

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/bimm143/>

The screenshot shows the homepage of the BIMM 143 course website. At the top left is the UC San Diego logo. The main title is "Bioinformatics (BIMM 143, Fall 2018)" with a magnifying glass icon. Below the title are sections for "Course Director" (Prof. Barry J. Grant), "Instructional Assistant" (Chao Shi), and "Course Syllabus" (Fall 2018 PDF). A sidebar on the left lists course sections: Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. Social media icons for Twitter, GitHub, and LinkedIn are at the bottom.

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

The screenshot shows the "Learning Goals" section of the course website. It features the same header and sidebar as the homepage. The "Learning Goals" section contains a list of objectives:

- 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 R environment to analyze bioinformatics data at scale.
- Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.
- Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genomics, Transcriptomics and Structural bioinformatics.

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

At the end of this course students will:

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

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

Specific Learning Goals....

What I want you to know by course end!

The screenshot shows the 'Specific Learning Goals' page for BIMM 143. The page title is 'Specific Learning Goals'. Below it, a text block states: 'Teaching toward the specific learning goals below is expected to occupy 60%-70% of class time. The remaining course content is at the discretion of the instructor with student body input. This includes student selected topics for peer presentation, as well as one student selected guest lecture from an industry based genomic scientist.' A table follows, titled 'Lecture(s):' with rows for each goal. The first row is: '1 Appreciate and describe in general terms the role of computation in hypothesis-driven discovery processes within the life sciences.' with 'Lecture(s): 1, 2, 20'. The sidebar on the left lists course sections: Overview, Lectures, Computer Setup, Learning Goals (highlighted with a red box), Assignments & Grading, and Ethics Code.

Course Structure

Derived from specific learning goals

The screenshot shows the 'Course Structure' page for BIMM 143. The page title is 'Course Structure'. Below it, a text block states: 'All Lectures are Tu/Th 9:00-12:00 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.' A table follows, titled 'Topics for Spring 2018' with rows for each lecture. The first row is: '1 Tu, 04/03 Welcome to Bioinformatics' with a detailed description of the course introduction, learning goals, and expectations. The sidebar on the left lists course sections: Overview, Lectures (highlighted with a red box), Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code.

Course Structure

Derived from specific learning goals

The screenshot shows the 'Course Structure' page for BIMM 143. The page title is 'Course Structure'. Below it, a text block states: 'All Lectures are Tu/Th 9:00-12:00 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.' A table follows, titled 'Topics for Spring 2018' with rows for each lecture. The first row is: '1 Tu, 04/03 Welcome to Bioinformatics' with a detailed description of the course introduction, learning goals, and expectations. The sidebar on the left lists course sections: Overview, Lectures (highlighted with a red box), Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code.

Class Details

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

The screenshot shows the 'Class Details' page for BIMM 143. The page title is 'Class Details'. Below it, a section titled '1: Welcome to Foundations of Bioinformatics' is shown. It includes a 'Topics:' section with a detailed description of the course introduction, learning goals, and expectations. It also includes a 'Goals:' section with a bulleted list of items: Understand course scope, expectations, logistics and ethics code, Increasing necessity for computation in modern life sciences research, Get introduced to how bioinformatics is practiced, Complete the pre-course questionnaire ↗, and Setup your laptop computer for this course. A 'Material:' section is also present with a bulleted list: Pre class screen casts (also see below): SC1: Welcome to BIMM-143 ↗, SC2: What is Bioinformatics? ↗ and SC3: How do we do Bioinformatics? ↗, Lecture Slides: Large PDF, Small PDF, and Handout: Class Syllabus ↗. The sidebar on the left lists course sections: Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code.

Homework

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

Homework:

- Questions [\(2\)](#)
- Readings:
 - PDF1: What is bioinformatics? An introduction and overview [\(2\)](#),
 - PDF2: Advancements and Challenges in Computational Biology [\(2\)](#),
 - Other: For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights [\(2\)](#) New York Times, 2014.

Screen Casts:

Welcome to "Foundations of Bioinformatics" (BGGN-213)

Computer Setup

Blank Laptop & Software Setup Instructions

1 Welcome to BIMM-143: Course introduction and logistics.

Homework

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

Homework:

- Questions [\(2\)](#)
- Readings:
 - PDF1: What is bioinformatics? An introduction and overview [\(2\)](#),
 - PDF2: Advancements and Challenges in Computational Biology [\(2\)](#),
 - Other: For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights [\(2\)](#) New York Times, 2014.

Screen Casts:

Welcome to "Foundations of Bioinformatics" (BGGN-213)

Computer Setup

Blank Laptop & Software Setup Instructions

1 Welcome to BIMM-143: Course introduction and logistics.

Homework

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

BIMM143 Lecture 1 Homework (W19)

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 *

Your email _____

UCSD PID number (exam number)

Your answer _____

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

1 point

Homework

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

BIMM143 Lecture 1 Homework (W19)

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

Homework is due before the next weeks class!

Email address *

Your email _____

UCSD PID number (exam number)

Your answer _____

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

1 point

Homework

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

The screenshot shows the DataCamp RStudio interface. The top navigation bar includes 'Exercise', 'Course Outline', and 'RDocumentation'. The main area displays an R script named 'script.R' with the following code:

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

Below the script, the R Console shows the following output:

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

The 'Instructions 1/2 50 XP' section contains two bullet points:- Run the code to transform the normalized counts.
- Perform PCA by plotting PC1 vs PC2 using the DESeq2 `plotPCA()` function on the DESeq2 transformed counts object, `vsd_smoc2`, and specify the `intgroup` argument as the factor to color the plot.

A 'Take Hint (-15 XP)' button is present.

Homework (35% of course grade)

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

The screenshot is identical to the one above, showing the DataCamp RStudio interface with the PCA analysis exercise. A large red diagonal annotation across the top right corner reads "Homework is due before the next weeks class!"

Projects

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

The screenshot shows the GitHub repository page for '9: Unsupervised Learning Mini-Project'. The page includes the following sections:

- 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: Large PDF, Small PDF
 - Lab: Hands-on section worksheet for PCA
 - Data file: WisconsinCancer.csv, new_samples.csv
 - Bio3D PCA App: <http://bio3d.ucsd.edu/pca-app/>
 - Feedback: Muddy point assessment
 - Bonus: Kevin's StackExchange Link on PCA

Projects

(20% of course grade)
Week long mini-projects (x2),
and 1 five week **main project**

The screenshot shows the GitHub repository page for '10: (Project) Find a Gene Assignment Part 1'. The page includes the following sections:

- Topics:** 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.
- Goals:** 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.
- Responses:** Your responses to questions Q1-Q4 are due at the beginning of class Thursday Nov 15th (11/15/18).
- Assignment:** The complete assignment, including responses to all questions, is due at the beginning of class Thursday Dec 4th (12/04/18).
- Late Responses:** Late responses will not be accepted under any circumstances.
- Bonus:** Hands-on with Git
- Description:** Today's lecture and hands-on sessions introduce Git, currently the most popular version control system. We will learn how to perform common operations with Git and RStudio. We will also cover the popular social code-hosting platforms GitHub and BitBucket.

Final Exam

Open-book, open-notes 150-minute test
(45% of course grade)

The screenshot shows the UC San Diego BIMM 143 website. The main navigation menu includes links for Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. Below the menu, there are social media icons for Twitter, GitHub, LinkedIn, and RSS. The main content area is titled "20: Final Exam". It contains instructions for the exam, a list of dos and don'ts, and a "Good luck!" message.

Bonus:

Bioinformatics & Genomics in industry

The screenshot shows the UC San Diego BIMM 143 website. The main navigation menu includes links for Overview, Lectures, Computer Setup, Learning Goals, Assignments & Grading, and Ethics Code. Below the menu, there are social media icons for Twitter, GitHub, LinkedIn, and RSS. The main content area is titled "21: Bonus: Bioinformatics & Genomics in industry". It contains a brief description of the event and a note from Barry J. Grant.

Bonus:

Online portfolio of **your** bioinformatics work!

The screenshot shows Jason Bennett's GitHub repository for Bioinformatics Class BIMM-143. The repository title is "Introduction to Bioinformatics Class S18". It features a DNA helix icon with a magnifying glass and the number "101 110". The repository description states it is a repository to store and display work completed during the Spring 2018 quarter in BIMM-143 at UCSD. It includes a "View the Project on Github" link and a note that the project is maintained by JasonP Bennett. The repository index lists various class topics from "Basic Data Exploration and Visualization in R" to "Transposons: A Sample Workflow".

Bonus:

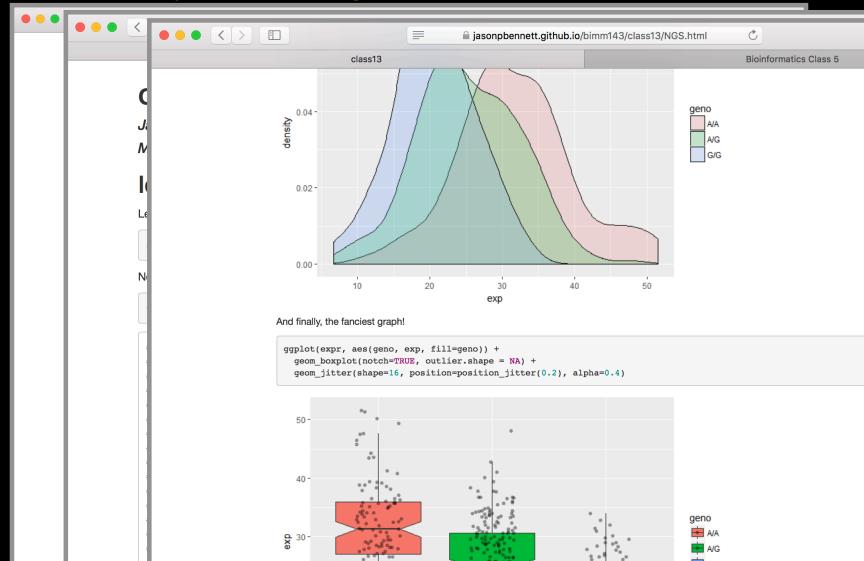
Online portfolio of **your** bioinformatics work!

The screenshot shows Jason Bennett's GitHub repository for class13. The repository title is "Identifying SNP's in a Population". It includes a note about analyzing SNP's from the Mexican-American population in Los Angeles. It shows R code for reading a CSV file and a table of genotype data. The data table includes columns for Population, Sex, and Genotype, with rows for various individuals (NA19648 through NA19664).

Individual	Population	Sex	Genotype
NA19648	ALL	AMR	A A G G
NA19649	AMR	Male	A A G G
NA19651	AMR	Female	A A G G
NA19652	AMR	Male	A A G G
NA19654	AMR	Female	A A G G
NA19655	AMR	Male	A A G G
NA19657	AMR	Female	A A G G
NA19658	AMR	Male	A A G G
NA19661	AMR	Female	A A G G
NA19663	AMR	Male	A A G G
NA19664	AMR	Female	A A G G

Bonus:

Online portfolio of **your** bioinformatics work!



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 biology majors 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 biology majors 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!**

BIMM-143 Learning Goals....

Data science R based learning goals

UC San Diego

BIMM 143

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 (highlighted with a red box)

Assignments & Grading

Ethics Code

Learning Goal	Skills
1 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
2 Use R to read and parse comma-separated (.csv) formatted files ready for subsequent analysis.	8, 9, 10, 11, 13, 15, 16
3 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
4 View and interpret the structural models in the PDB.	10, 11
5 Explain the outputs from structure prediction algorithms and small molecule docking approaches.	11
6 Appreciate and describe in general terms the rapid advances in sequencing technologies and the new areas of investigation that these advances have made accessible.	13, 14, 15
7 Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.	13
8 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
9 Given an RNA-Seq data file, find the set of significantly differentially expressed genes.	

BIMM-143 Learning Goals....

Delve deeper into “real-world” bioinformatics

The screenshot shows the GitHub repository for BIMM-143. The left sidebar includes links for Overview, Lectures, Computer Setup, Learning Goals (which is highlighted with a red box), Assignments & Grading, and Ethics Code. The main content area displays a table of learning goals numbered 9 through 17. A green box highlights the first six goals, which are described below:

- 9 Explain the outputs from structure prediction algorithms and small molecule docking approaches.
- 10 Appreciate and describe in general terms the rapid advances in sequencing technologies and the new areas of investigation that these advances have made accessible.
- 11 Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.
- 12 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.
- 13 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.
- 14 Perform a GO analysis to identify the pathways relevant to a set of genes (e.g. identified by transcriptomic study or a proteomic experiment).

Goals 15 through 17 are also listed but are not highlighted.

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 the R environment to analyze bioinformatics data at scale.
- Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genome informatics, Transcriptomics and Structural informatics.

IEEE 2016 Top Programming Languages

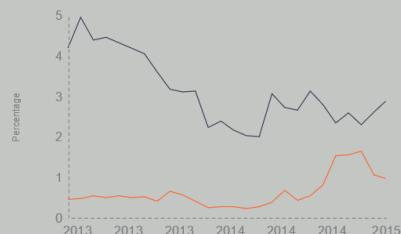


<http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages>

R and Python: The Numbers

Popularity Rankings

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



Python
R

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



Python

\$94,139

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

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

Past Student Opinions...

Q1. Did you enjoy this course in relation to others you have experienced at UCSD?
Hell Yeah!
Yes
it was too lit
Yes!
Yes!
yes
yes!
I do too!
One of the best
The best
yes
Ye
Yes
yes
Yes
Yes
Yes
yes!
yes, one of the most useful classes I've had
no but im just really bad at coding so thots just me <—Don't be discouraged! It takes time. No one starts as a master. :) Chat 0

Past Student Opinions...

Q1. Did you enjoy this course in relation to others you have experienced at UCSD?
Hell Yeah!
Yes
- Yes.
Yes
Yes
Yes
yes, quite.
yes
- I enjoyed this lab course better than my other lab courses
This is the best lab course I've taken at UCSD
Yes
Yes this course was very enjoyable and perhaps more relevant than others
Yes even as a beginner +1
Yes this course was interesting compared to other courses offered at UCSD+1
This is one of the most enjoyable classes offered here! (:+1
Yes
Yes. I very much enjoyed this course.
yes
Yes!
I enjoyed this course much more than many of my other courses at UCSD.
This is one of the best and most useful courses I have taken at UCSD.
Yes
yes, it was a very relaxing course and I love how helpful and passionate the professor and the TA were.

Past Student Opinions...

etherpad.net/p/bimm143_S18

Q1. Did you

Yes
- Yes.
Yes
Yes
Yes, quite.
- I enjoyed this
This is the best
Yes
Yes this course
Yes even as a b
Yes this course
This is one of the
Yes
Yes, I very much
Yes!
I enjoyed this c
This is one of th
Yes, one of th
no but im just
yes, it was a ve

etherpad.net/p/bgn213_S18

Q1. Did you enjoy this course in relation to others you have experience

Yes, very much
Yes, absolutely!
Yes
Yes, I like the focus on applying R to real world biological datasets
Yes
yes
Yes
It was a lot harder than I was expecting
yes
Yes!
yes
yes
Yes!
yes
Yes, I learned lots of things that are very useful in research but hard to learn ourselves
Yes this class was awesome!
Yes, this course was amazingly put together in a logical way and was extremely thorough.



Instructor	Course	Term	Rcmnd Class	Rcmnd Instr	Study Hrs/wk	Avg Grade Expected	Avg Grade Received
Grant, Barry J	BIMM 143 - Bioinformatics Laboratory (A)	FA18	100.0 %	100.0 %	4.50	B+ (3.53)	N/A
Grant, Barry J	BIMM 143 - Bioinformatics Laboratory (A)	SP18	94.7 %	94.7 %	5.66	B+ (3.63)	B+ (3.35)
Grant, Barry J	BIMM 143 - Bioinformatics Laboratory (A)	WI18	100.0 %	100.0 %	5.64	B+ (3.64)	N/A
Grant, Barry J	BIMM 194 - Adv Topics- Molecular Biol (C)	WI18	92.9 %	100.0 %	1.30	A (4.00)	A (4.00)

Average for BIMM143: 98.2 % 98.2 % 5.27

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.

Q. What is Bioinformatics?

Q. What is Bioinformatics?

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

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

Q. What is Bioinformatics?

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

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

... **Bioinformatics is computer aided biology!**

Q. What is Bioinformatics?

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

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

... **Bioinformatics is computer aided biology!**

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

MORE DEFINITIONS

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

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

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

MORE DEFINITIONS

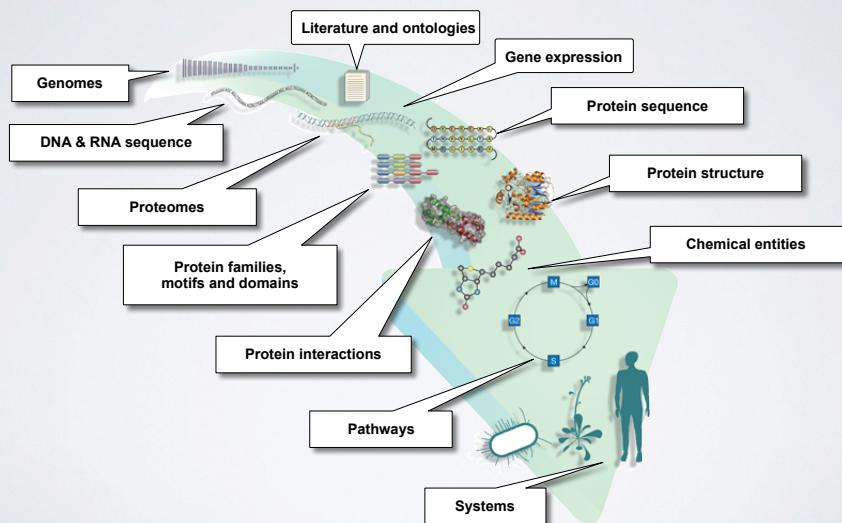
- “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “**informatics**” techniques (derived from disciplines such as applied mathematics, computer science, and statistics) to **understand** and **analyze** the information associated with these molecules, on a large-scale.

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

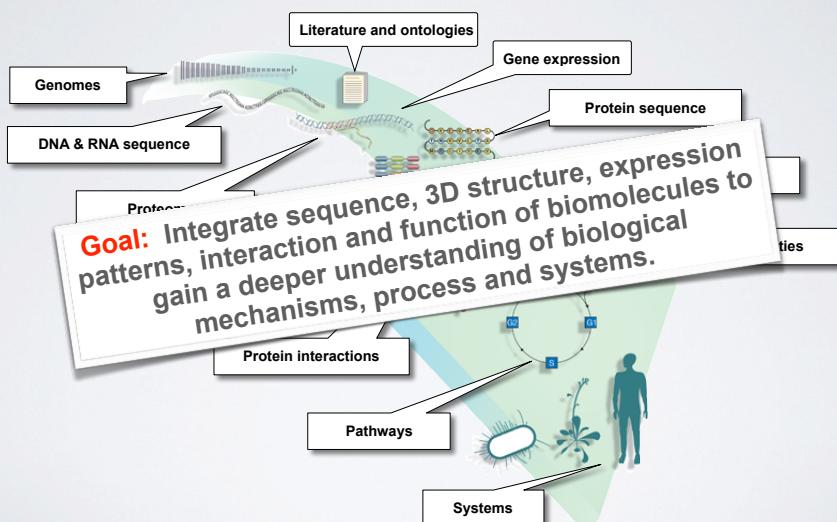
- “Bioinformatics is research, development, or application of **computational approaches** for expanding the use of **biological, medical, behavioral or health data**, including those to acquire, store, organize and analyze such data.”

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

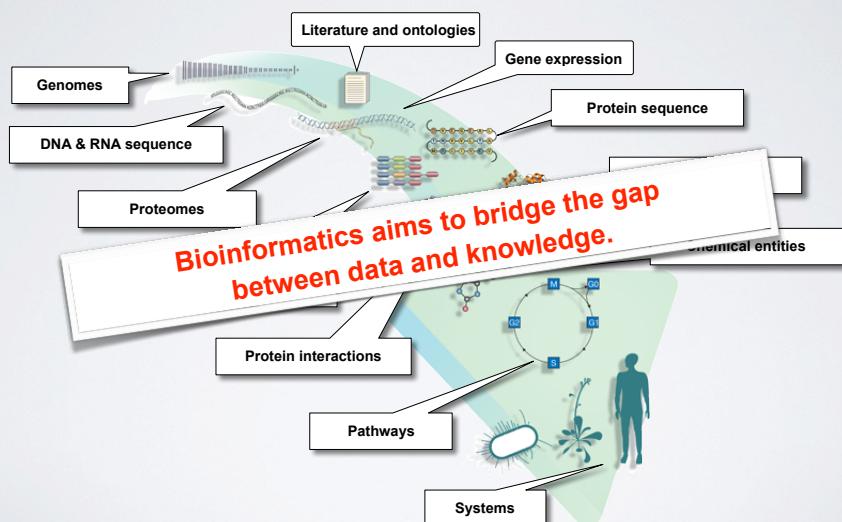
Major types of Bioinformatics Data



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

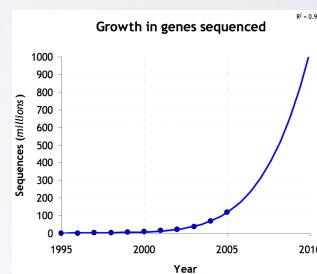
- DNA sequence determines protein sequence.
- Protein sequence determines protein structure.
- Protein structure determines protein function.
- Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function *in space and time*.

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

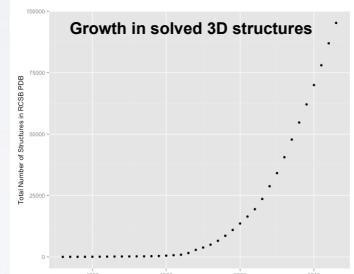


E.G. data from sequencing, structural genomics, proteomics, new high throughput assays, etc...

Why do we need Bioinformatics?

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

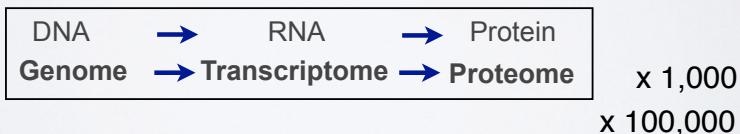
- Bioinformatics provides methods for the efficient:
 - storage
 - annotation
 - search and retrieval
 - data integration
 - data mining and analysis



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?

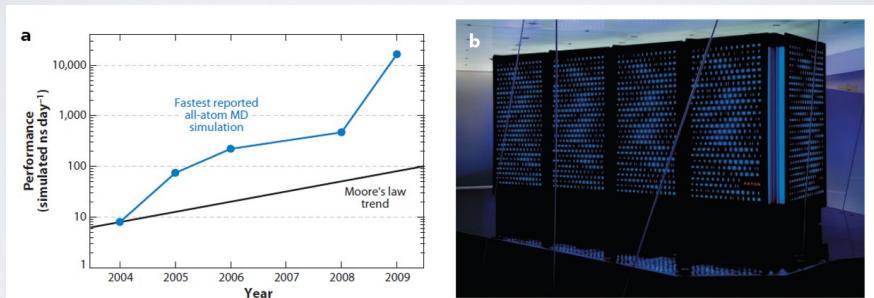
Pre-packaged tools and databases

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

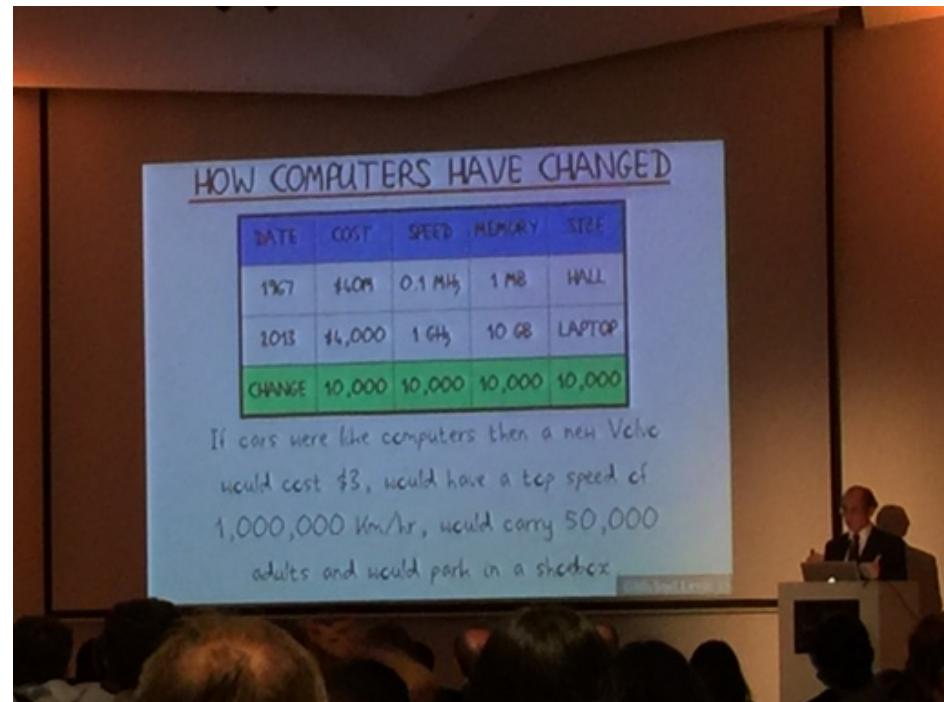
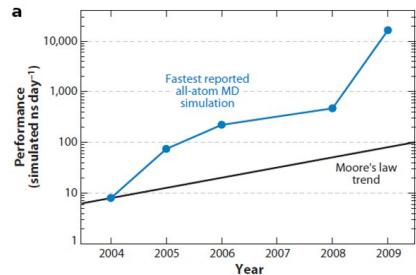
Advanced tool application & development

- 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



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*” 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 sidebar with links like 'NCBI Home', 'Resource List (A-Z)', 'All Resources', 'Chemical, Pharmacogenomics', 'Data & Software', 'DNA & RNA', 'Proteins', 'Proteins & Domains', 'Genetics & Medicine', 'Genomes & Maps', 'Hivology', 'Luminescence', 'Proteins', 'Protein Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Visualization'. The main content area displays 'Welcome to NCBI' and 'Popular Resources' including PubMed, Bookshelf, Nucleotide Central, PubMed Health, and more. A 'Get Started' section provides links to 'Analyze data using NCBI software', 'Download NCBI data or software', 'DataSets', 'Learn how to accomplish specific tasks at NCBI', and 'Search data in databases or other NCBI databases'. A '3D Structures' section shows a molecular model. A 'NCBI Announcements' box is also present.

The screenshot shows the European Bioinformatics Institute (EBI) homepage. It features a sidebar with links like 'Home', 'About EBI', 'Services', 'Research', 'Training', 'Contact', 'Log in', 'Help', 'Logout', and 'Feedback'. The main content area displays 'Welcome to the European Bioinformatics Institute' and 'Popular Resources' including PubMed, Bookshelf, Nucleotide Central, PubMed Health, and more. A 'Find a gene, protein or chemical' search bar is prominently displayed. Below it are sections for 'Services', 'Research', 'Training', 'Industry', 'European Bioinformatics', 'EMBL', 'News from EMBL-EBI', 'Upcoming events', 'Plant and Animal Genome conference (IPAG XXV)', 'IPAG XXV', 'SME Forum 2016', 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 bioinformatics.

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 bioinformatics.

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 bioinformatics.

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 bioinformatics.

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 2016)

SIB Forum 2016

Periodic

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

European Bioinformatics Institute (EBI)

- Created in 1997 as a part of the European Molecular Biology Laboratory (EMBL)
- EBI's mission includes:
 - providing freely available **data and bioinformatics services**
 - and providing advanced **bioinformatics training**
- We will briefly cover several EBI databases and tools that have advantages over those offered at NCBI



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

The screenshot shows the main website for the European Bioinformatics Institute. At the top, there is a navigation bar with links for Services, Research, Training, and About us. Below the header, a banner reads "The European Bioinformatics Institute Part of the European Molecular Biology Laboratory". A search bar is present with the placeholder "Find a gene, protein or chemical:". Below the search bar, there is a section titled "Popular" with links to Services, Research, Training, European Coordination, and EMBL ALUMNI. A red box highlights the "Services" link. To the right, there is a sidebar with links for Jobs, Visit us, EMBL News, and Contacts. Further down, there is a "Visit EMBL.org" section with a link to "EMBL 40 years". On the left, there is a "Upcoming events" section for the "Plant and Animal Genome conference (PAG XXIV)" from Sunday 10 - Tuesday 12 January 2016.

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

The screenshot shows the "Services" page of the EBI website. At the top, there is a navigation bar with links for Services, Research, Training, and About us. Below the header, there is a section titled "Bioinformatics services" with a brief description. A grid of icons represents various services: DNA & RNA, Gene expression, Proteins (which is highlighted with a red box), Structures, Systems, Chemical biology, Ontologies, Literature, and Cross domain. To the right, there is a "Popular" section with links to Ensembl, UniProt, PDBe, ArrayExpress, and ChEMBL. Below the "Popular" section, there is a "Training" section featuring an image of a person working on a computer.

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

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

The screenshot shows the "Proteins" page of the EBI website. At the top, there is a navigation bar with links for Services, Research, Training, and About us. Below the header, there is a section titled "Proteins" with a brief description. A "Popular services" section lists UniProt, InterPro, PRIDE, Pfam, Clustal Omega, HHMMER, and InterProScan 5. Each service has a small icon and a brief description. To the right, there is a "Quick links" section with links for Popular services in this category, All services in this category, and Project websites in this category.

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

The screenshot shows the main website of the European Bioinformatics Institute (EBI). At the top, there's a search bar and navigation links for Services, Research, Training, and About us. Below the header, there's a section titled "Popular" with links to Services, Research, Training, and News. A large "Visit EMBL.org" button is prominently displayed. To the right, there's a "Training" section with a red box highlighting the "Training" link. Further down, there's a "Upcoming events" section featuring the "Plant and Animal Genome conference (PAG XXIV)".

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

This screenshot shows a specific online tutorial page from the EBI website. The title is "Using sequence similarity searching tools at EMBL-EBI: webinar". The page includes a video player showing a presentation by Andrew Cowley, a description of the webinar's focus on using tools like BLAST and PSI-Search, and a sidebar with links to other training resources and events.

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

The screenshot shows the "Train online" section of the EBI website. It features a heading "Notable EBI databases include: ENA, UniProt, Ensembl" and a list of tools: FASTA, BLAST, InterProScan, MUSCLE, DALI, HMMER. Below this, there's a "Find a course" section with links to Genes and Genomes and Gene Expression.

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, Biolmage, 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, Pictr_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DATABASE, 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, HAEML, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVdb, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klobo, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, MycDB, NDB, NRSub, 0-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

There are lots of Bioinformatics Databases

For a annotated listing of major bioinformatics databases please see the online handout
< Major Databases.pdf >

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.

Side-note: Databases come in all shapes and sizes



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

Your Turn!

https://bioboot.github.io/bimm143_W19/lectures/#1

1: Welcome to Foundations of Bioinformatics

Topics:
Course introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Student 30-second introductions, Student computer setup.

Goals:

- Understand course scope, expectations, logistics and [ethics code](#).
- Understand the increasing necessity for computation in modern life sciences research.
- Get introduced to how bioinformatics is practiced.
- Complete the [pre-course questionnaire](#).
- Setup your [laptop computer](#) for this course.

Material:

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

BIMM-143: INTRODUCTION TO BIOINFORMATICS (Lecture 1)

Bioinformatics Databases and Key Online Resources

https://bioboot.github.io/bimm143_W18/lectures/#/

Dr. Barry Grant
Jan 2018

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
ATGGTCATCTGACTCTTGAGAACGCTCTCCCGTTACGCCCGTGGGGCGCAAGCTGCACTGTGAA
TTGGTGTTGAGGGCTTBSGCGACGCTGGGCTTPTPCTTAACCTTGAAACCCACGCTGCTTGGAGTCCTTG
GAACTGTTGTTGATGCTCCTGCCTCACCGGCAACCTTGCGCAACCTGGCTCAAGGAAAGTGCTGCTGG
GCTCTGCTGATGCTCCTGCCTCACCGGCAACCTTGCGCAACCTGGCTCAAGGAAAGTGCTGCTGG
GTGACAAGCTGGACACTGCAATCTGGAGAATCTTGCGCAACCTGGCTCTGGCAACCTGCTGCTGCTGG
CTGACTTGTGCAANGAATTCACCCCCACAGTCAGGCTTGGCTTACGAAAGTGCTGGCTGGCTGGCTTAA
GCCCTGGCCACAAATTCATGCTTACGCTTGGCTTACGAAAGTGCTGGCTGGCTGGCTTAA
```

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

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

YOUR TURN!

- There are five major hands-on sections including:

1. BLAST, GenBank and OMIM @ **NCBI**

[~35 mins]

2. GENE database @ **NCBI**

[~15 mins]

— BREAK —

3. UniProt & Muscle @ **EBI**

[~25 mins]

4. PFAM, PDB & NGL

[~30 mins]

— BREAK —

5. Extension exercises

[~30 mins]

► Please do answer the last review question (**Q19**).

► We encourage discussion and exploration!

YOUR TURN!

- There are five major hands-on sections including:

1. BLAST, GenBank and OMIM @ **NCBI**

2. GENE database @ **NCBI**

— BREAK —

3. UniProt & Muscle @ **EBI**

4. PFAM, PDB & NGL

— BREAK —

5. Extension exercises

End times:

[10:35 am]

[10:55 am]

— 11:05 am —

[11:30 am]

[12:00 pm]

— 12:10 am —

[12:40 pm]

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

https://bioboot.github.io/bimm143_W19/lectures/#1

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

