



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. The header features the UC San Diego logo and the course title "BGGN 213". Below the title, a brief description states: "A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD". A sidebar on the left contains links for "Overview", "Lectures", "Computer Setup", "Learning Goals", "Assignments & Grading", and "Ethics Code". Social media icons for Twitter, GitHub, and LinkedIn are also present. The main content area is titled "Bioinformatics (BGGN 213, Spring 2018)" and includes sections for "Course Director" (Prof. Barry J. Grant), "Instructional Assistant" (Yuanheng Zhou), and "Course Syllabus" (Spring 2018 PDF). A DNA helix icon is in the top right.

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

This screenshot is identical to the one above, showing the BGGN 213 homepage. However, the "Learning Goals" link in the sidebar has been highlighted with a red box. The rest of the page content, including the main bioinformatics section and the DNA helix icon, remains the same.

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

The screenshot shows the "Learning Goals" page for the BGGN 213 course. The sidebar highlights the "Learning Goals" link. The main content area starts with a statement: "At the end of this course students will:" followed by a bulleted list of nine items detailing what students should attain. Below this, 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." A DNA helix icon is in the top right.

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, 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!

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

BGGN 213

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

Lectures

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

#	Date	Topics for Spring 2018
1	Wed, 04/04	Welcome to Bioinformatics Course introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Hands on with major Bioinformatics databases and key online NCBI and EBI resources
2	Fri, 04/06	Sequence alignment fundamentals, algorithms and applications Homology, Sequence similarity, Local and global alignment, classic Needleman-Wunsch, Smith-Waterman and BLAST heuristic approaches, Hands on with dot plots, Needleman-Wunsch and BLAST algorithms highlighting their utility and limitations

Course Structure

Derived from specific learning goals

BGGN 213

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

Lectures

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

#	Date	Topics for Spring 2018
1	Wed, 04/04	Welcome to Bioinformatics Course introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Hands on with major Bioinformatics databases and key online NCBI and EBI resources
2	Fri, 04/06	Sequence alignment fundamentals, algorithms and applications Homology, Sequence similarity, Local and global alignment, classic Needleman-Wunsch, Smith-Waterman and BLAST heuristic approaches, Hands on with dot plots, Needleman-Wunsch and BLAST algorithms highlighting their utility and limitations

Class Details

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

The screenshot shows the course homepage with a sidebar containing links for Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, and Screen Cast Videos. The main content area displays the first lecture titled '1: Welcome to Foundations of Bioinformatics'. It includes sections for Topics, Goals, and Material, each with a bulleted list of items.

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 the course homepage with a sidebar containing links for Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, and Screen Cast Videos. The main content area displays the 'Homework' section. It includes sections for Homework, Readings, and Screen Casts.

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://integrabionline.org/bggm213>

Homework

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

The screenshot shows the course homepage with a sidebar containing links for Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, and Screen Cast Videos. The main content area displays the 'Homework' section. A red box highlights the 'Questions' link under the Homework heading.

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://integrabionline.org/bggm213>

Homework

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

The screenshot shows a Google Form titled 'BGGN213 Lecture 1 Homework (F17)'. The form has a purple header and asks for the user's UCSD username/email address. It also contains a question about the most frequently used operating system for bioinformatics tool development, with options for Windows, iOS, Unix, and Perl.

BGGN213 Lecture 1 Homework (F17)

Please answer the following questions

* Required

Your UCSD username/email address *

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

Your answer

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

Windows

iOS

Unix

Perl

Homework

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

The screenshot shows a Google Form interface. At the top, it says "BGGN213 Lecture 1 Homework". Below that, there's a red banner with the text "Homework is due before the next weeks class!". The form contains a question: "Which of the following operating systems is most frequently used for bioinformatics tool development?" with options: Windows, iOS, Unix, and Perl. There is a field for "Your answer" and a "Submit" button at the bottom.

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 a table of learning goals. The first column lists numbered goals from 5 to 12. The second column describes the goal, and the third column lists associated numbers. A green box highlights the last four goals (6, 7, 8, 9). A red arrow points down to the right side of the table.

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 a table of learning goals. The first column lists numbered goals from 13 to 20. The second column describes the goal, and the third column lists associated numbers. A green box highlights the last five goals (15, 16, 17, 18, 19). A red arrow points down to the right side of the table.

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

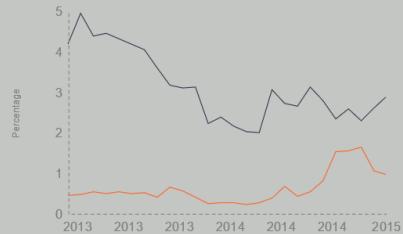
Language Rank	Types	Spectrum Ranking
1. C	⌚💻📱	100.0
2. Java	🌐⌚💻	98.1
3. Python	🌐💻	98.0
4. C++	⌚💻📱	95.9
5. R	💻	87.9
6. C#	🌐⌚💻	86.7
7. PHP	🌐	82.8
8. JavaScript	🌐⌚	82.2
9. Ruby	🌐💻	74.5
10. Go	🌐💻	71.9

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

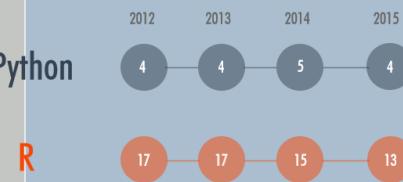
R and Python: The Numbers

Popularity Rankings

R and Python's popularity between 2013 and February 2015 (Tibbe 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

- 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 notification sidebar on the right. A red circle highlights the notification icon in the top right corner of the header. The sidebar lists several notifications:

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

See all notifications

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

The screenshot shows a DataCamp course page titled "What is an IDE anyway?". The question text is:

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 (-1xp)

A red circle highlights the "Integrated Development Environment" option. At the bottom right, a red circle highlights the "Submit Answer" button.

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

The screenshot shows a DataCamp course page with a modal window titled "Exercise Completed". The message says:

Exercise Completed
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!

PRESS ENTER TO CONTINUE

A red circle highlights the "PRESS ENTER TO CONTINUE" button. Below it is a "Become a power user!" section with a "Submit Answer" button, which is also highlighted with a red circle.

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

The screenshot shows the DataCamp website's group details page for 'Foundations of Bioinformatics (BGGN-213)'. The 'Groups' tab in the top navigation bar is circled in red. The main content area displays a leaderboard and assignment details.

Leaderboard

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

My Assignments

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

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

The screenshot shows the same DataCamp website page, but the 'My Assignments' tab in the top navigation bar is circled in red. The main content area displays the same assignment list as the previous screenshot.

My Assignments

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

Today's Menu

Course Logistics	Website, screencasts, survey, ethics, assessment and grading.
Learning Objectives	What you need to learn to succeed in this course.
Course Structure	Major lecture topics and specific learning goals.
Introduction to Bioinformatics	Introducing the <i>what</i> , <i>why</i> and <i>how</i> of bioinformatics?
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

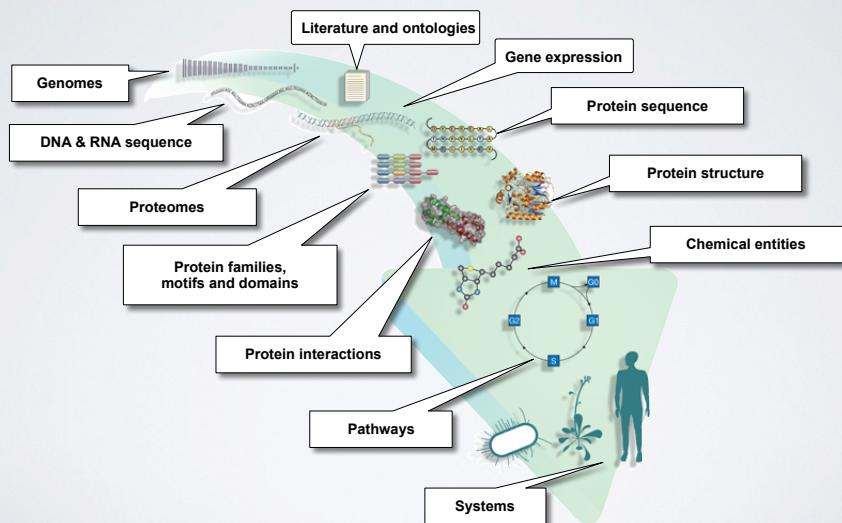
- “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “**informatics**” techniques (derived from disciplines such as applied mathematics, computer science, and statistics) to **understand** and **analyze** the information associated with these molecules, on a large-scale.

Luscombe NM, et al. Methods 1999; 21:40:346.

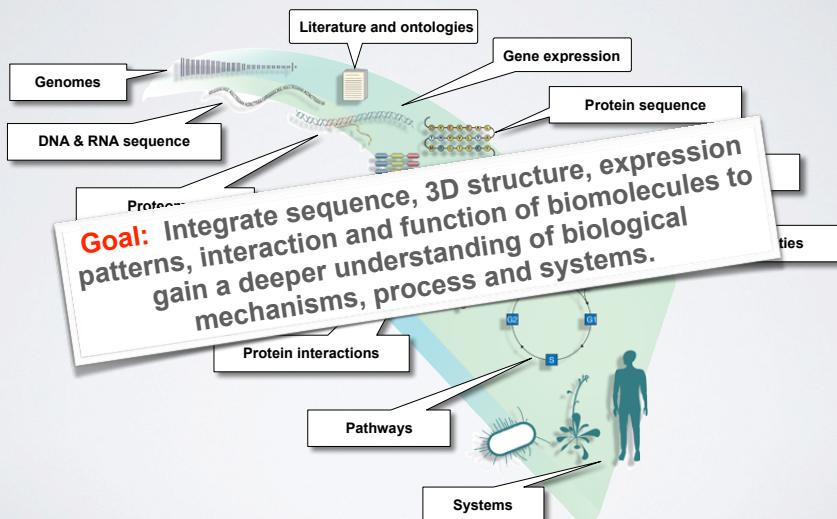
- “Bioinformatics is research, development, or application of **computational approaches** for expanding the use of **biological, medical, behavioral or health data**, including those to acquire, store, organize and analyze such data.”

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

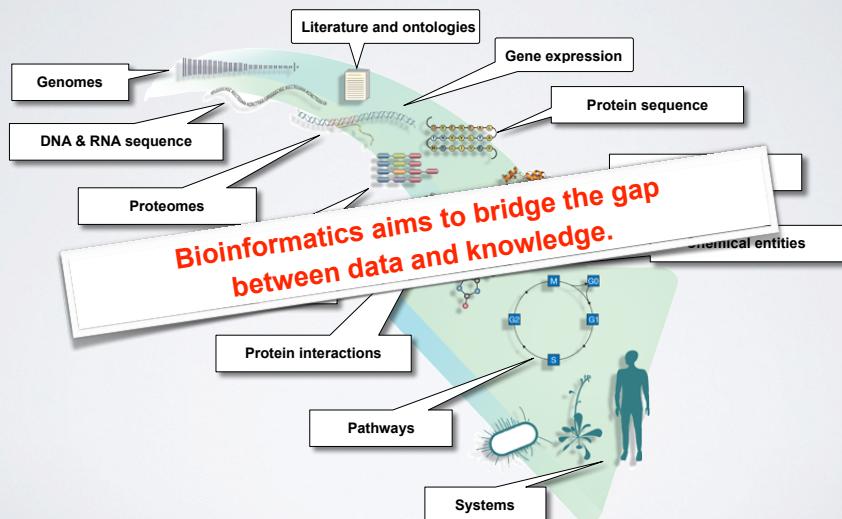
Major types of Bioinformatics Data



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

Where did bioinformatics come from?

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

Recap: The key dogmas of molecular biology

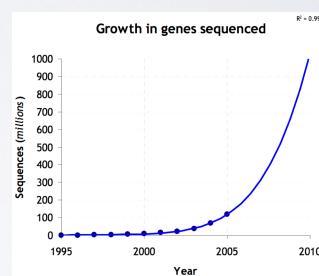
- DNA sequence determines protein sequence.
- Protein sequence determines protein structure.
- Protein structure determines protein function.
- Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function *in space and time*.

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

Why do we need Bioinformatics?

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

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

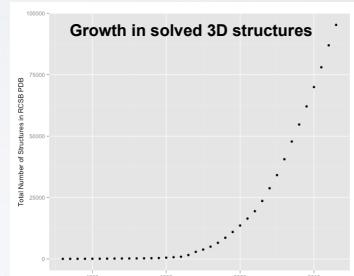


E.G. data from sequencing, structural genomics, proteomics, new high throughput assays, etc...

Why do we need Bioinformatics?

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

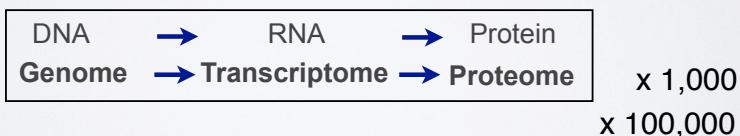
- Bioinformatics provides methods for the efficient:
 - storage
 - annotation
 - search and retrieval
 - data integration
 - data mining and analysis



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

How do we do Bioinformatics?

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



How do we *actually* do Bioinformatics?

Pre-packaged tools and databases

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

Advanced tool application & development

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

How do we *actually* do Bioinformatics?

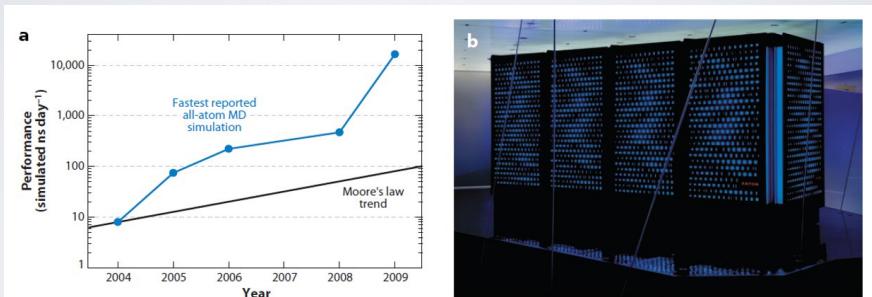
Pre-packaged tools and databases

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

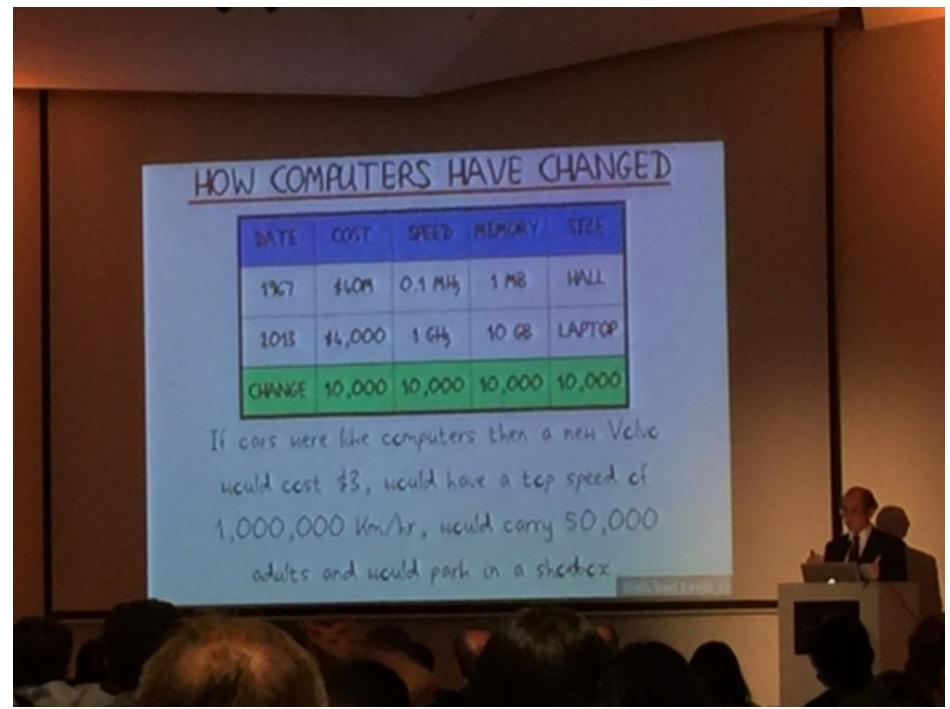
Advanced tool application & development

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

SIDE-NOTE: SUPERCOMPUTERS AND GPUS



SIDE-NOTE: SUPERCOMPUTERS AND GPUS



NSF Extreme Science and Engineering Discovery Environment (XSEDE)

www.xsede.org/community-engagement/educator-p...

XSEDE

About For Users Ecosystem Community Engagement News XUP

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

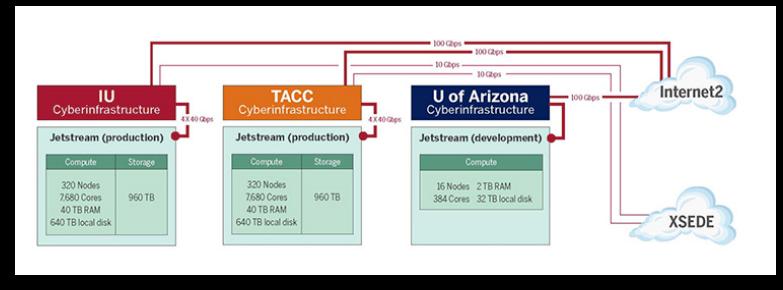
- XSEDE sponsors full-semester online courses
- Collaborations with faculty at participating institutions
- Campus visits offer guidance concerning course content

Related Links

- Diversity and Inclusion
- Student Engagement
- Campus Champions
- XSEDE Scholars Program

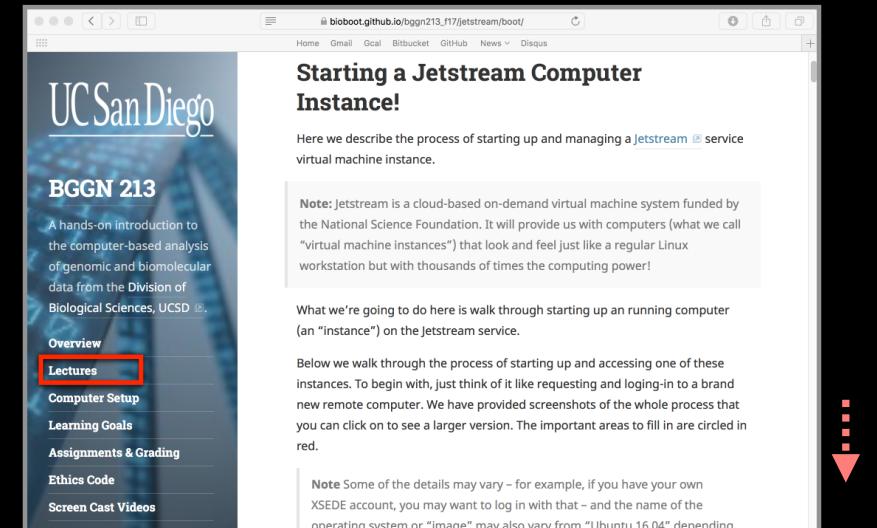
What is *Jetstream*?

- 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.

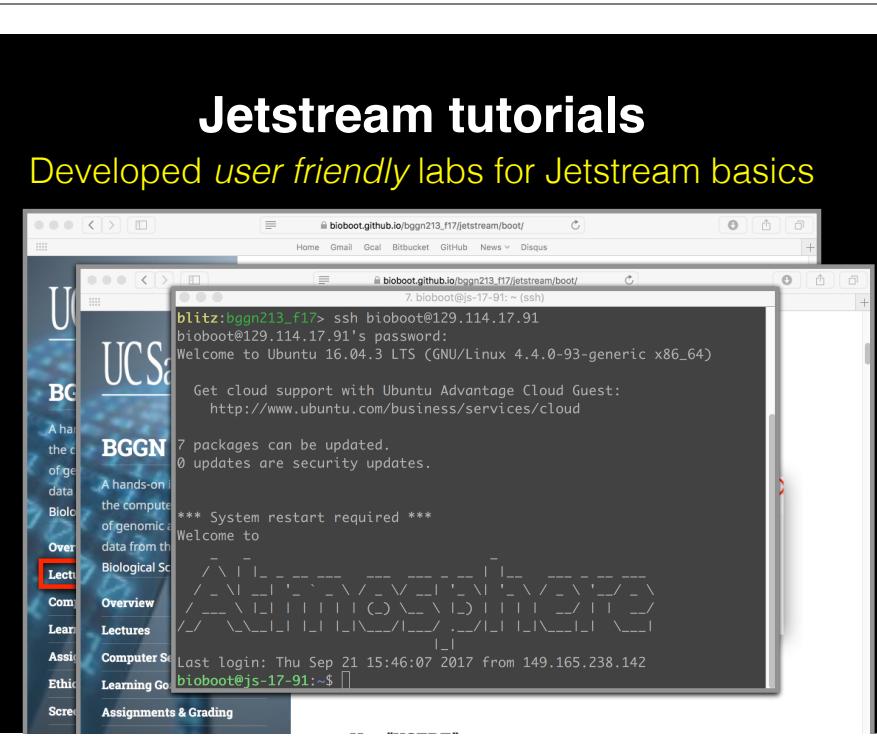


Jetstream tutorials

Developed *user friendly* labs for Jetstream basics



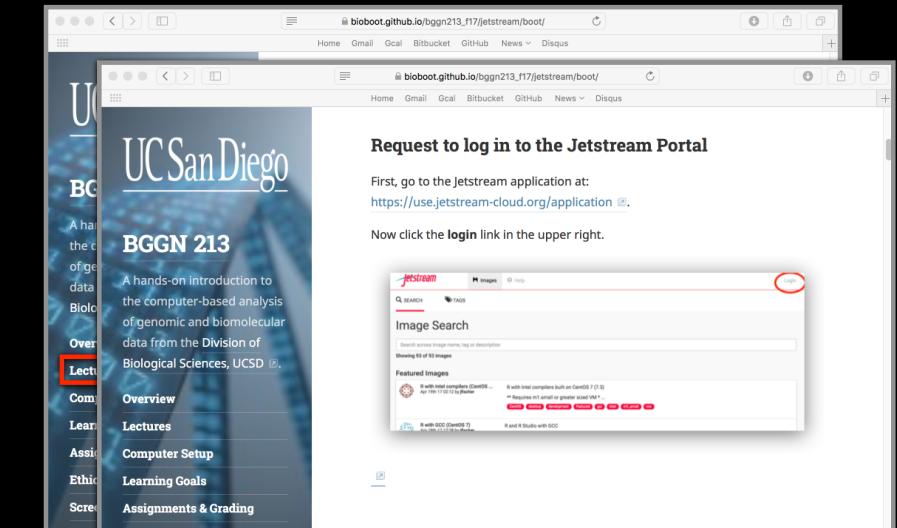
The screenshot shows a web browser window with the URL https://bioboot.github.io/bggm213_f17/jetstream/boot/. The main content is titled "Starting a Jetstream Computer Instance!". It describes the process of starting up and managing a Jetstream service virtual machine instance. A note states that Jetstream is a cloud-based on-demand virtual machine system funded by the National Science Foundation. Below this, there's a section titled "What we're going to do here is walk through starting up an running computer (an "instance") on the Jetstream service." Another note at the bottom indicates that some details may vary depending on the operating system or "image". The sidebar on the left has a navigation menu with "Lectures" highlighted with a red box.



The screenshot shows a terminal session within a browser window. The user has logged in via SSH to a Jetstream instance, with the command `blitz:bggm213_f17> ssh bioboot@129.114.17.91`. After entering the password, the user is welcomed to Ubuntu 16.04.3 LTS (GNU/Linux 4.4.0-93-generic x86_64). The user then runs the command `sudo apt-get update`, which shows 7 packages can be updated and 0 updates are security updates. Finally, the user runs `sudo apt-get upgrade`, which requires a system restart. The terminal session ends with the command `bioboot@js-17-91:~$`.

Jetstream tutorials

Developed *user friendly* labs for Jetstream basics



The screenshot shows a web browser window with the URL https://bioboot.github.io/bggm213_f17/jetstream/boot/. The main content is titled "Request to log in to the Jetstream Portal". It instructs the user to go to the Jetstream application at <https://use.jetstream-cloud.org/application> and click the "login" link in the upper right. Below this, there's a screenshot of the Jetstream portal's "Image Search" page, which shows a search bar and a list of featured images related to Intel computers built on CentOS 7.0 R.

Skepticism & Bioinformatics

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

- Do they make sense?
- Is it what we expected?
- Do we have adequate controls, and how did they come out?
- Modeling is modeling, but biology is different...
What does this model actually contribute?
- Avoid the miss-use of 'black boxes'

Skepticism & Bioinformatics

Gunnar von Heijne in “*Sequence Analysis in Molecular Biology; Treasure Trove or Trivial Pursuit*” states:

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

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

The screenshot shows the Protein BLAST search interface. It includes sections for General Parameters (Max target sequences: 500, Short queries checked, 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 checked, Mask: Mask for lookup table only checked, Mask lower case letters checked), and PSI/PHI/DELTA BLAST (Upload PSSM Optional, PSI-BLAST Threshold: 0.005, Pseudocount: 0). A callout box highlights the text "Even Blast has many settable parameters". Another callout box highlights the text "Related tools with different terminology" pointing to a section labeled "STEP 3 - Set your PROGRAM" with options FASTA and HMMER.

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)

Key Online Bioinformatics Resources: NCBI & EBI

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

The screenshot shows the NCBI homepage. It features a search bar, links for "All Databases", "Get Started", and "Popular Resources" (PubMed, BioProject, PubMed Central, PubMed Health, BLAST, NCBI Books, Genome, SNP, Gene, Variation, PubChem). There are also sections for "3D Structures" and "NCBI Announcements". A callout box highlights the URL <http://www.ncbi.nlm.nih.gov>.

The screenshot shows the European Bioinformatics Institute (EMBL-EBI) homepage. It features a search bar, links for "Services", "Research", "Industry", "European Cooperation", "Visit EMBL.org", "Upcoming events", "Part of the European Molecular Biology Laboratory", and "Periodicals". A callout box highlights the URL <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>

The screenshot shows the NCBI homepage with the "Popular Resources" section highlighted. It includes links to PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. To the right, there is a "Welcome to NCBI" summary and a "Get Started" section with links to various tools and databases. A "3D Structures" section is also visible.

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

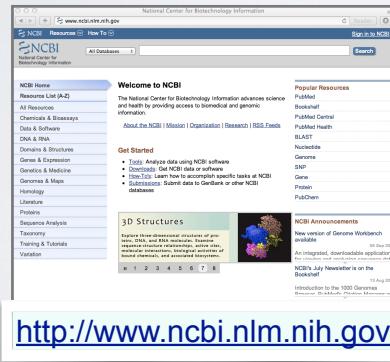
The screenshot shows the NCBI homepage with several red arrows pointing to specific links in the "Popular Resources" section. Arrows point to "PubMed", "BLAST", "Nucleotide", "Gene", and "Protein". A bracket on the right side groups "Nucleotide", "Gene", and "Protein". Other links in the "Popular Resources" section include Bookshelf, PubMed Central, PubMed Health, and PubChem. The "Welcome to NCBI" summary and "Get Started" section are also visible.

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

The screenshot shows the NCBI homepage with a large callout box in the center. The text in the box reads: "Notable NCBI databases include: GenBank, RefSeq, PubMed, dbSNP and the search tools ENTREZ and BLAST". The rest of the page includes the "Popular Resources" section, the "Welcome to NCBI" summary, and the "Get Started" section.

Key Online Bioinformatics Resources: NCBI & EBI

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

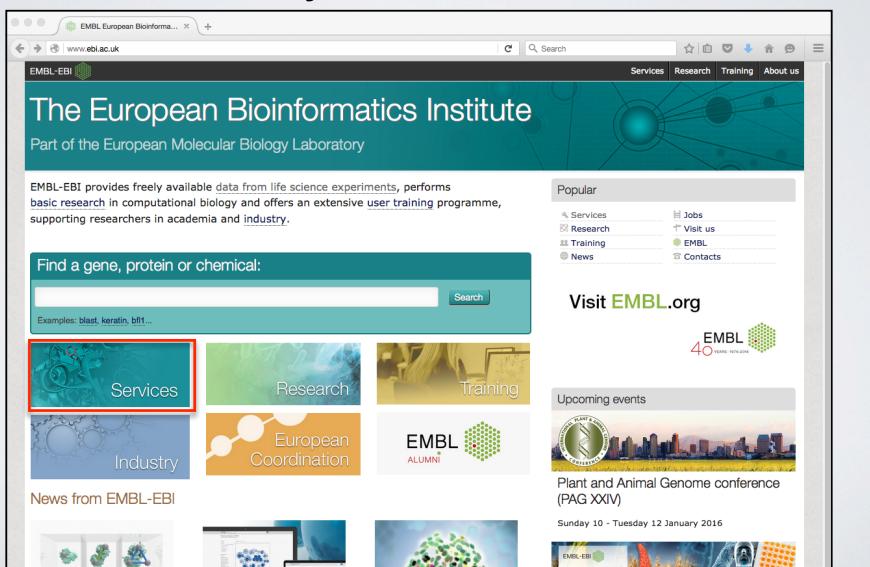


<http://www.ncbi.nlm.nih.gov>



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

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

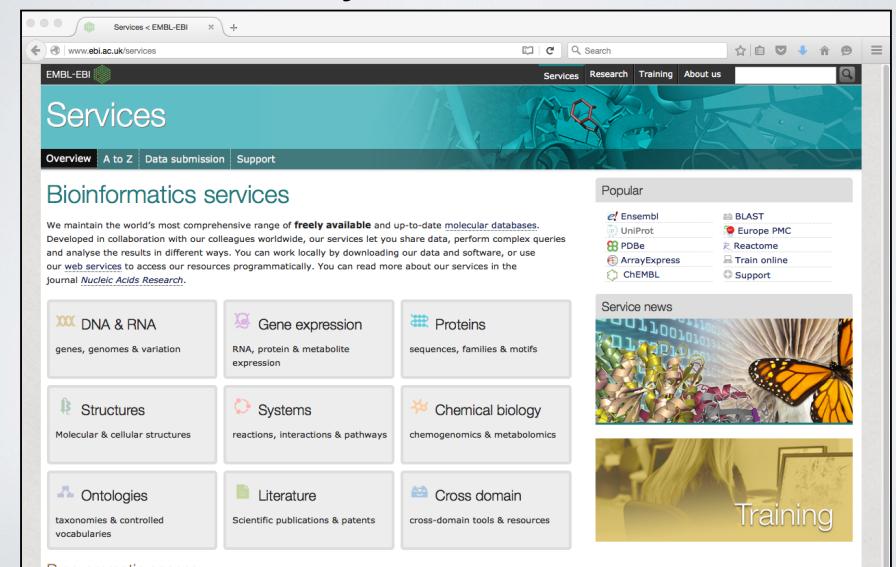


European Bioinformatics Institute (EBI)

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

Services

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*.

Popular

- Ensembl
- UniProt
- PDBe
- ArrayExpress
- ChEMBL
- Proteins
- DNA & RNA
- Gene expression
- Structures
- Systems
- Chemical biology
- Ontologies
- Literature
- Cross domain

Training

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

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

Proteins

Popular services

- UniProt: The Universal Protein Resource
- 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

Find a gene, protein or chemical:

Examples: blast, keratin, bfl1...

Services

Research

Training

European Coordination

EMBL ALUMNI

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

Train online

Using sequence similarity searching tools at EMBL-EBI: webinar

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

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

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

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

Find a course
Browse by subject
Genes and Genomes
Gene Expression
Interactions, Pathways and Networks

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, 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, Genilesne, 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, Klothe, 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 !!!

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, 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, Genilesne, 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, Klothe, 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 !!!

*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 <i>what, why</i> and <i>how</i> of bioinformatics?
Bioinformatics Database	Hands-on exploration of several major databases and their associated tools.

Your Turn!

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

The screenshot shows a web browser displaying the UC San Diego BGGM 213 course page. The page has a dark theme with blue and white text. At the top, there are navigation links for Home, Gmail, Calendar, Bitbucket, GitHub, News, Disqus, BGGM-213, BGGM-143, Atmosphere, Blink, Google Docs, and Galaxy. Below the header, there is a banner for the course. The main content area has a sidebar on the left with 'Goals', 'Material', and 'Homework' sections. The 'Goals' section lists course objectives. The 'Material' section highlights the 'Lecture Slides' (both large and small PDFs) and 'Lab: Hands-on section worksheet' (which is highlighted with a red box). The 'Homework' section lists 'Questions' and 'Readings' (with two PDF links). The 'Lectures' section is currently selected and highlighted with a red box. Other visible sections include 'Computer Setup', 'Learning Goals', 'Assignments & Grading', and 'Ethics Code'. Social media icons for Twitter, GitHub, and LinkedIn are at the bottom.

BGGN-213: FOUNDATIONS OF BIOINFORMATICS (Lecture 1)

Bioinformatics Databases and Key Online Resources
https://bioboot.github.io/bggm213_S18/lectures/#1

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.

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**
 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 —

[3:59 pm]

— 4:10 pm —

[4:40 pm]

- ▶ 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 | [~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!

SUMMARY

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

HOMEWORK

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

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

THANK YOU