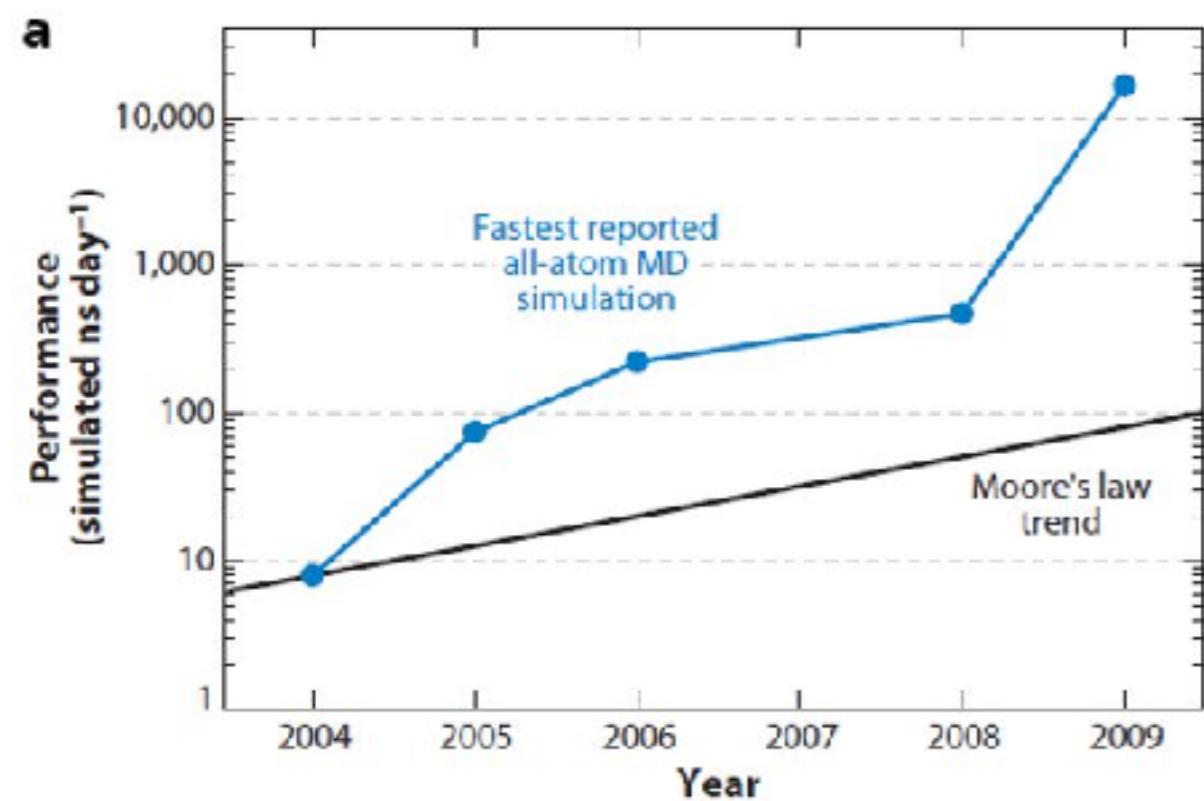
Advanced tool application & development

- 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	\$10M	0.1 MHz	1 MB	WALL
2013	\$14,000	1 GHz	10 GB	LAPTOP
CHANGE	10,000	10,000	10,000	10,000

If cars were like computers then a new Vehc
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-pr. The page features a dark blue header with the XSEDE logo and navigation links for Home, Gmail, Gcal, Bitbucket, GitHub, News, and Disqus. A secondary navigation bar includes About, For Users, Ecosystem, Community Engagement (which is highlighted in black), 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. Below it, a dark text box contains the text: 'XSEDE pursues innovation and collaboration in computational science education.' To the right, a 'Key Points' section lists four bullet points about campus visits. At the bottom right, there's a 'Related Links' section with four blue links: Diversity and Inclusion, Student Engagement, Campus Champions, and XSEDE Scholars Program.

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

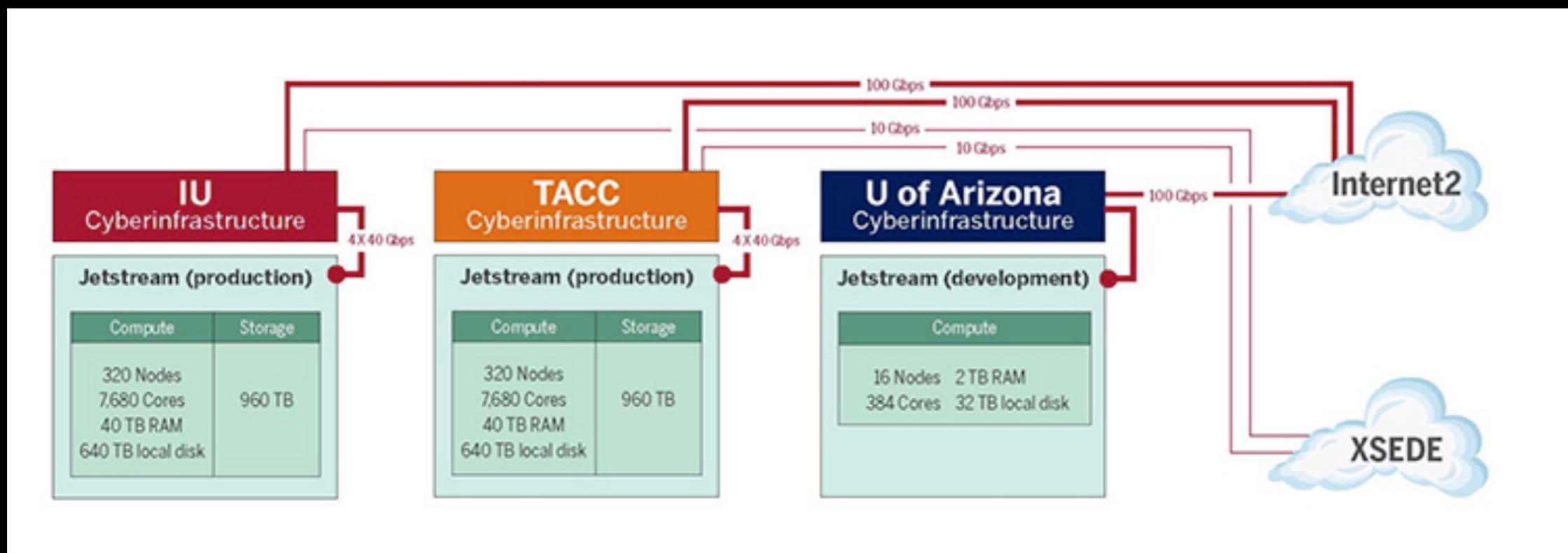
- XSEDE sponsors full-semester online courses
- Collaborations with faculty at participating institutions
- Campus visits offer guidance concerning course content

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):

- UCSanDiego 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 box)
- Computer Setup
- Learning Goals
- Assignments & Grading
- Ethics Code
- Screen Cast Videos

Right Side (Tutorial Content):

The URL in the address bar is `bioboot.github.io/bggn213_f17/jetstream/bboot/`.

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!

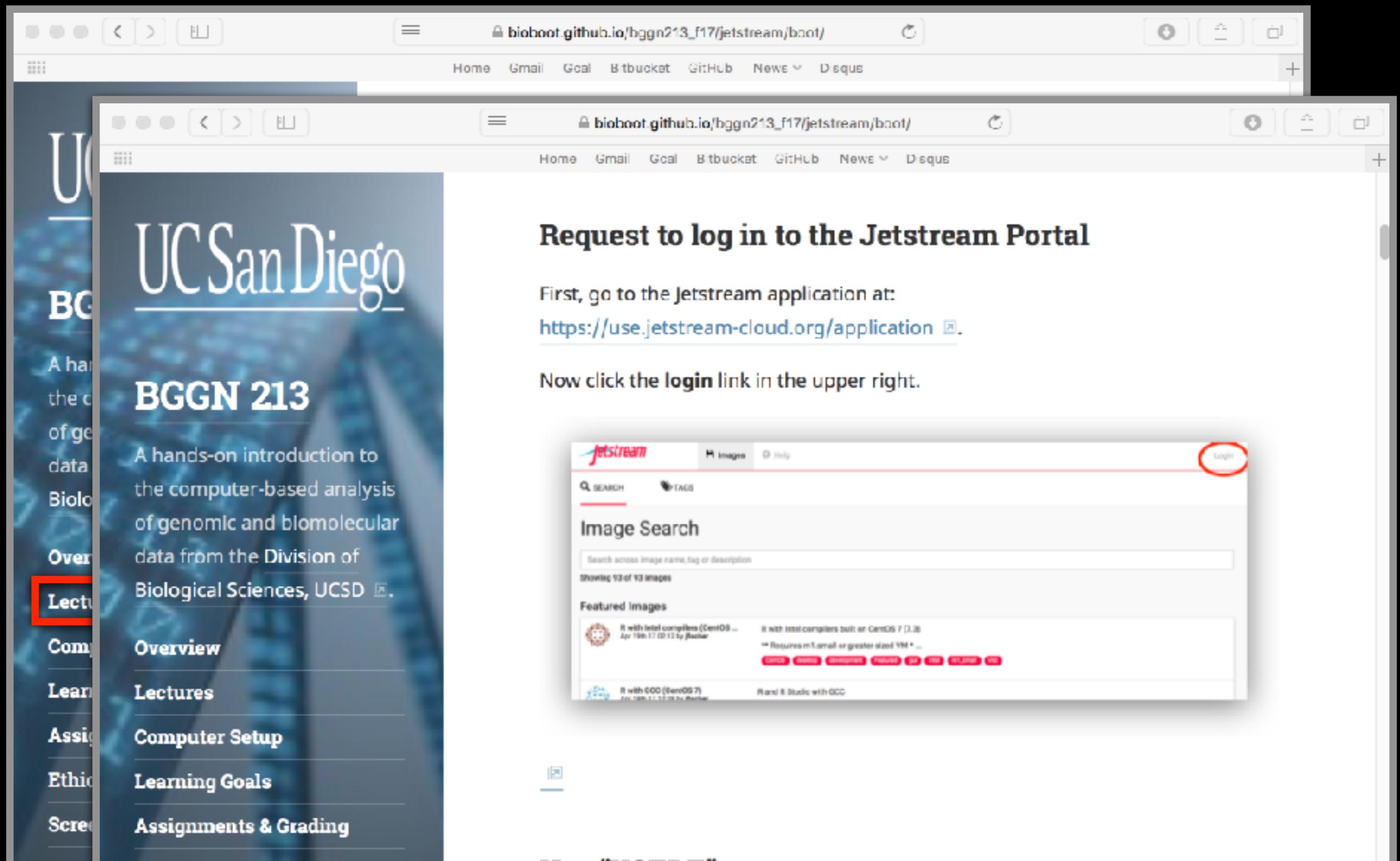
What we're going to do here is walk through starting up and running computer (an "instance") on the Jetstream service.

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 with two tabs open. The left tab is for the UC San Diego BGGN 213 course, featuring a sidebar with 'Lectures' highlighted in red. The right tab is a 'Request to log in to the Jetstream Portal' page, showing the Jetstream application interface with a 'Login' button circled in red.

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

Over
Lectures
Computer Setup
Learning Goals
Assignments & Grading

Request to log in to the Jetstream Portal

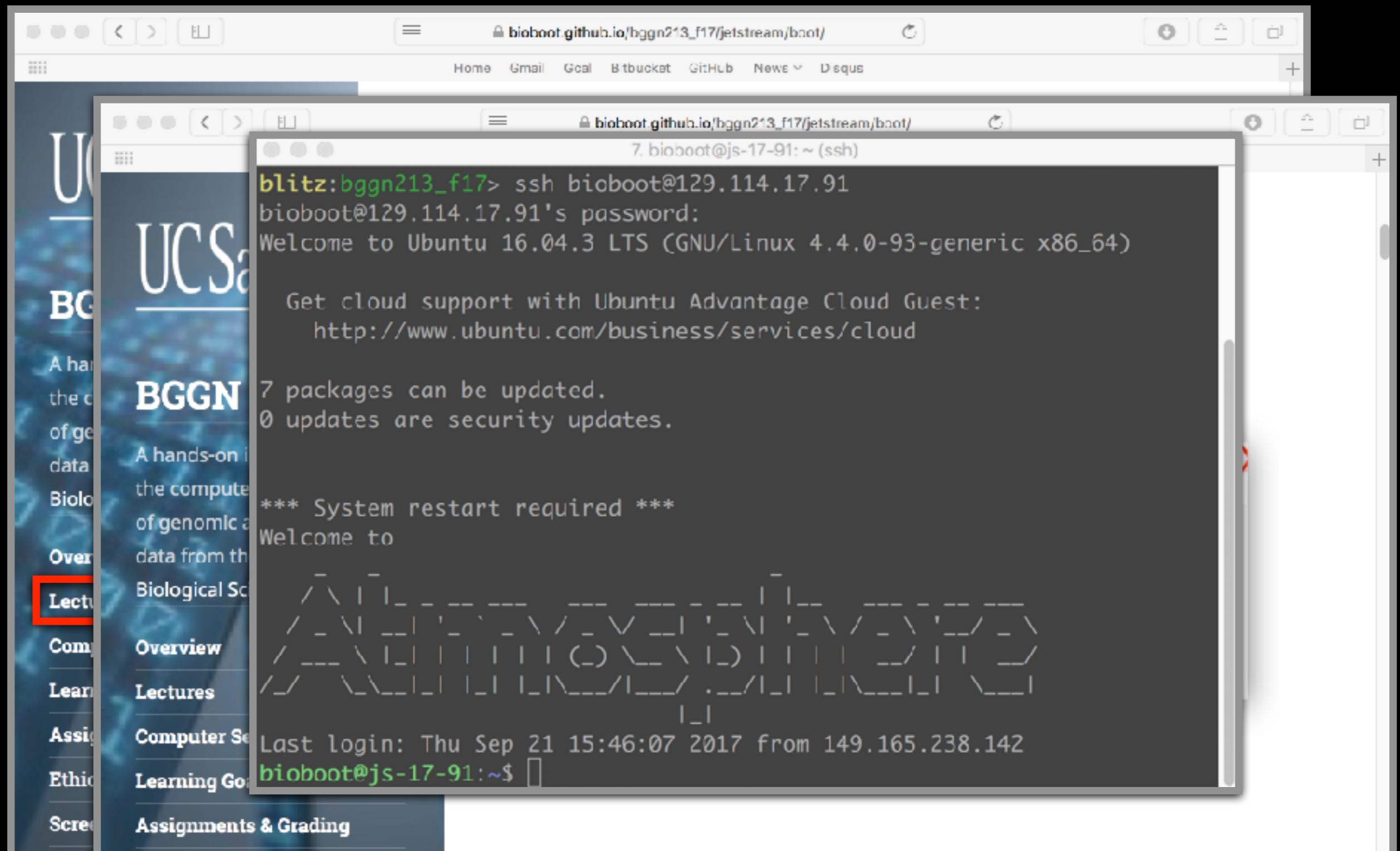
First, go to the Jetstream application at:
<https://use.jetstream-cloud.org/application>.

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

Jetstream  Images Help
SEARCH TAGS
Image Search
Search across image name, tag or description
Showing 93 of 93 images
Featured Images
R with Intel compilers (CentOS 7) Apr 19th 17:03:13 by Rainer
R with Intel compilers built on CentOS 7 (3.8) Requires m1.small or greater sized VM * ...
R with GCO (CentOS 7) Apr 19th 17:11:23 by Rainer R and R Studio with GCO

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

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

Common problems with Bioinformatics

Confusing multitude of tools available

- ▶ Each with many options and settable parameters

Most tools and databases are written by and for nerds

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

Most are developed independently

Notable exceptions are found at the:

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

General Parameters

Max target sequences

Select the maximum number of aligned sequences to display

Short queries Automatically adjust parameters for short input sequences

Expect threshold

Word size

Max matches in a query range

Scoring Parameters

Matrix

Gap Costs Existence: 11 Extension: 1

Compositional adjustments Conditional compositional sco

Filters and Masking

Filter Low complexity regions

Mask Mask for lookup table only

Mask lower case letters

PSI/PHI/DELTA BLAST

Upload PSSM Optional Choose File no file selected

PSI-BLAST Threshold

Pseudocount

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

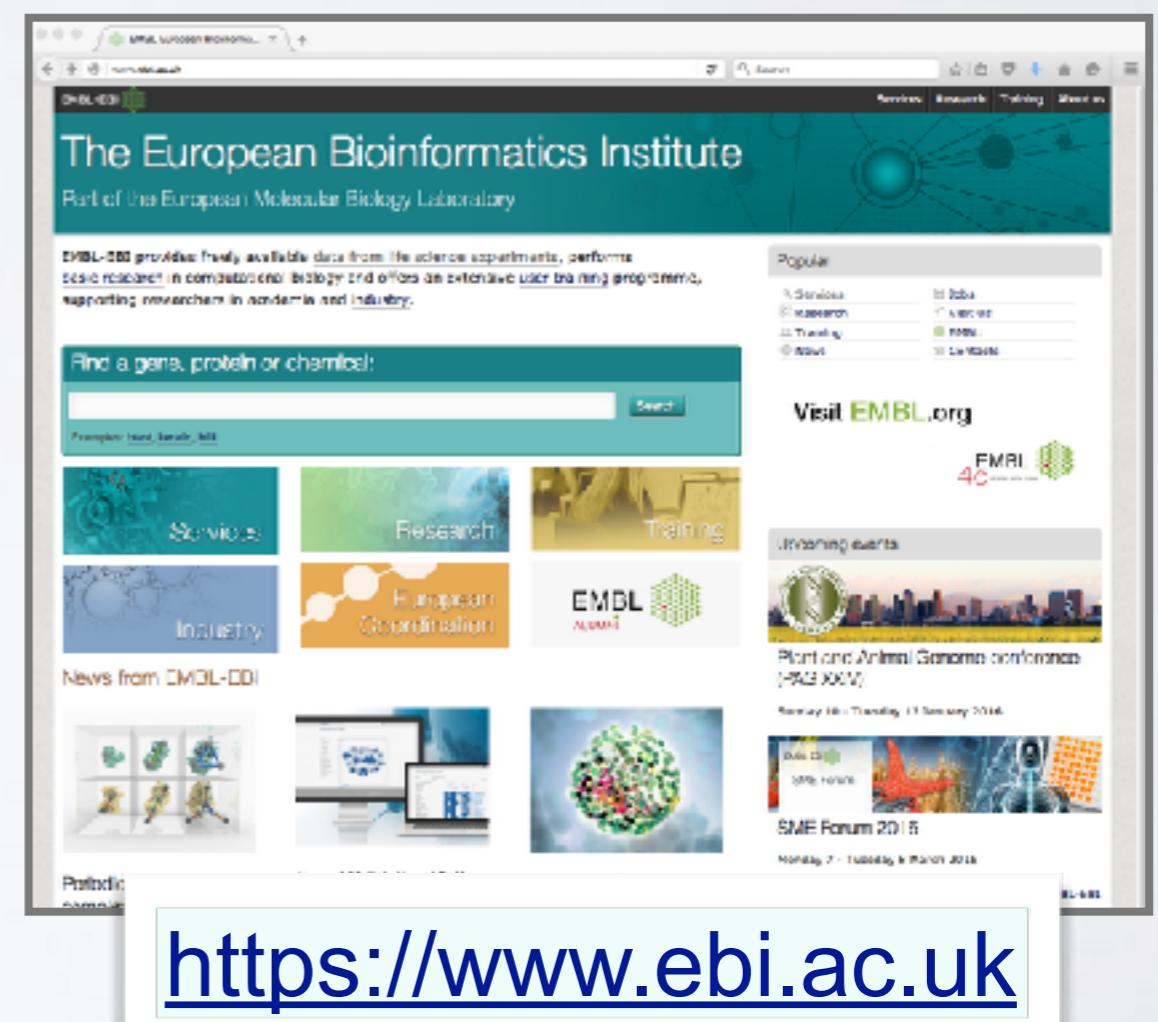
Related tools with different terminology

Key Online Bioinformatics Resources: NCBI & EBI

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



The screenshot shows the NCBI homepage with a blue header "National Center for Biotechnology Information". Below it is a navigation bar with links to "NCBI Resources", "How To", "Sign in to NCBI", "All Databases", and a search bar. The main content area includes a "Welcome to NCBI" section, a "Get Started" section with links to tools, downloads, how-to guides, and submissions, and a "3D Structures" section featuring a 3D molecular model. On the left is a sidebar with links to various NCBI databases like Resource List (A-Z), All Resources, Chemicals & Biosassays, Data & Software, DNA & RNA, Domains & Structures, Genes & Expression, Genetics & Medicine, Genomes & Maps, Homology, Literature, Proteins, Sequence Analysis, Taxonomy, Training & Tutorials, and Variation.



The screenshot shows the EMBL-EBI homepage with a green header "The European Bioinformatics Institute, Part of the European Molecular Biology Laboratory". Below it is a navigation bar with links to "Services", "Research", "Training", and "About us". The main content area includes a search bar for "Find a gene, protein or chemical", sections for "Services", "Research", "Training", "Industry", "European Coordination", and "News from EMBL-EBI". There are also sections for "Upcoming events" (Plant and Animal Genome conference) and "Past events" (SME Forum 2015). On the right is a sidebar with links to "Popular" resources like Adonis, ChIP-seq, iGEM, PANTHER, and UniProt, and a link to "Visit EMBL.org".

<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

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

Welcome to NCBI

The National Center for Biotechnology Information provides access to unique information, tools and resources in the fields of medicine, health and biology.

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.

Popular Resources

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

Resources

Central Health

Announcements

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 containing 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 'Gene', 'Protein', and 'PubChem' links, while three separate red arrows point to the 'PubMed', 'BLAST', and 'Gene' links respectively.

<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

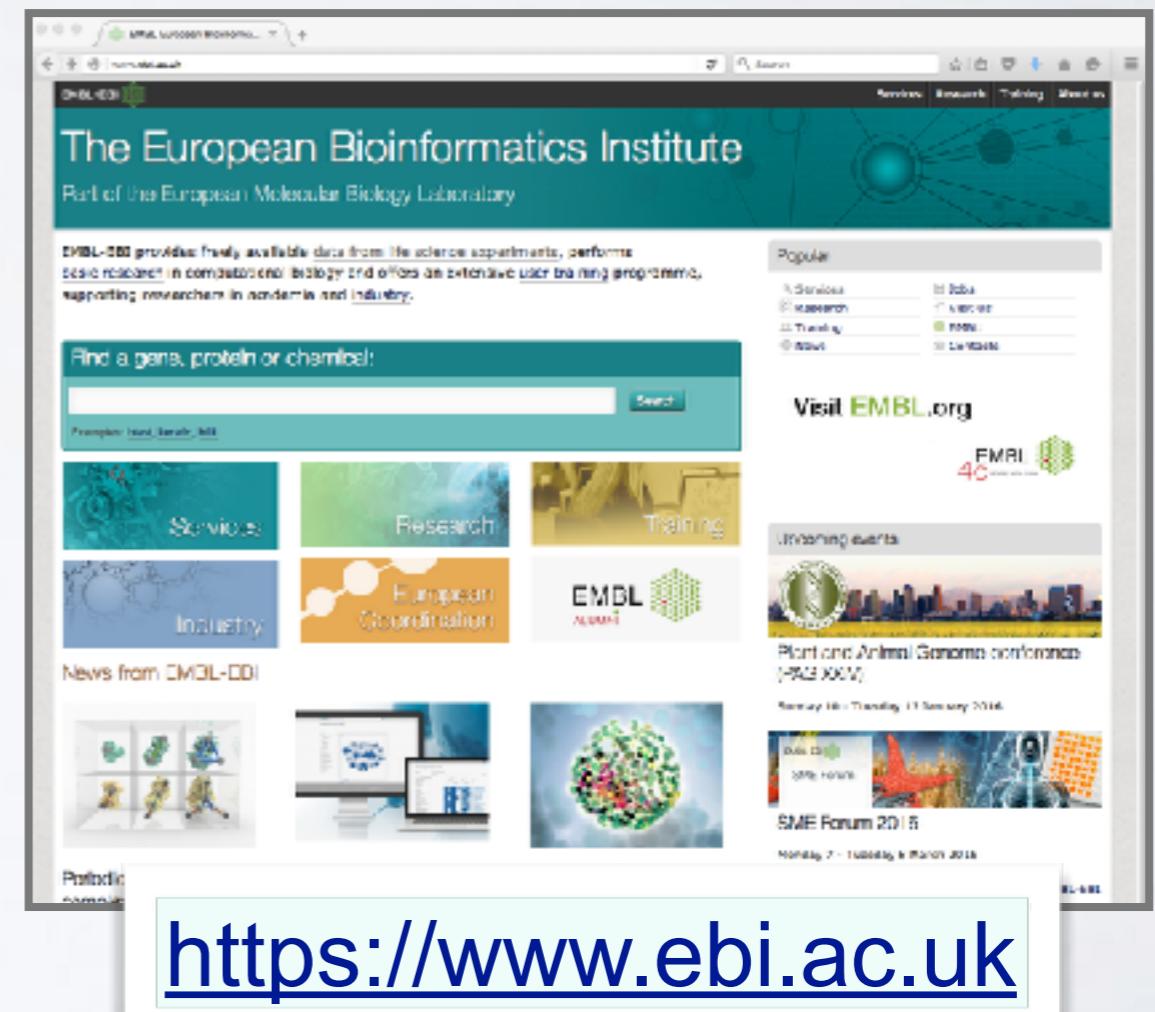
Key Online Bioinformatics Resources: NCBI & EBI

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



The screenshot shows the NCBI homepage with a blue header "National Center for Biotechnology Information". Below it is a navigation bar with links for "NCBI Resources", "How To", "Sign in to NCBI", "All Databases", and a search bar. The main content area includes sections for "Welcome to NCBI", "Get Started", "3D Structures", and "NCBI Announcements". The "Popular Resources" sidebar lists PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, NUCOTIDE, GenBank, SNP, Gene, Protein, and PubChem.

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



The screenshot shows the EBI homepage with a green header "The European Bioinformatics Institute". Below it is a navigation bar with links for "Reviews", "Research", "Training", and "About us". The main content area includes a search bar for "Find a gene, protein or chemical", sections for "Services", "Research", "Training", "Industry", "European Coordination", and "News from EMBL-EBI". The right sidebar features a "Popular" section with links for "Datasets", "Search", "Training", and "News", and a "Visit EMBL.org" section with a link to "EMBL-EBI".

<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, a search bar, and navigation links for Services, Research, Training, and About us. The main content area has a teal background with the text: "The European Bioinformatics Institute Part of the European Molecular Biology Laboratory". Below this, a paragraph describes the institute's mission: "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 allows users to "Find a gene, protein or chemical" with examples like "blast, keratin, bfl1...". To the right, a "Popular" sidebar lists links to Services, Research, Training, News, Jobs, Visit us, EMBL, and Contacts. A "Visit EMBL.org" section features the EMBL 40th anniversary logo. Another sidebar for "Upcoming events" promotes 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

Industry

European Coordination

News from EMBL-EBI

Popular

Services

Research

Training

News

Jobs

Visit us

EMBL

Contacts

Visit EMBL.org

EMBL 40 YEARS 1974-2014

Upcoming events

Plant and Animal Genome conference (PAG XXIV)

Sunday 10 - Tuesday 12 January 2016

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

The screenshot shows the EMBL-EBI Services website. At the top, there's a navigation bar with links for Services, Research, Training, and About us. Below the header, a large banner features a blue-toned image of molecular structures and gears. The main content area is titled "Bioinformatics services". It lists several categories: DNA & RNA, Gene expression, Proteins, Structures, Systems, Chemical biology, Ontologies, Literature, and Cross domain. Each category has a small icon and a brief description. To the right, there's a "Popular" sidebar with links to Ensembl, UniProt, PDBc, ArrayExpress, CHEMBL, BLAST, Europe PMC, Reactome, Train online, and Support. Below the sidebar, there's a "Service news" section featuring a photo of a butterfly and a "Training" section.

Services < EMBL-EBI

www.ebi.ac.uk/services

Search

Services | Research | Training | About us

Services

Overview A to Z Data submission Support

Bioinformatics services

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

DNA & RNA
genes, genomes & variation

Gene expression
RNA, protein & metabolite expression

Proteins
sequences, families & motifs

Structures
Molecular & cellular structures

Systems
reactions, interactions & pathways

Chemical biology
chemogenomics & metabolomics

Ontologies
taxonomies & controlled vocabularies

Literature
Scientific publications & patents

Cross domain
cross-domain tools & resources

Popular

- Ensembl
- UniProt
- PDBc
- ArrayExpress
- CHEMBL
- BLAST
- Europe PMC
- Reactome
- Train online
- Support

Service news

Training

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

The screenshot shows the EBI Services website (www.ebi.ac.uk/services) with a banner for 'Bioinformatics services'. Below the banner, there are nine service categories arranged in a grid:

- DNA & RNA (genes, genomes & variation)
- Gene expression (RNA, protein & metabolite expression)
- Proteins (sequences, families & motifs) - This box is highlighted with a red border.
- 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)

To the right of the services grid is a 'Popular' sidebar containing links to Ensembl, UniProt, PDBe, ArrayExpress, and ChEMBL, each with its respective logo. A large image of a monarch butterfly is visible in the background of the sidebar area. At the bottom right, there is a 'Training' section with a blurred image of a person working at a computer.

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

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

Proteins

Popular services

 UniProt	UniProt: The Universal Protein Resource The gold-standard, comprehensive resource for protein sequence and functional annotation data.	Quick links <ul style="list-style-type: none">○ Popular services in this category○ All services in this category○ Project websites in this category
 InterPro	InterPro A database for the classification of proteins into families, domains and conserved sites.	
 PRIDE	PRIDE: The Proteomics Identifications Database An archive of protein expression data determined by mass spectrometry.	
 Pfam	Pfam A database of hidden Markov models and alignments to describe conserved protein families and domains.	
 Clustal Omega	Clustal Omega Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.	
 HMMER	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.	

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 blue header with the EMBL-EBI logo, a search bar, and navigation links for Services, Research, Training, and About us. The main content area has a teal background with the text: "The European Bioinformatics Institute Part of the European Molecular Biology Laboratory". Below this, a paragraph describes EMBL-EBI's mission: "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 followed by a "Find a gene, protein or chemical:" input field with examples like "blast, keratin, bfl1...". To the right is a "Popular" sidebar with links to Services, Research, Training, News, Jobs, Visit us, EMBL, and Contacts. A large yellow button labeled "Training" is highlighted with a red border. Other sections include "Services", "Research", "European Coordination", "Industry", "EMBL ALUMNI", and "Upcoming events" for 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

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

Plant and Animal Genome conference (PAG XXIV)

Sunday 10 - Tuesday 12 January 2016

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

The screenshot shows a web browser displaying the EMBL-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 features a video player showing a thumbnail of the webinar presentation. The thumbnail has a blue header with the text "Using sequence similarity search tools at EMBL-EBI" and "Finding homologous sequences with BLAST, FASTA, PSI-Search etc.". Below the header is a photo of a man (Andrew Cowley) and his contact information: andrew.cowley@ebi.ac.uk and support@ebi.ac.uk. The video player shows a progress bar at 0:03 / 37:42. To the left of the video, there is a sidebar with "Course content" sections: "Using sequence similarity searching tools at EMBL-EBI: webinar" (which is highlighted in yellow), "Contributors", and a "Print Course" button. At the bottom of the page, there is a summary text: "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."

Using sequence similarity e... X

www.ebi.ac.uk/training/online/course/using-sequence-similarity-searching-tools-embl-ebi

EMBL-EBI

Services Research Training About us

Train online

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

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

Course content

Using sequence similarity searching tools at EMBL-EBI: webinar

Contributors

Print Course

Using sequence similarity searching tools at EMBL-EBI: webinar

Using sequence similarity searching tools at EMBL-EBI: webinar

Using sequence similarity search tools at EMBL-EBI

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

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

0:03 / 37:42

Popular

Train online

Find us

Funding

Find us at...

Open days and career days

Conference exhibitions

EMBL courses and events

Genome campus events

Science for schools

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

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

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

Notable EBI databases include:
ENA, **UniProt**, **Ensembl**

and the tools **FASTA**, **BLAST**, **InterProScan**,
MUSCLE, **DALI**, **HMMER**

Find a course

Browse by subject



[Genes and Genomes](#)



[Gene Expression](#)



[Interactions, Pathways, and Networks](#)

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

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, RCCP, Beanref, CANSITE, CarbBank, CARBHYD, CATH, CAZy, ChickGBASE, Colibri, COPE, CottonDB, dbSTS, DDBJ, DGP, DictyDb, ECGC, EC02DBASE, FlyBase, GDB, HEPDB, HumanProteome, KEGG, MHCDB, MycoDB, PDBe, PDB, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc ..!!!!

There are lots of Bioinformatics Databases

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

< Major Databases.pdf >

Side-note: Databases come in all shapes and sizes



Databases can be of variable quality and often there are multiple databases with overlapping content.

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

The screenshot shows a web browser window with the URL https://bioboot.github.io/bggn213_S18/lectures/#1 in the address bar. 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 (highlighted with a red border)

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

Goals:

- Understand course scope, expectations, logistics and [ethics code](#).
- Understand the increasing necessity for computation in modern life sciences research.
- Get introduced to how bioinformatics is practiced.
- Complete the [pre-course questionnaire](#).
- 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](#) (highlighted with a red border)
- 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
TTGGTGGTGAGGCCCTGGCAGGCTGCTGGTGGTCTACCCCTGGACCCAGAGGTTCTTGAGTCCTTGG
GGATCTGTCCACTCCTGATGCAGTTATGGCAACCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTAGTGTAGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTGCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACCTCAGGCTCCTGGCAACGTGCTGGTCTGTGTGCTGGCCA
TCACCTTGGCAAAGAACATTCCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
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:

	End times:
1. BLAST, GenBank and OMIM @ NCBI	[10:45 am]
2. GENE database @ NCBI	[11:00 am]
— BREAK —	— 11:10 am —
3. UniProt & Muscle @ EBI	[11:35 am]
4. PFAM, PDB & NGL	[12:05 pm]
— BREAK —	— 12:15 am —
5. Extension exercises	[12:45 pm]

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

SUMMARY

- Bioinformatics is computer aided biology.
- Bioinformatics deals with the collection, archiving, organization, and interpretation of a wide range of biological data.
- There are a large number of primary, secondary and tertiary bioinformatics databases.
- The NCBI and EBI are major online bioinformatics service providers.
- Introduced Gene, UniProt, PDB databases as well as a number of ‘boutique’ databases including PFAM and OMIM.

HOMEWORK

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

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

THANK YOU