

BGGN 213
Foundations of Bioinformatics
 Barry Grant
 UC San Diego
<http://thegrantlab.org/bggn213>

HELLO
my name is
BARRY
 bjgrant@ucsd.edu

HELLO
HER name is
DANIELA
 dsamanie@ucsd.edu

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</i> and <i>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 course website for BGGN 213. The left sidebar contains a navigation menu with 'Learning Goals' highlighted. The main content area shows the course title, instructor information, and an 'Overview' section. The 'Overview' text describes the course as a hands-on introduction to computational and analytical methods for biological problems.

http://thegrantlab.org/bggn213/

This screenshot is identical to the previous one, but the 'Learning Goals' link in the left sidebar is highlighted with a red box. The main content area remains the same, showing the course overview.

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

The screenshot shows the 'Learning Goals' page of the course website. The left sidebar has 'Learning Goals' highlighted. The main content area lists the learning objectives for the course, such as understanding the necessity for computation and using bioinformatics tools.

At the end of this course students will:

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

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

Specific Learning Goals....

What I want you to know by course end!

The screenshot shows the 'Specific Learning Goals' section of the BGGN 213 course page. It includes a table with the following data:

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

Course Structure

Derived from specific learning goals

The screenshot shows the 'Lectures' section of the BGGN 213 course page. It includes a table with the following data:

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

Course Structure

Derived from specific learning goals

The screenshot shows the 'Lectures' section of the BGGN 213 course page. It includes a table with the following data:

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

Class Details

Goals, Class material, Screencasts & Homework

The screenshot shows a web browser displaying the course page for BGGN 213. The page is titled "1: Welcome to Foundations of Bioinformatics". It includes a "Topics" section with a course introduction, "Goals" section with a list of objectives, and a "Material" section with links to pre-class screen cast, lecture slides, and handouts. A sidebar on the left contains navigation links for Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, Ethics Code, and Screen Cast Videos.

Homework

Goals, Class material, Screencasts & Homework

The screenshot shows the "Homework" section of the BGGN 213 course page. It lists "Questions" and "Readings" (PDF1, PDF2, and a New York Times article). Below the readings is a "Screen Casts" section featuring a video thumbnail titled "Welcome to 'Foundations of Bioinformatics' (BGGN-21...)" by Barry Grant. A comment below the video reads "1 Welcome to BGGN-213: Course introduction and logistics."

Homework

Goals, Class material, Screencasts & Homework

This screenshot is similar to the previous one but highlights the "Questions" link in the homework section with a red box. The rest of the page content, including the sidebar and video thumbnail, remains the same.

Homework

Goals, Class material, Screencasts & Homework

The screenshot shows a Google Forms assignment titled "Lecture 1 Homework". It asks students to answer questions including their email address and UCSD PID number. A required question asks: "Which of the following operating systems is most frequently used for bioinformatics tool development" with radio button options for "Windows" and "mac". The question is worth 1 point.

Homework

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

https://docs.google.com/forms/d/e/1FAIpQLSe4HkV2MmuPV/... 133%

Lecture 1 Homework

Please answer the following questions including your main @ucsd.edu email address and UCSD PID number so you can receive credit for your responses.

*** Required**

Email address *

UCSD PID number (exam number)

Your answer

Which of the following operating systems is most frequently used for bioinformatics tool development 1 point

- Windows
- Mac

Homework is due before the next weeks class!

Projects

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

BGGN 213

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

9: Unsupervised learning mini-project

Topics: 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:

- Be able to import data and prepare for unsupervised learning analysis.
- Be able to apply and test combinations of PCA, k-means and hierarchical clustering to high dimensional datasets and critically review results.

Material:

- Lecture Slides: **To Update** Large PDF, Small PDF
- Lab: Hands-on Worksheet
- Data file: WisconsinCancer.csv, new_samples.csv
- Bio3D PCA App: <http://bio3d.ucsd.edu/pca-app/>
- Feedback: Muddy-Point-Assessment

Projects

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

BGGN 213

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

18: Cancer genomics

Topics: 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.

N.B. Find a gene assignment due before next class!

Material:

- Lecture Slides: Large PDF, Small PDF
- Lab: **TO UPDATE** Hands-on Worksheet Part 1
- Lab: **TO UPDATE** Hands-on Worksheet Part 2
- Data files:
 - [lecture18_sequences.fa](#)

Projects

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

BGGN 213

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

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

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

You may wish to consult the scoring rubric at the end of the above linked project description and the **example report** for format and content guidance.

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

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

Late responses will not be accepted under any circumstances.

Why Projects?

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

Online portfolio of **your** bioinformatics work!

The screenshot shows a GitHub repository page for 'Bioinformatics Class BIMM-143'. The page title is 'Introduction to Bioinformatics Class S18'. Below the title is a logo featuring a DNA double helix and a magnifying glass over the numbers '101110'. The page content includes an 'Index of Material' with a list of 16 classes, from 'Introductory Material: Working With R' to 'Class 16 - Transposons: A Sample Workflow'. The repository is maintained by 'jasonPBennett'.

Online portfolio of **your** bioinformatics work!

The screenshot shows a GitHub repository page for 'class13'. The page title is 'class13' by 'Jason Patrick Bennett', dated 'May 15, 2018'. The main heading is 'Identifying SNP's in a Population'. Below the heading is a code block showing R code to read a CSV file and a table of the data. The table has columns for 'Sample..Male.Female.Unknown.', 'A|A|G|G|G|G', and 'Genotype..forward.strand.'. The code block shows the following R code:

```
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 = -  
##  
##      Sample..Male.Female.Unknown. A|A|G|G|G|G  
## NA19648 (F) 1 0 0 0  
## NA19649 (M) 0 0 0 1  
## NA19651 (F) 1 0 0 0  
## NA19652 (M) 0 0 0 1  
## NA19654 (F) 0 0 0 1  
## NA19655 (M) 0 1 0 0  
## NA19657 (F) 0 1 0 0  
## NA19658 (M) 1 0 0 0  
## NA19661 (M) 0 1 0 0  
## NA19663 (F) 1 0 0 0  
## NA19664 (M) 0 0 1 0  
## NA19666 (M) 1 0 0 0
```

Online portfolio of **your** bioinformatics work!

The screenshot shows a GitHub repository page for 'class13'. The page title is 'class13' by 'Jason Patrick Bennett', dated 'May 15, 2018'. The main heading is 'Identifying SNP's in a Population'. Below the heading is a density plot showing the distribution of 'exp' values for three genotypes: AA (red), AG (green), and GG (blue). The x-axis is 'exp' (ranging from 10 to 50) and the y-axis is 'density' (ranging from 0.00 to 0.04). Below the density plot is a boxplot showing the distribution of 'exp' values for the same three genotypes. The x-axis is 'exp' (ranging from 30 to 50) and the y-axis is 'exp' (ranging from 30 to 50). The boxplot shows that the AA genotype has the highest median 'exp' value, followed by AG, and then GG. Below the boxplot is a code block showing the R code used to create the density plot and boxplot:

```
ggplot(expr, 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

UC San Diego

BGGN 213

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

- Overview
- Lectures
- Computer Setup
- Learning Goals
- Assignments & Grading
- Ethics Code

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

UC San Diego

BGGN 213

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

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

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

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

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

Goal Number	Goal Description	Page Numbers
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 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.

Why use R?

Productivity
Flexibility
Genomic data analysis

IEEE 2016 Top Programming Languages

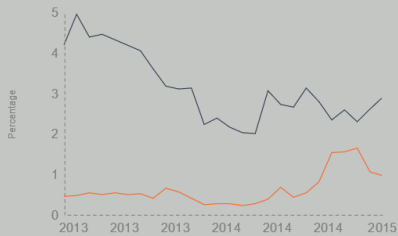
Language Rank	Types	Spectrum Ranking
1. C	📱 🖥️ 🧠	100.0
2. Java	🌐 📱 🖥️	98.1
3. Python	🌐 🖥️	98.0
4. C++	📱 🖥️ 🧠	95.9
5. R	🖥️	87.9
6. C#	🌐 📱 🖥️	86.7
7. PHP	🌐	82.8
8. JavaScript	🌐 📱	82.2
9. Ruby	🌐 🖥️	74.5
10. Go	🌐 🖥️	71.9

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

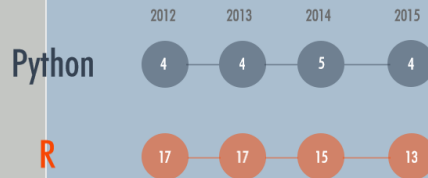
R and Python: The Numbers

Popularity Rankings

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



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



Jobs And Salary?

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



\$115,531



\$94,139

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

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/> >

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

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

The screenshot shows a browser window with a DataCamp course page. On the left, a dark overlay displays a message: "Exercise Completed" with a green checkmark and a "Stop" button. Below it, text says "Nice job! Move onto the next video to start learning more about the RStudio IDE!" and "PRESS ENTER TO Continue" with a "Continue" button. At the bottom, it says "Become a power user!" with "Submit Answer" and "Ctrl + Shift + Enter" buttons. The main content is an RStudio IDE interface showing the R console with text about R version 3.3.1 and its license. The environment pane shows "Global Environment" and "Environment History".

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

Homework assignments will be via DataCamp

The screenshot shows a DataCamp exercise page titled "PCA analysis". The text explains the goal: "To continue with the quality assessment of our samples, in the first part of this exercise, we will perform PCA to look how our samples cluster and whether our condition of interest corresponds with the principal components explaining the most variation in the data." It then provides instructions: "To assess the similarity of the smoc2 samples using PCA, we need to transform the normalized counts then perform the PCA analysis. Assume all libraries have been loaded, the DESeq2 object created, and the size factors have been stored in the DESeq2 object, dds_smc2." Below the text is a code editor with R code:

```
1 # Transform the normalized counts
2 vsd_smc2 <- vst(dds_smc2, blind = TRUE)
3
4 # Plot the PCA of PC1 and PC2
5 ...(..., intgroup=...)
```

 There are "Run Code" and "Submit Answer" buttons. Below the code, there are instructions: "Run the code to transform the normalized counts." and "Perform PCA by plotting PC1 vs PC2 using the DESeq2 plotPCA() function on the DESeq2 transformed counts object, vsd_smc2, and specify the intgroup argument as the factor to color the plot." There is also a "Take Hint (-15 XP)" button.

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

The screenshot shows a DataCamp leaderboard page for course "BGGN213_F19". The page has a search bar and tabs for "30 DAYS", "90 DAYS", and "PAST YEAR". A table lists students with their email, name, courses completed, chapters completed, and XP points. A green circle highlights a plus sign icon in the left sidebar.

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

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

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>)

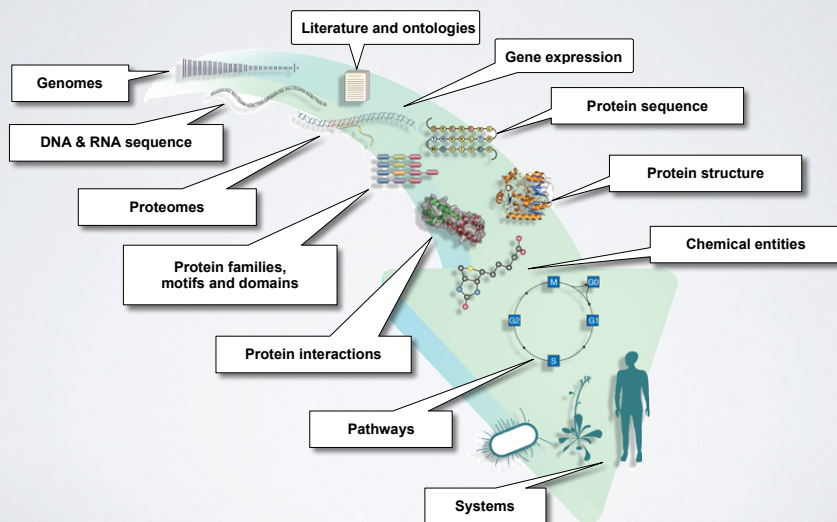
Side-Note:

There are many useful definitions...

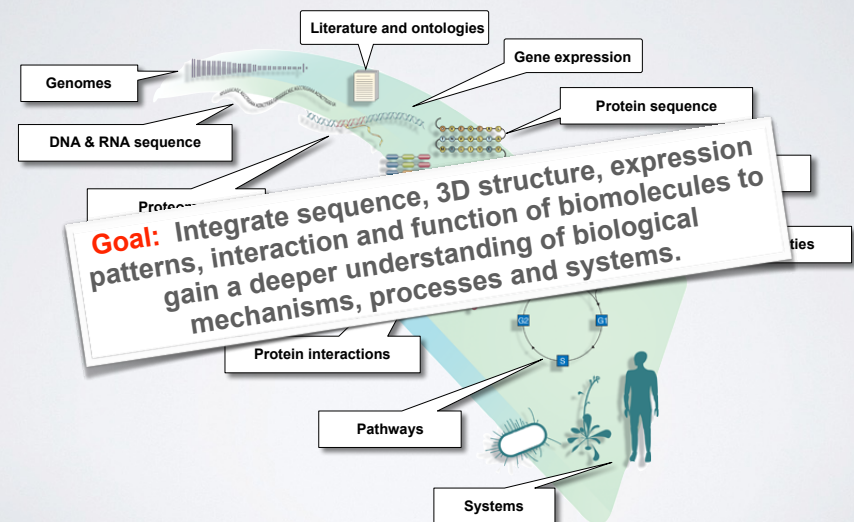
- "Computer based **management** and **analysis** of biological and biomedical data with useful applications in many disciplines, particularly **genomics**, **proteomics**, **metabolomics**, and related fields."
(BGGN-213)
- "Bioinformatics is conceptualizing biology in terms of **macromolecules** and 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>)

Key Point: Bioinformatics is Computer Aided Biology

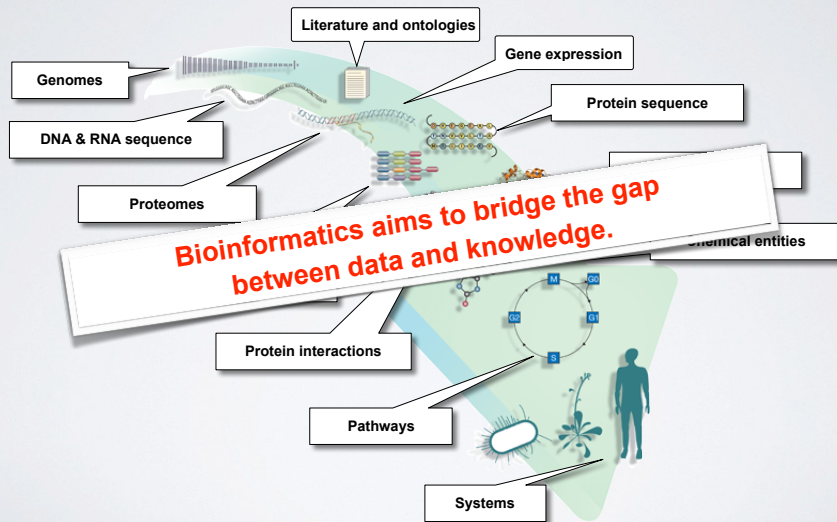
Major types of Bioinformatics Data



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

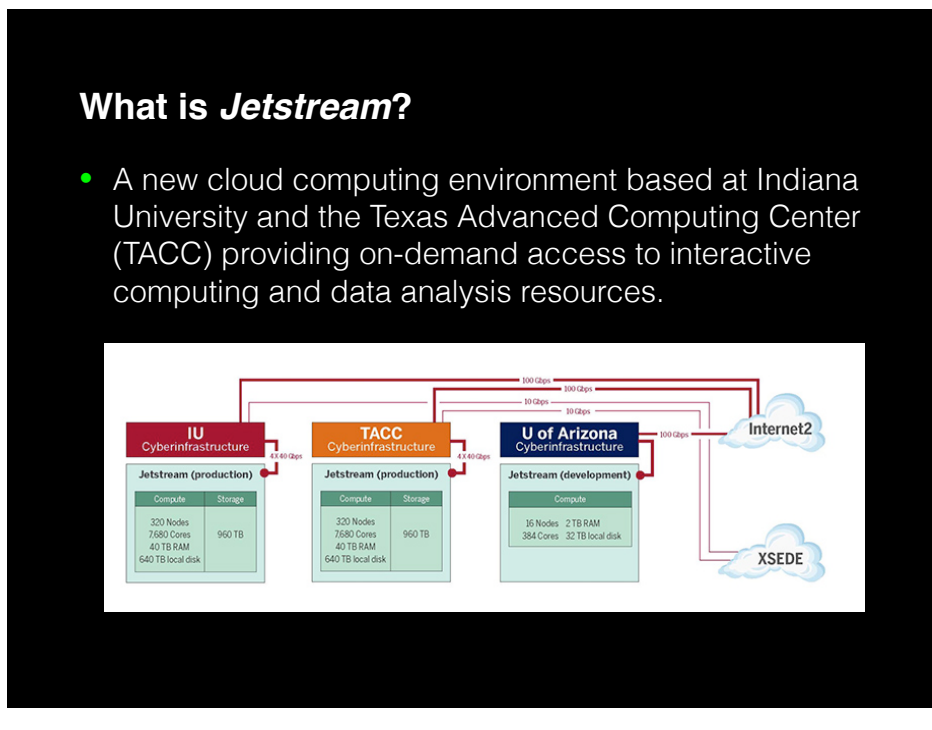
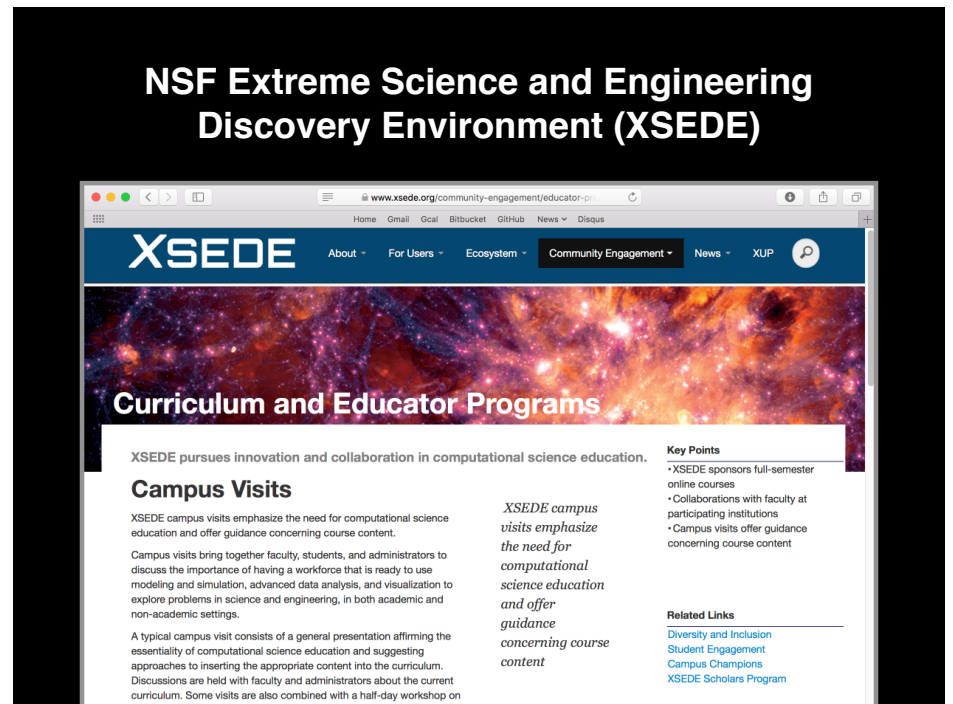
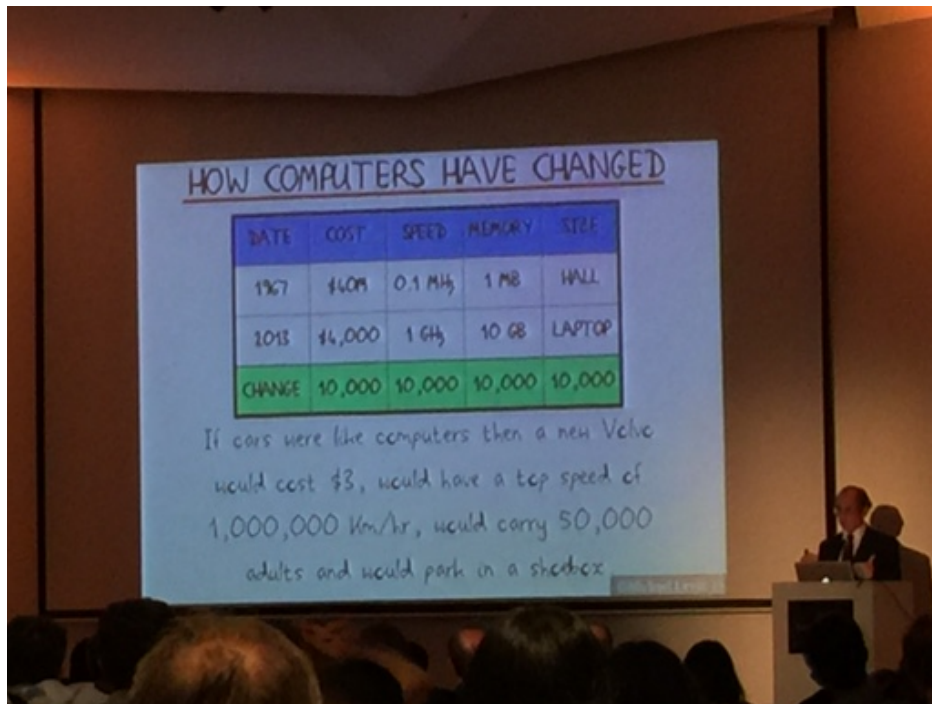
How do we *actually* do Bioinformatics?

Pre-packaged tools and databases

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

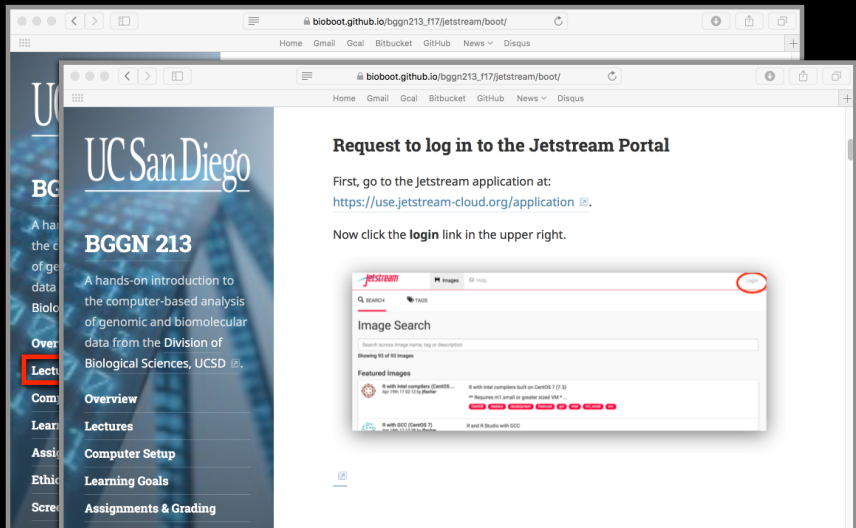
Advanced tool application & development

- 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...



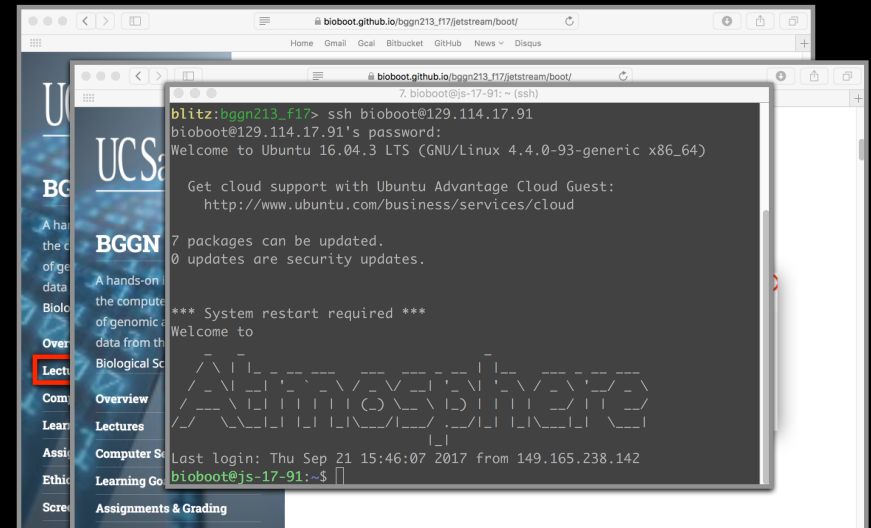
Jetstream tutorials

Developed *user friendly* labs for Jetstream basics



Jetstream tutorials

Developed *user friendly* labs for Jetstream basics



Skepticism & Bioinformatics

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

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

Skepticism & Bioinformatics

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

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

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

Common problems with Bioinformatics

Confusing multitude of tools available

- ▶ Each with many options and settable parameters

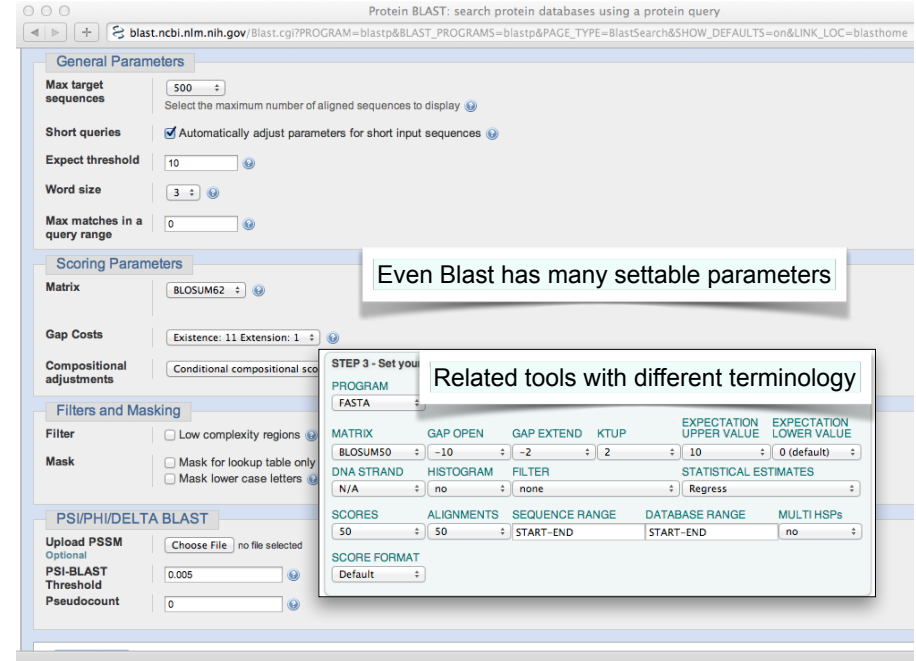
Most tools and databases are written by and for nerds

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

Most are developed independently

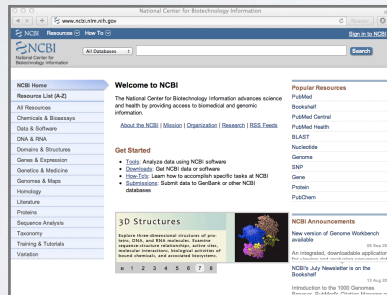
Notable exceptions are found at the:

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

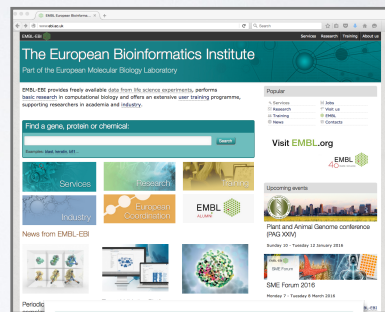


Key Online Bioinformatics Resources: NCBI & EBI

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



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



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

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>

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

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

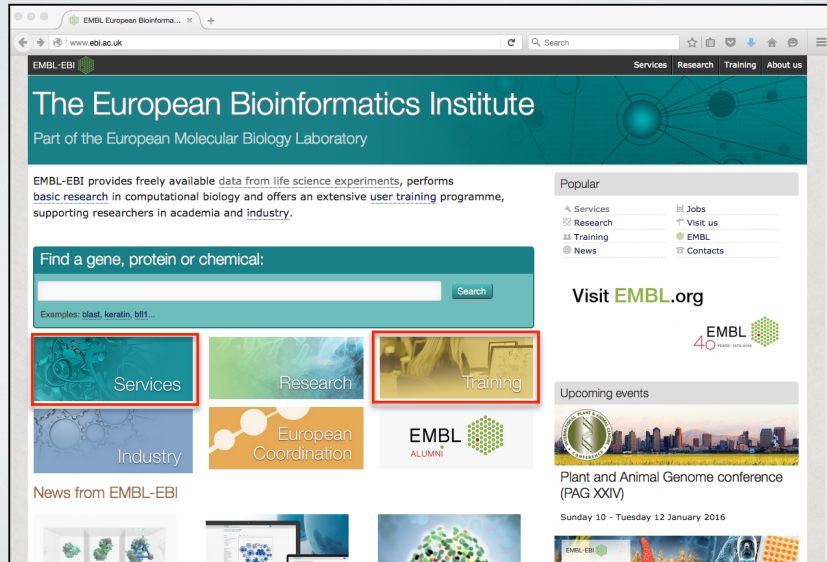
Key Online Bioinformatics Resources: NCBI & EBI

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

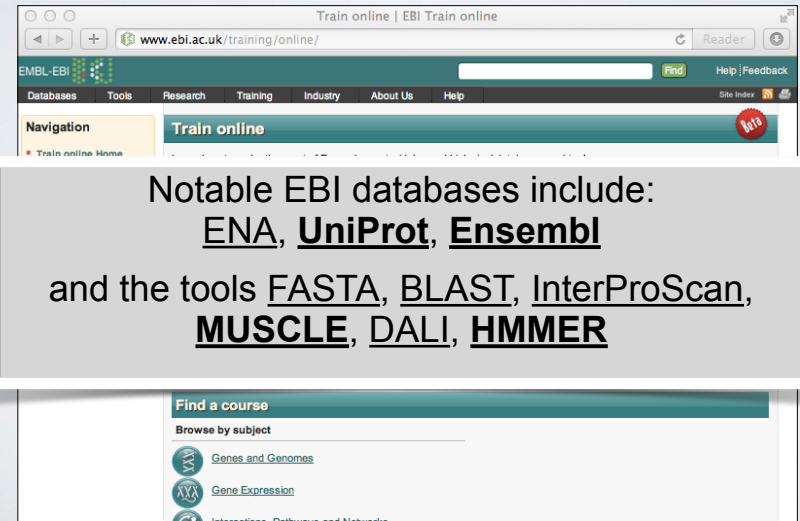
<http://www.ncbi.nlm.nih.gov>

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

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



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

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, Biomag, 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, DPlnteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlilesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIBD, HICD, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, MycDB, NDB, NRSdb, 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, Biomag, 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, DPlnteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlilesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIBD, HICD, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, MycDB, NDB, NRSdb, 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 an 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 leaning 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.

Hands-on section

<http://thegrantlab.org/bgg213/>

Your Turn!

UC San Diego

BGGN 213

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

Overview

Lectures

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

Goals:

- Understand course scope, expectations, logistics and ethics code.
- Understand the increasing necessity for computation in modern life sciences research.
- Get introduced to how bioinformatics is practiced.
- Complete the pre-course questionnaire.
- Setup your laptop computer for this course.
- The goals of the hands-on session is to introduce a range of core bioinformatics databases and associated online services whilst actively investigating the molecular basis of several common human disease.

Material:

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

Homework:

- Questions
- Readings:
 - PDF1: What is bioinformatics? An introduction and overview
 - PDF2: Advancements and Challenges in Computational Biology

BGGN-213: FOUNDATIONS OF BIOINFORMATICS (Lecture 1)

Bioinformatics Databases and Key Online Resources
https://bioboot.github.io/bgg213_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
ATGGTGCATCTGACTCCTGTGGGAAAGTCTGCCCTTACTGCCCTGTGGGCAAGGTGAACGTGGATGAAG
TTGGTGGTGGAGCCCTGGGAGGCTGCTGGTGTCTACCTTGGACCCAGAGGTCTTTGAGTCTTTGG
GGTCTGTCCACTCTCTGACGTTATGGGCAACCTAAAGTGAAGCTCTAGGCGAAGAAAGTCTGGT
GCCTTAGTGATGGCTGGCTACCTGGACAACCTCAAGGGCACCTTTGCACACTGAGTGAAGTGCAC
GTGACAAGCTGCACCTGGTCTGGAAGTCTAGGCTCTGGGCAAGCTGCTGGTCTGTGTGGGCCCA
TCACCTTGGCAAAGATTACCCCAACAGTCAAGGCTGCTATCAGAAAGTGGTGGCTGTGGCTAAT
GCCCTGGCCACAAAGTACTAAAGCTGGCTTTCTTGGCTCAATTT
```

The only information you are given is the above sequence so you must begin your investigation with a sequence search - for this example we will use NCBI's **BLAST** service at: <http://blast.ncbi.nlm.nih.gov/>

Note that there are several different "basic BLAST" programs available at NCBI (including nucleotide BLAST, protein BLAST, and BLASTx).

YOUR TURN!

- There are five major hands-on sections including:

1. BLAST, GenBank and OMIM @ **NCBI** [~35 mins]
2. GENE database @ **NCBI** [~15 mins]
— BREAK —
3. UniProt & Muscle @ **EBI** [~25 mins]
4. PFAM, PDB & NGL [~30 mins]
— BREAK —
5. Extension exercises [~30 mins]

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

YOUR TURN!

- There are five major hands-on sections including:

- | | |
|--|-------------------------|
| 1. BLAST, GenBank and OMIM @ NCBI | End times:
[2:35 pm] |
| 2. GENE database @ NCBI | [2:55 pm] |
| — BREAK — | — 3:10 pm — |
| 3. UniProt & Muscle @ EBI | [3:30 pm] |
| 4. PFAM, PDB & NGL | [4:00 pm] |
| — BREAK — | — 4:10 pm — |
| 5. Extension exercises | [4:40 pm] |

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

SUMMARY

- Bioinformatics is computer aided biology.
- Bioinformatics deals with the collection, archiving, organization, and interpretation of a wide range of biological data.
- 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/bgggn213/>

- ✓ Complete the initial course questionnaire:
- ✓ Check out the "background reading" material online:
- ✓ Complete the lecture 1 homework questions:



THANK YOU