

05 : 00

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 is a subheader "Bioinformatics (BGGN 213, Winter 2020)". A DNA helix icon is positioned to the right of the title. The main content area includes sections for "Course Director" (Prof. Barry J. Grant), "Instructional Assistant" (Daniela Castruita), and "Course Syllabus" (Fall 2020 PDF). A sidebar on the left lists course navigation links: Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. Social media icons for Twitter, GitHub, and LinkedIn are at the bottom.

<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 course title, director information, and syllabus, remains the same.

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

The screenshot shows the "Learning Goals" page for BGGN 213. The header includes the UC San Diego logo and the course title. The main content is titled "Learning Goals" and states: "At the end of this course students will:". Below this is a bulleted list of nine learning objectives. At the bottom, a note says: "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 sidebar on the left contains the same course navigation links as the homepage, with "Learning Goals" also highlighted by a red box. Social media icons are at the bottom.

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 '1: Welcome to Foundations of Bioinformatics' page. It includes a 'Topics' section with course introduction, learning goals, and expectations; a 'Goals' section listing course objectives; and a 'Material' section with links to pre-class screen casts, lecture slides, handouts, and computer setup instructions.

Homework

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

The screenshot shows the 'Homework' section of the website. It lists 'Questions', 'Readings', and a 'Screen Casts' section featuring a video of Barry Grant discussing the course.

Homework

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

The screenshot shows the 'Homework' section of the website. A red box highlights the 'Questions' link under the 'Homework:' heading. Other sections shown include 'Readings' and 'Screen Casts'.

Homework

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

The screenshot shows a Google Form titled 'Lecture 1 Homework'. It includes fields for 'Email address *', 'UCSD PID number (exam number)', and 'Your answer'. A question asks which operating system is most frequently used for bioinformatics tool development, with options for Windows and Mac.

Homework

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

The screenshot shows a Google Forms survey titled "Lecture 1 Homework". It includes fields for "Email address *", "UCSD PID number (exam number)", and "Your answer". A question asks: "Which of the following operating systems is most frequently used for bioinformatics tool development?" with options "Windows" and "ios". A red diagonal watermark across the form reads "Homework is due before the next weeks class!".

Projects

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

The screenshot shows a UC San Diego BGGN 213 course page. The "9: Unsupervised learning mini-project" section is highlighted. It includes a description: "Longer hands-on session with unsupervised learning analysis of cancer cells, Practical considerations and best practices for the analysis and visualization of high dimensional datasets.", goals, and a list of resources.

Projects

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

The screenshot shows a UC San Diego BGGN 213 course page. The "18: Cancer genomics" section is highlighted. It includes a description: "Cancer genomics resources and bioinformatics tools for investigating the molecular basis of cancer. Large scale cancer sequencing projects; NCI Genomic Data Commons; What has been learned from genome sequencing of cancer? Immunoinformatics, Immunotherapy and cancer; Using genomics and bioinformatics to harness a patient's own immune system to fight cancer. Implications for the development of personalized medicine.", a note about finding a gene assignment, material, and a list of resources.

Projects

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

The screenshot shows a UC San Diego BGGN 213 course page. The "10: Project: Find a gene assignment (Part 1)" section is highlighted. It includes a description: "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.", a note about consulting the scoring rubric, a note about responses due at the beginning of class on Feb 22nd, a note about the complete assignment due at the beginning of class on March 13th, and a note about late responses not being accepted.

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!

The screenshot shows a GitHub repository page titled "Bioinformatics Class BIMM-143". The main heading is "Introduction to Bioinformatics Class S18". Below it is a stylized DNA helix icon with a magnifying glass over it, showing the numbers "101" and "110". A text box states: "A repository to store and display my work completed during the Spring 2018 quarter in BIMM-143 at UCSD." A link "View the Project on GitHub" and "jasonPBennett/bimm143" are provided. To the right is a sidebar titled "Index of Material" listing 16 class topics from "Working With R" to "Transposons: A Sample Workflow".

Online portfolio of **your** bioinformatics work!

The screenshot shows a project page for "class13" by Jason Patrick Bennett, dated May 15, 2018. The title is "Identifying SNP's in a Population". It includes a code snippet for reading a CSV file and a table output. Below the table is a large block of R code for analyzing SNP data from the Mexican-American population in Los Angeles.

```
genotype <- read.csv("373531-SampleGenotypes-Homo_sapiens_Variation_Sample_rs8067378.csv")  
  
Now lets look at a table of the data:  
  
table(genotype)  
  
## , Population.s. = ALL, AMR, MXL, Father = -, Mother = -  
## Genotype..forward.strand.  
## Sample..Male.Female.Unknown, A|A A|G G|A G|C  
## 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!

The screenshot shows a project page for "class13" featuring two plots. The top plot is a density distribution of "exp" values for three genotypes: AA (pink), AG (light green), and GG (light blue). The bottom plot is a faceted boxplot of "exp" values for the same three genotypes, with each facet containing individual data points jittered around the boxplots. A legend indicates the genotypes: AA (red), AG (green), and GG (blue).

```
ggplot(exp, aes(geno, exp, fill=geno)) +  
  geom_boxplot(notch=TRUE, outlier.shape = NA) +  
  geom_jitter(shape=16, position=position_jitter(0.2), alpha=0.4)
```

Bonus:

Bioinformatics & Genomics in industry

21: Bonus: Bioinformatics & Genomics in industry

Friday March 15th at 1pm come and enjoy a set of short open ended guest lectures from leading genomic scientists at Illumina Inc., Synthetic Genomics Inc., Samumed and the La Jolla Institute for Allergy and Immunology. Come prepared for networking and to have your questions about industry careers in Bioinformatics and Genomics answered.

© 2019 Barry J. Grant. All rights reserved. A UCSD Division of Biological Sciences Course

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!**

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

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.

Learning Goal	Number
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

Why use R?

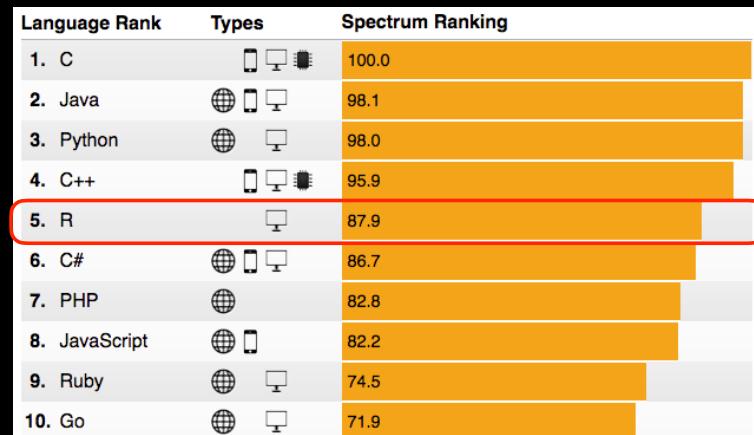
Productivity
Flexibility
Genomic data analysis

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

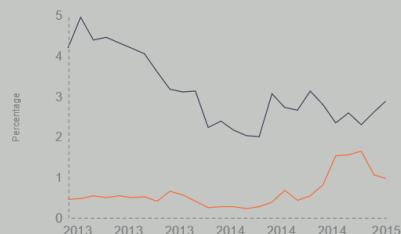


<http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages>

R and Python: The Numbers

Popularity Rankings

R and Python's popularity between 2013 and February 2015 (Folbe 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 designed specifically for data analysis

- Large friendly user and developer community.
- As of Jan 6th 2019 there are 15,352 add on **R packages** on **CRAN** and 1,823 on **Bioconductor** - much 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 RStudio

You have a new assignment: Working with the RStudio IDE... 16 days ago

You have a new assignment: Introduction to R... 16 days ago

bigrant invited you to the group Foundations of Data Science... 16 days ago

You have a new assignment: Orientation... 9 months ago

See all notifications

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

Console

R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"

Copyright (C) 2016 The R Foundation for Statistical Computing

Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.

You are welcome to redistribute it under certain conditions.

Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.

Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications.

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

Type 'q()' to quit R.

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

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 exercise 1: What is an IDE anyway?". Below this, there is a "Continue" button with a red circle around it. On the right, the RStudio environment shows the console, environment, and global history panes. The console pane contains R code and output related to the exercise.

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

A screenshot of the DataCamp Leaderboard page for the course "BGGN213_F19". The page displays a table of student performance over the past year. The columns include Email, Name, Courses Completed, Chapters Completed, and XP Points. A search bar and filter options for "30 DAYS", "90 DAYS", and "PAST YEAR" are visible at the top. A red circle highlights the "PAST YEAR" filter button.

Email	Name	Courses Completed	Chapters Completed	XP Points
akoehler@ucsd.edu	Alanna Koehler	8	40	48980
osongste@ucsd.edu	Livia Songster	8	39	48320
picheng@ucsd.edu	Pin-Chung (Tony) Cheng	10	47	47324
pberube@ucsd.edu	Peter Berube	7	33	35398
k7lee@ucsd.edu	Kat Lee	6	28	30000
ktmiyamo@ucsd.edu	Kiana Miyamoto	4	19	26600
ttsin@ucsd.edu	Tat Hei Tsin	4	19	26305
amferry@ucsd.edu	Amir Ferry	4	23	24608

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

Homework assignments will be via DataCamp

A screenshot of the DataCamp RStudio IDE interface. The main pane shows an exercise titled "PCA analysis" with instructions to perform a PCA on a dataset. The code editor pane contains R code for transforming the dataset and performing PCA. The R Console pane shows the execution of the code, with an error message indicating that the object 'vsd_smoc2' was not found. The Slides pane is also visible.

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.

“What is Bioinformatics?”

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

... A hybrid of biology and computer science

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

Bioinformatics is computer aided biology!

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

Bioinformatics is computer aided biology!

Goal: Data to Knowledge

There are many useful definitions...

- "Computer based **management** and **analysis** of biological and biomedical data with useful applications in many disciplines, particularly **genomics, proteomics, metabolomics**, and related fields."
(BGGN-213)
- "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 et al. 2001)
- "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 ...<cut>..."
(National Institutes of Health: <http://tinyurl.com/l3gxr6b>)

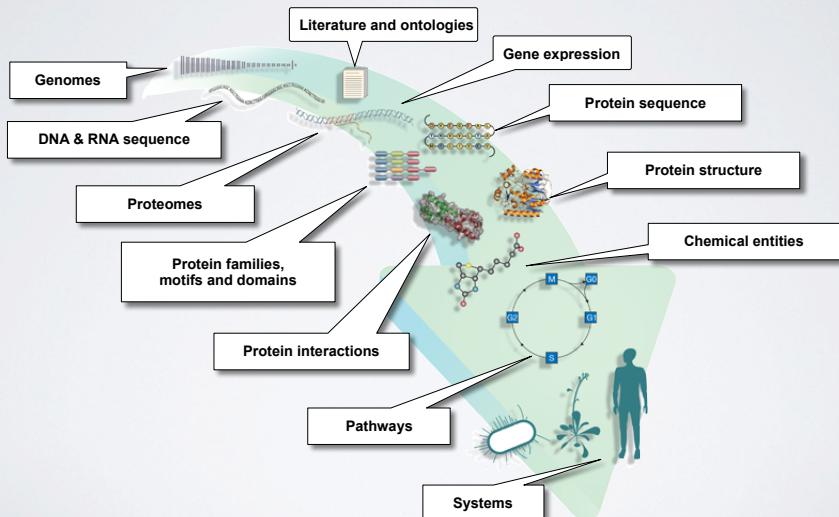
Side-Note:

There are many useful definitions...

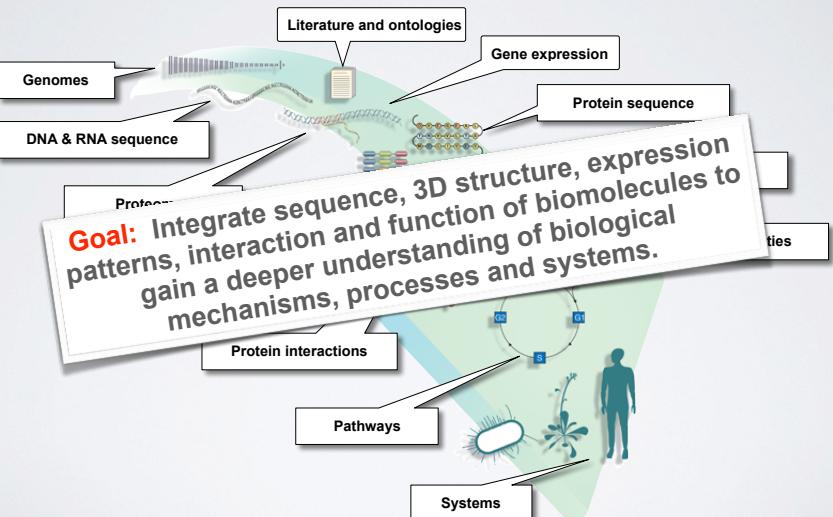
- "Computer based **management** and **analysis** of biological and biomedical data with useful applications in many disciplines, particularly **genomics, proteomics, metabolomics**, and related fields."
(BGGN-213)
- "Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying "**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 et al. 2001)
- "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 ...<cut>..."
(National Institutes of Health: <http://tinyurl.com/l3gxr6b>)

Side-Note:

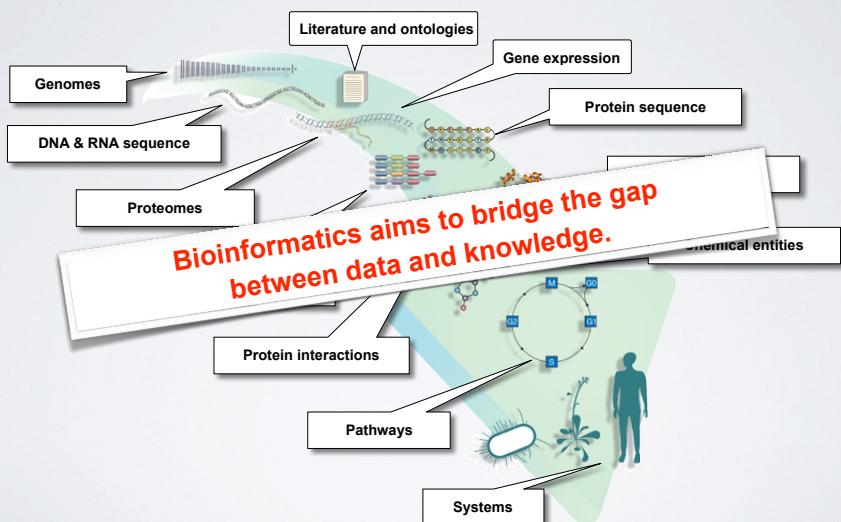
Major types of Bioinformatics Data



Major types of Bioinformatics Data

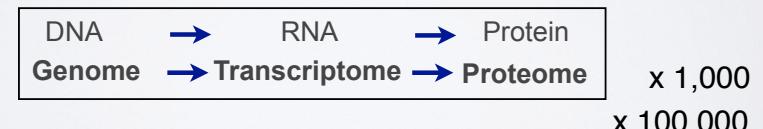


Major types of Bioinformatics Data



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 programing languages frequently required (e.g. R, Python, Perl, C, Java, Fortran)
- May require specialized high performance computing...

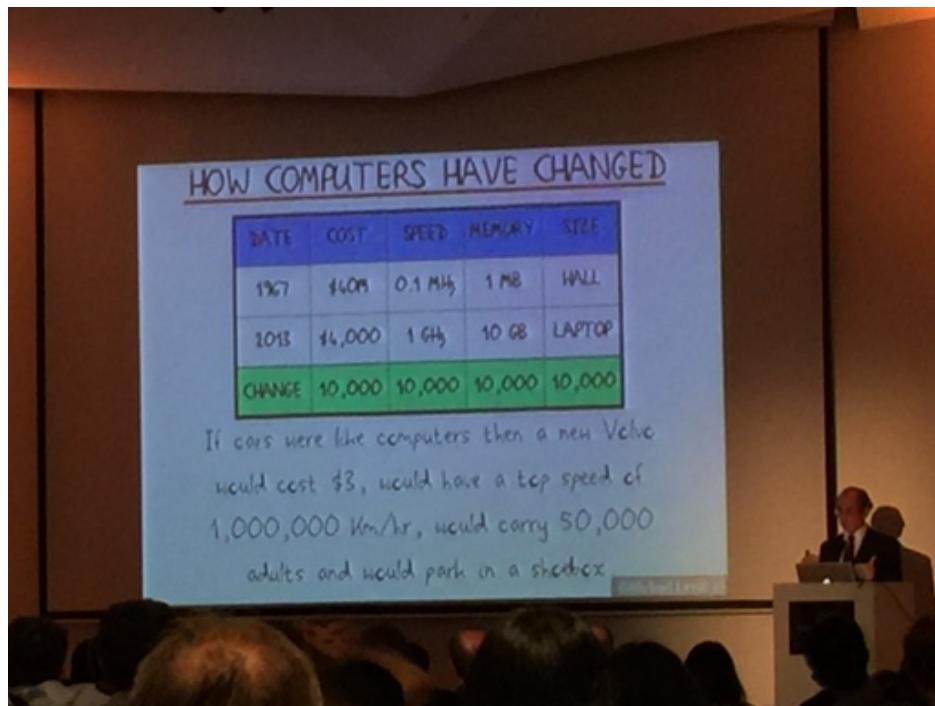
How do we *actually* do Bioinformatics?

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 programing languages frequently required (e.g. R, Python, Perl, C, Java, Fortran)
- May require specialized high performance computing...



NSF Extreme Science and Engineering Discovery Environment (XSEDE)

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

The XSEDE website features a dark header with navigation links: Home, Gmail, Gcal, Bitbucket, Github, News, Discus, About, For Users, Ecosystem, Community Engagement (which is highlighted), News, and XUP. The main content area has a background image of a starry galaxy. It includes sections for 'Curriculum and Educator Programs' and 'Campus Visits'. A sidebar on the right lists 'Key Points' and 'Related Links'.

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

Jetstream tutorials

Developed *user friendly* labs for Jetstream basics

The image contains two side-by-side screenshots of a web browser window. Both screenshots show the URL bioboot.github.io/bggm213_f17/jetstream/boot/.
The left screenshot displays a "Request to log in to the Jetstream Portal" page. It includes instructions to go to the Jetstream application at <https://use.jetstream-cloud.org/application> and click the "login" link in the upper right. Below this is a screenshot of the Jetstream search interface, showing a search bar and a grid of images labeled "Featured Images".
The right screenshot shows a terminal session on a cloud instance. The terminal prompt is "blitz:ggm213_f17>". It shows the user logging in with their password, entering "Welcome to Ubuntu 16.04.3 LTS (GNU/Linux 4.4.0-93-generic x86_64)". The user then runs "apt update" and "apt upgrade", which lists 7 packages to be updated and 0 security updates. A message indicates a system restart is required. The session ends with "Last login: Thu Sep 21 15:46:07 2017 from 149.165.238.142" and the user's name "bioboot@js-17-91:~\$".

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'

Jetstream tutorials

Developed *user friendly* labs for Jetstream basics

The image shows a single screenshot of a terminal session on a cloud instance. The terminal prompt is "blitz:ggm213_f17>". The user logs in with their password and enters "Welcome to Ubuntu 16.04.3 LTS (GNU/Linux 4.4.0-93-generic x86_64)". The user then runs "apt update" and "apt upgrade", which lists 7 packages to be updated and 0 security updates. A message indicates a system restart is required. The session ends with "Last login: Thu Sep 21 15:46:07 2017 from 149.165.238.142" and the user's name "bioboot@js-17-91:~\$".

Skepticism & Bioinformatics

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

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

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

Common problems with Bioinformatics

Confusing multitude of tools available

- Each with many options and settable parameters

Most tools and databases are written by and for nerds

- Same is true of documentation - if any exists!

Most are developed independently

Notable exceptions are found at the:

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

The screenshot shows the 'Protein BLAST' search interface from NCBI. 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, Mask: Mask for lookup table only, Mask lower case letters), and 'PSI/PHI/DELTA BLAST' (Upload PSSM Optional, PSI-BLAST Threshold: 0.005, Pseudocount: 0). A callout box highlights the 'Even Blast has many settable parameters'.

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, navigation links like 'All Resources', 'Databases', 'Software', 'Gene & Variation', 'Genomes & Maps', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Visualization'. Below the search bar, there's a '3D Structures' section with a 3D molecular model and a link to 'http://www.ncbi.nlm.nih.gov'.

The screenshot shows the European Bioinformatics Institute (EBI) homepage. It features a search bar, navigation links like 'Popular Resources', 'Services', 'Research', 'Training', and 'About us'. Below the search bar, there's a 'Find a gene, protein or chemical?' search form and a 'Visit EMBL.org' button. The page also displays news from EMBL-EBI and upcoming events like 'Plant and Animal Genome conference (PAG XXV)' and 'SME Forum 2016'.

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

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

National Center for Biotechnology Information (NCBI)

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



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

National Center for Biotechnology Information

NCBI Home | Resources | How To | Sign in to NCBI

All Databases | Search

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

Popular Resources

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

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 bioactivities.

NCBI Announcements

New version of Genome Workbench available

06 Sep

An integrated, downloadable application

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

National Center for Biotechnology Information

NCBI Home | Resources | How To | Sign in to NCBI

All Databases | Search

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

Popular Resources

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

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 bioactivities.

NCBI Announcements

New version of Genome Workbench available

06 Sep

An integrated, downloadable application

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

National Center for Biotechnology Information

NCBI Home | Resources | How To | Sign in to NCBI

All Databases | Search

Welcome to NCBI

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

Notable NCBI databases include:

GenBank, RefSeq, PubMed, dbSNP

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

Popular Resources

- PubMed

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 bioactivities.

NCBI Announcements

New version of Genome Workbench available

06 Sep

An integrated, downloadable application

Key Online Bioinformatics Resources: NCBI & EBI

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

National Center for Biotechnology Information

NCBI Home | Resources | How To | Sign in to NCBI

All Databases | Search

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

Popular Resources

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

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 bioactivities.

NCBI Announcements

New version of Genome Workbench available

06 Aug 2011

NCBI Newsletter is on the Bookshelf!

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

The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

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

Find a gene, protein or chemical:

Services | Research | Training | About | News

Visit EMBL.org

Upcoming events

Part and Animal Genome conference (PANG X00)

SIB Forum 2016

Periodic

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

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

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

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, BiolImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlesne, 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, ICTVbase, 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 !!!

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, 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, Genlesne, 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, ICTVbase, 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.

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.

Hands-on section

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

BGGN-213: FOUNDATIONS OF BIOINFORMATICS (Lecture 1)

Bioinformatics Databases and Key Online Resources
https://bioboot.github.io/bggn213_S18/lectures/#1
Dr. Barry Grant

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

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

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

Section 1
The following transcript was found to be abundant in a human patient's blood sample.

```
>example1
ATGGTGCATCTGACTCTCTGTGGAGAAGTCGCGTTACTGCCCTGTGGGGCAAGGTGAACCGTGATGAA
TTGGTGGTGGAGCCCTGGCAGGCTGCTGGTGGTCTACCCCTGGACCCAGAGGTCTTGGAGTCCTTGG
GGATCTGTCACACTCCGTAGCAAGCTTAAGTGAAGGCTCATGGCAAGAAAGTGCCTCGGT
GGCTTCTAGTGTAGCCGTGGCTCACCTGGACACCTCAAGGGCACCTTGGCACACTGAGCTGAGCTGCAC
GTGACAAGCTGACCTGGTGGAGACTTCAGGCTCTGGCAAGGGCACCTGGTGGCTGTCAGTGTGCTGGCC
TCACTTGGCAAAAGATTCAACCCCACAGTGCAGGCTGCTTACAGAAGTGGTGGCTGTGTCAGGCTAAT
GCCCTGGCCACACAGTATCATAAGTGTGCTTCTGCTGTCATTT
```

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	[2:35 pm]
2. GENE database @ NCBI	[2:55 pm]
— BREAK —	— 3:10 pm —
3. UniProt & Muscle @ EBI	[3:30 pm]
4. PFAM, PDB & NGL	[4:00 pm]
— BREAK —	— 4:10 pm —
5. Extension exercises	[4:40 pm]

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:

