



HELLO
my name is

BARRY

bjgrant@ucsd.edu

HELLO
HIS → name is

KEVIN

kkchau@ucsd.edu

Office Hours: [SignUp](#)

Location: TATA, #2501

Introduce Yourself!

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

Today's Menu

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

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

The screenshot shows the homepage of the BGGN 213 course. At the top, there's a navigation bar with links to Home, Gmail, Gcal, Bitbucket, GitHub, News, Disqus, BGGN-213, BIMM-143, BIMM-194, Atmosphere, Blink, GDocs, and Galaxy. Below the navigation is the UC San Diego logo and the course title "BGGN 213". A sub-header "Bioinformatics (BGGN 213, Spring 2018)" is followed by a magnifying glass icon with a DNA helix inside it. On the left, a sidebar lists course navigation: Overview, Lectures, Computer Setup, Learning Goals (which is highlighted with a red box), Assignments & Grading, and Ethics Code. Below the sidebar, social media icons for Twitter, GitHub, Email, and RSS are shown. The main content area contains sections for Course Director (Prof. Barry J. Grant), Instructional Assistant (Yuansheng Zhou), Course Syllabus (Spring 2018 PDF), and an "Overview" section describing bioinformatics as a field of computational methods applied to biological problems.

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

The screenshot shows the "Learning Goals" page for the BGGN 213 course. The page has a similar header and sidebar to the homepage. The main content area starts with a heading "Learning Goals" and a sub-section "At the end of this course students will:". Below this is a bulleted list of learning objectives. After the list, a summary states: "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." The "Learning Goals" link in the sidebar is also highlighted with a red box.

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 the UC San Diego BGGN 213 website. The left sidebar includes links for Overview, Lectures (highlighted with a red box), Computer Setup, Learning Goals (highlighted with a red box), Assignments & Grading, Ethics Code, and Screen Cast Videos. The main content area is titled "Specific Learning Goals". It states: "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." Below this, it says: "All students who receive a passing grade should be able to:" followed by a table:

	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 the UC San Diego BGGN 213 website. The left sidebar includes links for Overview, Lectures (highlighted with a red box), Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. The main content area is titled "Lectures". It says: "All Lectures are Wed/Fri 1:00-4:00 pm in Warren Lecture Hall 2015 (WLH 2015) (Map)". Below this is a table:

#	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

The screenshot shows the UC San Diego BGGN 213 website. The left sidebar includes links for Overview, Lectures (highlighted with a red box), Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, and Screen Cast Videos. The main content area is titled "Lectures". It says: "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." Below this is a table:

#	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 the UC San Diego BGGN 213 website. The left sidebar includes links for Overview, Lectures (highlighted with a red box), Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, and Screen Cast Videos. The main content area is titled "1: Welcome to Foundations of Bioinformatics". It says: "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." Below this is a section titled "Goals:" with a bulleted list:

- 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.

Below this is a section titled "Material:" with a bulleted list:

- [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 the UC San Diego BGGN 213 course website. The main content area displays the 'Homework' section, which includes a list of 'Questions' (marked with a red box) and 'Readings'. Under 'Readings', there are three items: 'PDF1: What is bioinformatics? An introduction and overview', 'PDF2: Advancements and Challenges in Computational Biology', and 'Other: For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights'. Below this is a 'Screen Casts' section featuring a video thumbnail of Barry Grant discussing 'Foundations of Bioinformatics'.

Homework

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

The screenshot shows a Google Form titled 'BGGN213 Lecture 1 Homework'. It contains a question asking students to answer the following questions. A red banner across the form reads 'Homework is due before the next weeks class!'. The form also includes a question about the most frequently used operating systems for bioinformatics tool development, with options for Windows, iOS, Unix, and Perl.

Projects

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

The screenshot shows the UC San Diego BGGN 213 course website. The main content area displays a '9: Unsupervised learning mini-project'. The 'Topics' section describes a longer hands-on session on unsupervised learning analysis of cancer cells. The 'Goals' section lists the ability to import data and prepare for unsupervised learning analysis, apply PCA, k-means, and hierarchical clustering, and critically review results. The 'Material' section provides links to lecture slides, a lab worksheet, data files, a Bio3D PCA app, and feedback.

Projects

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

The screenshot shows the UC San Diego BGGN 213 course website. The main content area displays a '18: Cancer genomics' section. The 'Topics' section covers cancer genomics resources and bioinformatics tools for investigating the molecular basis of cancer, mentioning NCI Genomic Data Commons and immunotherapy. The 'N.B.' section notes a gene assignment due before the next class. The 'Material' section lists lecture slides, a lab worksheet, and data files.

Projects

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

10: Project: Find a gene assignment (Part 1)

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

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

Your responses to questions Q1-Q4 are due at the beginning of class **Fri Feb 22nd** (02/22/19).

The complete assignment, including responses to all questions, is due at the beginning of class **Wed March 13th** (03/13/19).

Late responses will not be accepted under any circumstances.

Why Projects?

- Projects allow you to practice your new Bioinformatics skills in a less guided environment.
- In Projects, we provide datasets and ask you questions about them; just like a research project.
- Projects help build a personal portfolio and showcase your new skills, as well as help put what we have learned into practice.

Online portfolio of **your** bioinformatics work!

Bioinformatics Class BIMM-143

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

Index of Material

Introductory Material: Working With R

- Class 5 - Basic Data Exploration and Visualization in R
- Class 6 - Creating R Functions
- Class 7 - R Packages, working with CRAN, and working with Bioconductor
- Using R and Other Tools for Bioinformatics Analysis
- Class 8 - An Introduction to Machine Learning (Hierarchical Clustering)
- Class 9 - Analyzing High Dimensional Datasets and Unsupervised Learning
- Class 11 - Structural Bioinformatics: Analyzing Protein Structure and Function
- Class 12 - Drug Discovery: Techniques and Analysis
- Class 13 - Genome Informatics and High Throughput Sequencing (NGS, RNA-Seq, and FastQC)
- Class 14 - Transcriptomics and RNA-Seq Analysis
- Class 15 - Genome Annotation and Using Functional Databases (KEGG and GO - Gene Ontology)
- Class 16 - Transposons: A Sample Workflow

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

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

This project is maintained by [jasonPBennett](#)

Online portfolio of **your** bioinformatics work!

class13

Jason Patrick Bennett
May 15, 2018

Identifying SNP's in a Population

Lets analyze SNP's from the Mexican-American population in Los Angeles:

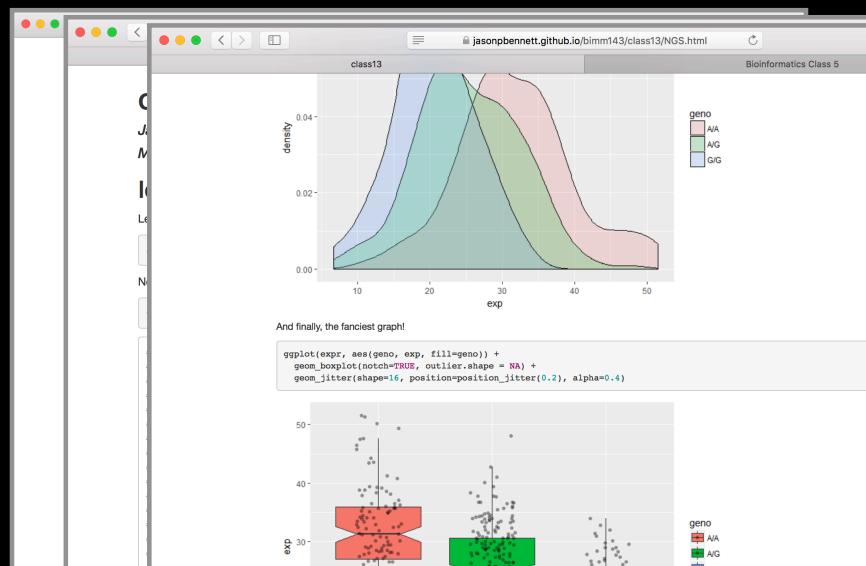
```
genotype <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")
```

Now lets look at a table of the data:

```
table(genotype)
```

```
## , , Population.s. = ALL, AMR, MXL, Father = -, Mother = -
##                               Genotype..forward.strand.
## Sample..Male.Unknown. A|A A|G C|A G|G
## # NA19648 (F) 1 0 0 0
## # NA19649 (M) 0 0 0 1
## # NA19651 (F) 1 0 0 0
## # NA19652 (M) 0 0 0 1
## # NA19654 (F) 0 0 0 1
## # NA19655 (M) 0 1 0 0
## # NA19657 (F) 0 1 0 0
## # NA19658 (M) 1 0 0 0
## # NA19661 (M) 0 1 0 0
## # NA19663 (F) 1 0 0 0
## # NA19664 (M) 0 0 1 0
```

Online portfolio of **your** bioinformatics work!



Bonus:

Bioinformatics & Genomics in industry

The screenshot shows the UC San Diego BGGN 213 course page. It features a sidebar with links to Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. The main content area includes a title 'BGGN 213', a description of the course, and a section for 'Friday March 15th at 1pm come and enjoy a set of short open ended guest lectures from leading genomic scientists at Illumina Inc., Synthetic Genomics Inc., Samumed and the La Jolla Institute for Allergy and Immunology. Come prepared for networking and to have your questions about industry careers in Bioinformatics and Genomics answered.'

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 programming experience or high level math skills.

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

BGGN-213 Learning Goals....

Advanced UNIX and R based learning goals

The screenshot shows the BGGN 213 Learning Goals page. The sidebar has a red box around the 'Learning Goals' link. The main content lists 12 learning goals numbered 5 through 12, each with a brief description and associated numbers.

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	13, 14, 15

BGGN-213 Learning Goals....

Delve deeper into “real-world” bioinformatics

The screenshot shows the UC San Diego BGGN 213 course website. The 'Learning Goals' section is highlighted with a red box. A green box highlights a subset of learning goals, specifically items 15 through 19. These goals involve analyzing RNA-Seq data, performing GO analysis, using KEGG pathway databases, applying graph theory to biological networks, and understanding social impacts of genomic data. The entire list of 20 learning goals is visible in the table.

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

Why use R?

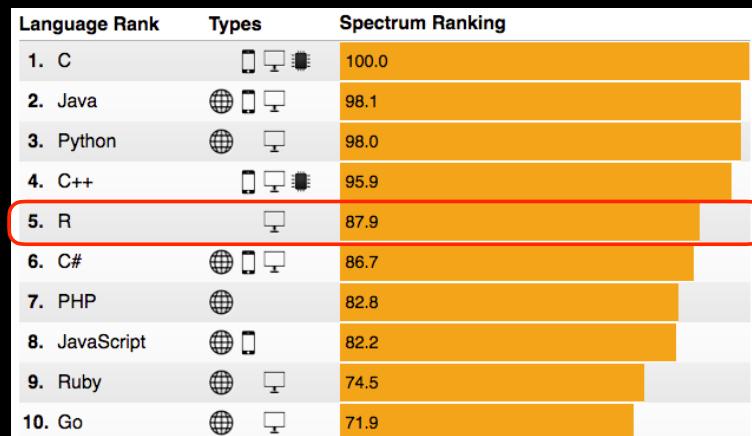
Productivity
Flexibility
Genomic data analysis

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 UNIX and 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.

IEEE 2016 Top Programming Languages

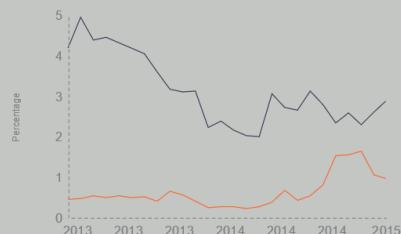


<http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages>

R and Python: The Numbers

Popularity Rankings

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



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

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

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

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.

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

What is an IDE anyway?

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

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

A screenshot of the DataCamp RStudio IDE interface. On the left, a modal window titled "Exercise Completed" shows a green checkmark icon and the text "by completing the first exercise". Below it, a "Possible Answers" section has a "Continue" button highlighted with a red circle. On the right, the RStudio environment shows the console, environment, and global history panes. A message in the console pane says "Environment is empty". The R console shows the following code:

```
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.
Y
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/ >

Homework assignments will be via DataCamp

A screenshot of the DataCamp RStudio IDE interface. The left pane shows an "Exercise" titled "PCA analysis". The right pane shows the R console with the following code and output:

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

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

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

A screenshot of the DataCamp Groups page for the "Foundations of Bioinformatics (BGN-213)" group. The top navigation bar has a "Groups" button highlighted with a red circle. The page displays a table of group members with their names, XP points, courses, and chapters completed. The table data is as follows:

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

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.

Q. What is Bioinformatics?

Q. What is Bioinformatics?

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

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

Q. What is Bioinformatics?

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

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

Q. What is Bioinformatics?

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

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

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

MORE DEFINITIONS

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

- “Bioinformatics is research, development, or application of **computational approaches** for expanding the use of **biological, medical, behavioral or health data**, including those to **acquire, store, organize and analyze** such data.”

National Institutes of Health (NIH) (<http://tinyurl.com/l3gxr6b>)

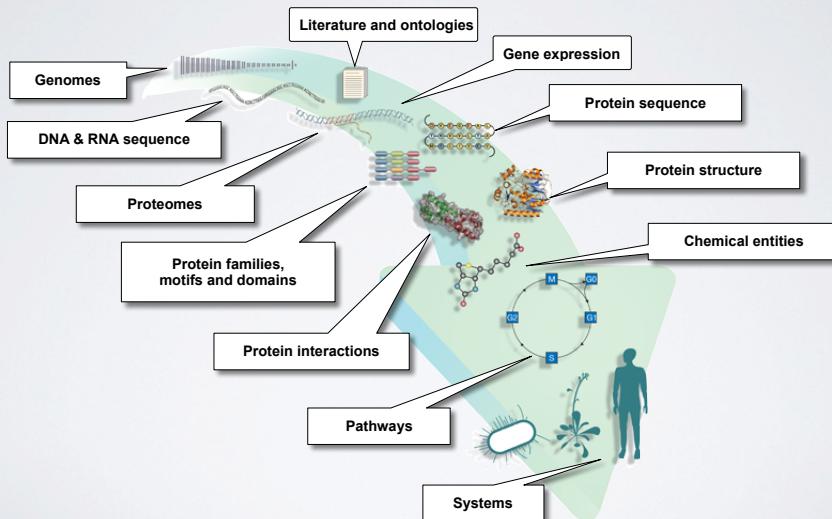
MORE DEFINITIONS

- “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “**informatics**” techniques (derived from disciplines such as applied 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.

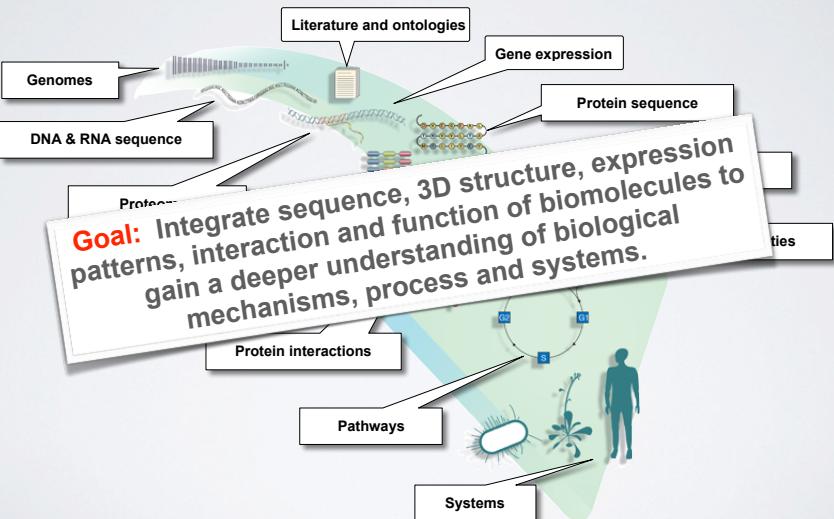
- “Bioinformatics is research, development, or application of **computational approaches** for expanding the use of **biological, medical, behavioral or health data**, including those to **acquire, store, organize and analyze** such data.”

National Institutes of Health (NIH) (<http://tinyurl.com/l3gxr6b>)

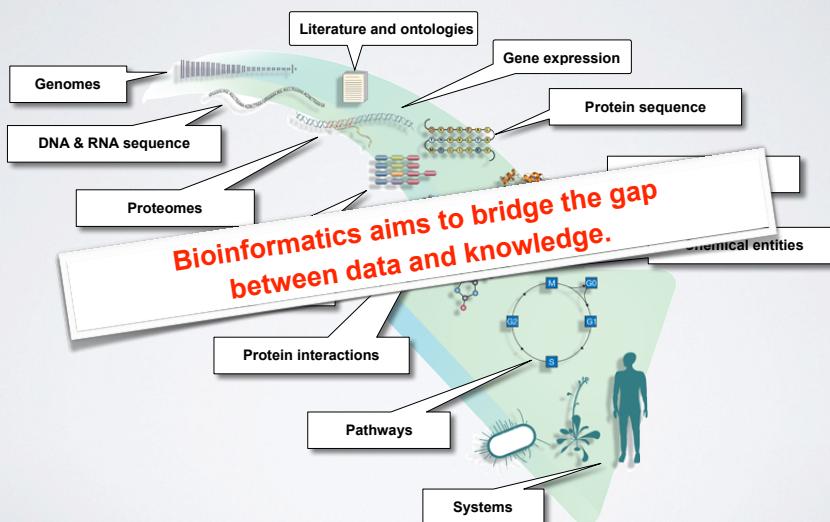
Major types of Bioinformatics Data



Major types of Bioinformatics Data



Major types of Bioinformatics Data



BIOINFORMATICS RESEARCH AREAS

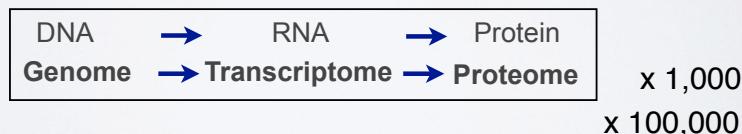
Include but are not limited to:

- 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

How do we do Bioinformatics?

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



How do we *actually* do Bioinformatics?

Pre-packaged tools and databases

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

Advanced tool application & development

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

How do we *actually* do Bioinformatics?

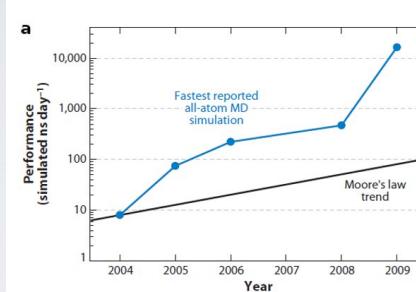
Pre-packaged tools and databases

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

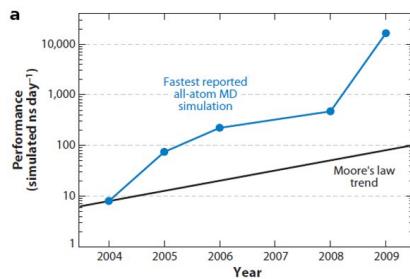
Advanced tool application & development

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

SIDE-NOTE: SUPERCOMPUTERS AND GPUS



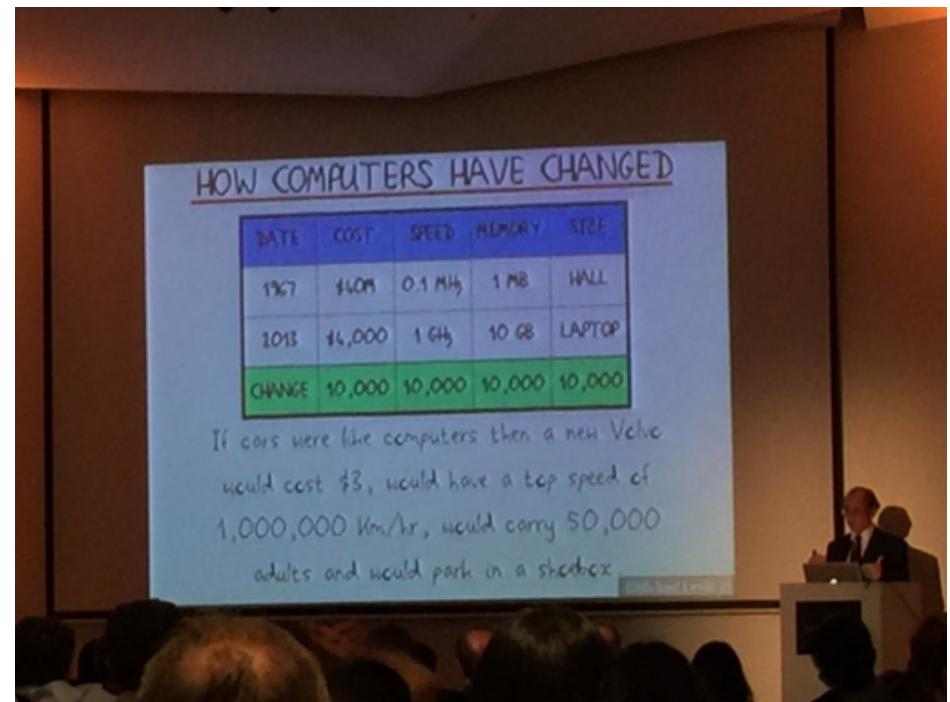
SIDE-NOTE: SUPERCOMPUTERS AND GPUS



HOW COMPUTERS HAVE CHANGED

DATE	COST	SPEED	MEMORY	SIZE
1967	\$60B	0.1 MHz	1 MB	HALL
2013	\$6,000	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 shed.



NSF Extreme Science and Engineering Discovery Environment (XSEDE)

The screenshot shows the XSEDE website's "Curriculum and Educator Programs" page. The header includes links for About, For Users, Ecosystem, Community Engagement, News, and XUP. A search bar is also present. The main content features a large image of a galaxy and sections for "Campus Visits" and "Key Points".

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

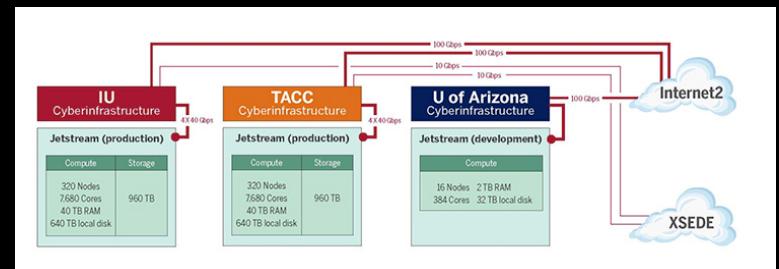
- XSEDE sponsors full-semester online courses
- Collaborations with faculty at participating institutions
- Campus visits offer guidance concerning course content

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 tutorial page from the UC San Diego BGGN 213 website. The page title is "Starting a Jetstream Computer Instance!". It describes the process of starting up and managing a Jetstream service virtual machine instance. The text notes that Jetstream is a cloud-based on-demand virtual machine system funded by the National Science Foundation, providing computers 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 screenshot shows a tutorial page from the UC San Diego BGGN 213 website. The page title is "Request to log in to the Jetstream Portal". It instructs users to first go to the Jetstream application at <https://use.jetstream-cloud.org/application>. It then says to click the "login" link in the upper right.

Request to log in to the Jetstream Portal

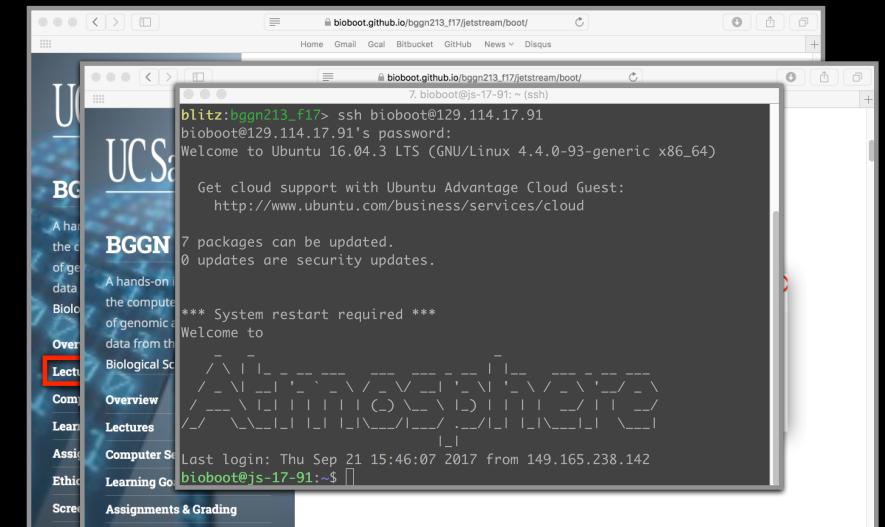
First, go to the Jetstream application at:
<https://use.jetstream-cloud.org/application>.

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

The screenshot shows the Jetstream login page. At the top, there is a search bar with a "SEARCH" button and a "LOG IN" button. Below the search bar, there is a "Image Search" section with a "Search images image name, tag or description" input field and a "Showing 10 of 10 images" message. Under "Featured Images", there are two images: one for "Ubuntu 16.04 (CentOS 7)" and another for "Ubuntu 16.04 (CentOS 7) with GCC".

Jetstream tutorials

Developed *user friendly* labs for Jetstream basics



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’!

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’

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)

Protein BLAST: search protein databases using a protein query
blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome

General Parameters

Max target sequences: 500 Select the maximum number of aligned sequences to display

Short queries: 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 score

Filters and Masking

Filter: Low complexity regions

Mask: Mask for lookup table only
 Mask lower case letters

PSI/PHI/DELTA BLAST

Upload PSSM: Choose File no file selected

Optimal PSI-BLAST Threshold: 0.005

Pseudocount: 0

Even Blast has many settable parameters

STEP 3 - Set your search parameters

Related tools with different terminology

PROGRAM	SEARCH PARAMETERS				EXPECTATION LOWER VALUE	EXPECTATION UPPER VALUE
FASTA	MATRIX	GAP OPEN	GAP EXTEND	KTUP	0 (default)	10
	BLOSUM50	-10	-2	2		
	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					

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

Bethesda, MD



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 search bar at the top containing the query "NCBI". Below the search bar, there are several search results listed in a table format. The first result is a link to the "National Center for Biotechnology Information". Other results include links to "NCBI Bookshelf", "PubMed Central", "PubMed Health", "BLAST", "Nucleotide", "GenBank", "Gene", "Protein", and "PubMed". On the left side of the page, there is a sidebar with links to "NCBI Home", "Resource List [A-Z]", "All Resources", "Chemicals & Bioassays", "Data & Software", "Databases", "Domain & Structure", "Gene & Expression", "Genetics & Medicine", "Gene Maps", "Homology", "Literature", "Proteins", "Protein Analysis", "Taxonomy", "Training & Tutorials", and "Variation". On the right side, there are sections for "Popular Resources", "Recent Publications", and "NCBI Announcements". The footer contains links to "NCBI News", "About NCBI", "Mission | Organization", "Research", "NCBI Tools", "Help", and "Feedback".

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

National Center for Biotechnology Information

www.ncbi.nlm.nih.gov

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-Tos: 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 bioinformatics.

Popular Resources

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

NCBI Announcements

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

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

The screenshot shows the NCBI homepage with a sidebar on the left containing links to various resources like NCBI Home, Resource List (A-Z), and 3D Structures. The main content area features a "Popular Resources" section with links to PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. Arrows point from the text descriptions to the corresponding links. A "Get Started" section below provides tools for analyzing data using NCBI software, downloading data, learning how to access specific tools, and submitting data. A "3D Structures" section is also present.

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

This screenshot shows the same NCBI homepage as above, but with a large callout box highlighting the "Popular Resources" section. The box contains the text "Notable NCBI databases include: GenBank, RefSeq, PubMed, dbSNP and the search tools ENTREZ and BLAST". Below this, the "databases" section is visible, listing Homology, Literature, Proteins, Sequence Analysis, Taxonomy, Training & Tutorials, and Variation. A "3D Structures" section is also present at the bottom.

Key Online Bioinformatics Resources: NCBI & EBI

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

A screenshot of the NCBI homepage with a red box highlighting the URL "http://www.ncbi.nlm.nih.gov". The page content is identical to the one above, showing the Popular Resources section and other database links.

A screenshot of the EBI homepage with a red box highlighting the URL "https://www.ebi.ac.uk". The page content includes links to Popular Resources, Services, Research, and Events, along with news and upcoming events like the Plant and Animal Genome conference (PPG XXIX) and the GM4 Forum 2016.

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 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...

Services (highlighted with a red box)

Research

Training

Industry

European Coordination

EMBL ALUMNI

News from EMBL-EBI

Upcoming events

Plant and Animal Genome conference (PAG XXIV)
Sunday 10 - Tuesday 12 January 2016

EMBL-EBI 40 years research

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

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 (highlighted with a red box)

Gene expression

Proteins

Structures

Systems

Chemical biology

Ontologies

Literature

Cross domain

BLAST

UniProt

PDB

ArrayExpress

ChEMBL

Europe PMC

Reactome

Train online

Support

Service news

Training

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

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 (highlighted with a red box)

Gene expression

Proteins (highlighted with a red box)

Structures

Systems

Chemical biology

Ontologies

Literature

Cross domain

<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

InterPro

PRIDE: The Proteomics Identifications Database

Pfam

Clustal Omega

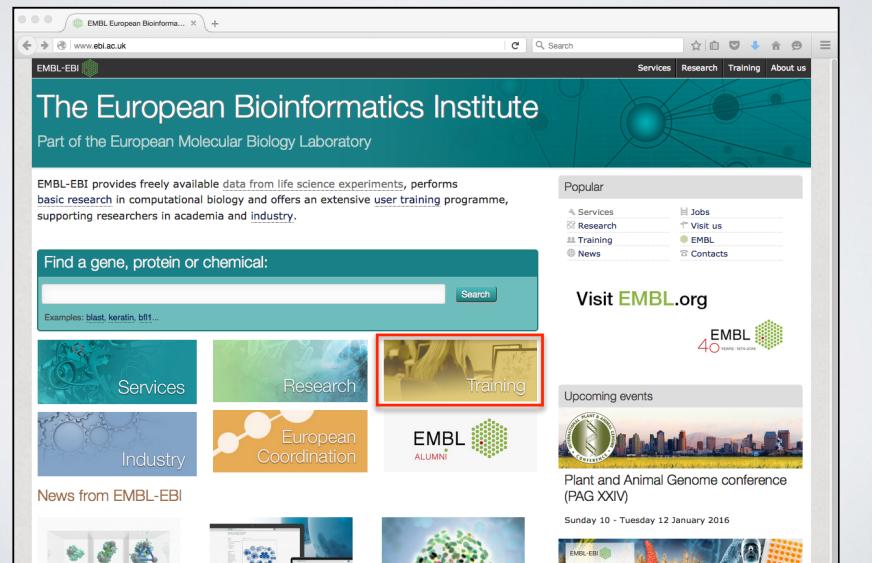
HMMER - protein homology search

InterProScan 5

Quick links

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

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



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...

Services Research Training Industry

European Coordination EMBL ALUMNI

News from EMBL-EBI

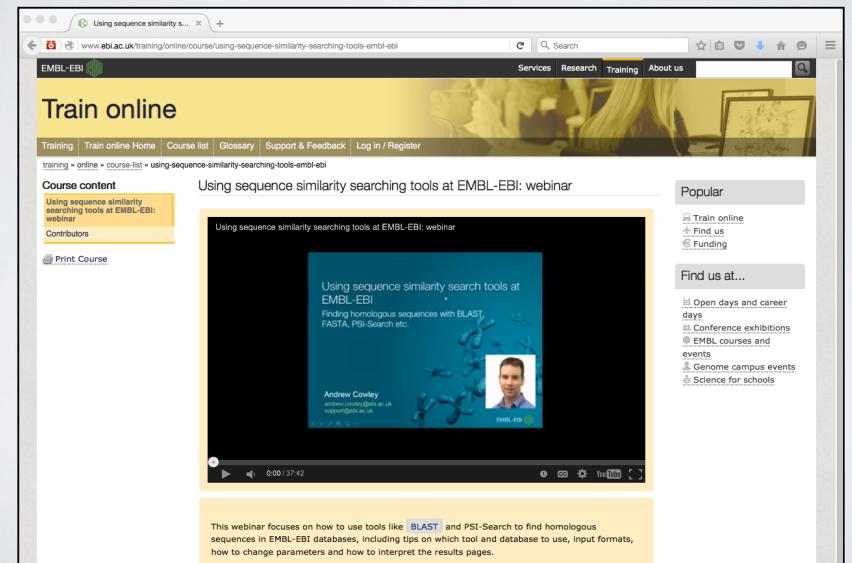
Visit EMBL.org

Upcoming events

Plant and Animal Genome conference (PAG XXIV)
Sunday 10 - Tuesday 12 January 2016

EMBL-EBI 40 years research

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



Train online

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

Course content

Using sequence similarity searching tools at EMBL-EBI: webinar

Contributors

Print Course

Using sequence similarity search tools at EMBL-EBI
Finding homologous sequences with BLAST, FASTA, PSI-Search etc.

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

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.

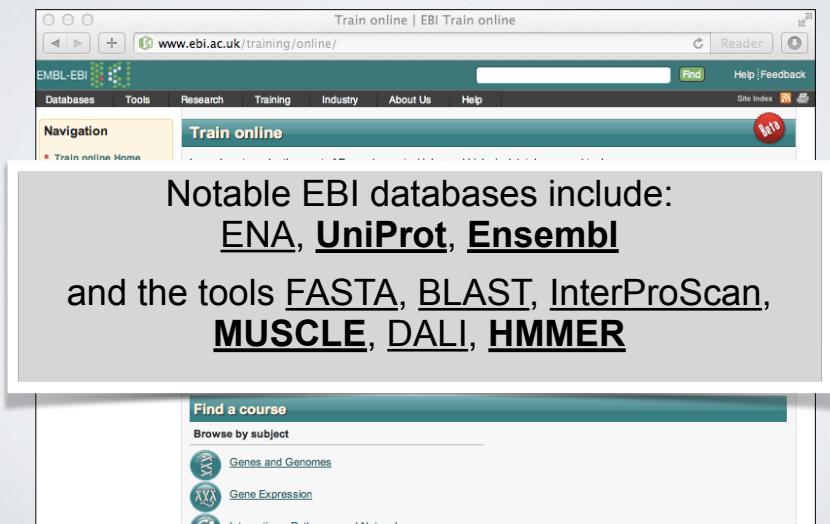
Popular

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

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



Train online | EBI Train online

www.ebi.ac.uk/training/online/

Databases Tools Research Training Industry About Us Help Site Index

Navigation 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 Databases 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, DPLInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genilexes, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, MycDB, NDB, NRSub, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS- MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc !!!

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, 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, DPLInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genilexes, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, MycDB, NDB, NRSub, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS- MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc !!!

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

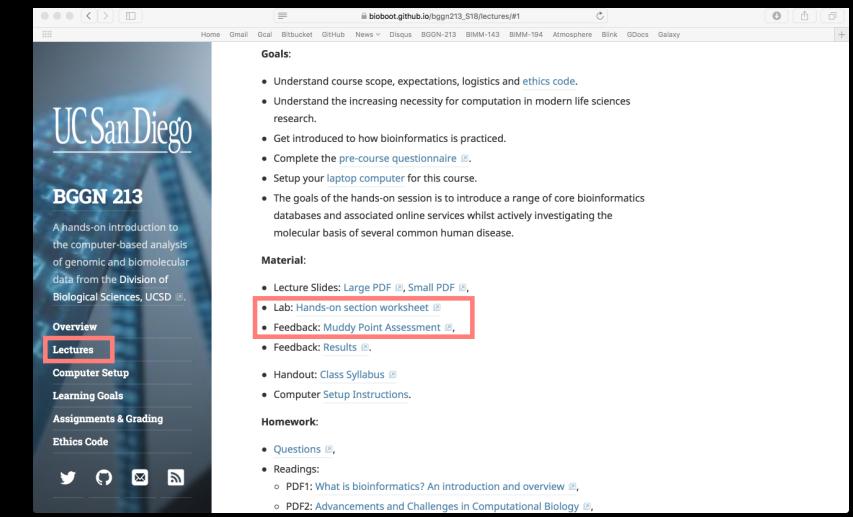
Introducing the *what, why and how* of bioinformatics?

Bioinformatics Database

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

Your Turn!

https://bioboot.github.io/bggm213_S18/lectures/#1



The screenshot shows a dark-themed website for BGGN 213. At the top, there is a navigation bar with links to Home, Gmail, Gcal, Bitbucket, GitHub, News, Discus, BGGN-213, BMM-143, BMM-194, Atmosphere, Blink, GDocs, and Galaxy. Below the navigation, there is a header with the UC San Diego logo and the course title 'BGGN 213'. A sidebar on the left lists sections: Overview, Lectures (which is currently selected and highlighted in red), Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. The main content area contains a list of goals, material, and homework sections, each with a brief description and a link.

BGGN-213: FOUNDATIONS OF BIOINFORMATICS (Lecture 1)

Bioinformatics Databases and Key Online Resources

https://bioboot.github.io/bggm213_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, GENIE 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
ATGGCTGCACTCTGACTCTGTGGAGAACGTCGGCTTACCTGCCCTGTGGGGCAAGGTGAACGTTGGATGAAG
TTGGTGCGTGAGGCCCTGGCAGGCTGCTGGTGGCTACCCCTGGACGCCAGGGTTCTTGAGTCCTTGG
GGATCTGTCGACTCTCGTGAATGGCAACCTTAAGGTGAAGGCTCATGGCACAGAAAGTGTCTGGT
GCCCTTAACTGATGGCCCTGGCTCACCTGGCACACCTCAAGGGCACCTTGGCACACCTGAGGTGAGCTGCACT
GTGACAGCTGCACGCTGGAGACTTCAAGGCTCCGGGCAACCTGCTGGTGTGGTGTGGCCA
TCACCTTGGCAAAGAATTCACCCCAACAGTCAGGCTGCCATCAGAAAGTGGTGGCTGGTGTGGCTAAAT
GCCCTGGCCACAAAGTACTACAAGCTCGCTTCTGGCTGTGCAATT
```

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**
2. GENE database @ **NCBI**
— BREAK —
3. UniProt & Muscle @ **EBI**
4. PFAM, PDB & NGL
— BREAK —
5. Extension exercises

End times:
[2:35 pm]
[2:55 pm]
— 3:10 pm —
[3:30 pm]
[4:00 pm]
— 4:10 pm —
[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.
- 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**:

