



Office Hours:
[SignUp](#)

Location:
Bonner hall,
#2140

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 Bioinformatics (BIMM 143, Fall 2018) website. The header features the UC San Diego logo and the course title. Below the header, there's a sidebar with links for Course Director (Prof. Barry J. Grant), Instructional Assistant (Chao Shi), and Course Syllabus (Fall 2018 PDF). The main content area includes an 'Overview' section with a brief introduction to bioinformatics and its applications in modern life sciences and medical research. A sidebar on the right lists course sections: Overview, Lectures, Computer Setup, Learning Goals (which is currently selected and highlighted in red), Assignments & Grading, and Ethics Code. Social media icons for Twitter, GitHub, LinkedIn, and RSS feed are at the bottom.

<http://thegrantlab.org/bimm143/>

This screenshot is identical to the one above, showing the homepage of the Bioinformatics (BIMM 143, Fall 2018) website. The difference is that the 'Learning Goals' link in the sidebar has been highlighted with a red box to indicate it has been clicked or is the current page.

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

The screenshot shows the 'Learning Goals' page. At the top, it says 'At the end of this course students will:' followed by a bulleted list of learning objectives. The list includes: '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.', and 'Be familiar with the research objectives of the bioinformatics related sub-disciplines of Genomics, Transcriptomics and Structural bioinformatics.' Below this, a note states: 'In short, students will develop a solid foundational knowledge of bioinformatics and be able to evaluate new biomolecular and genomic information using existing bioinformatic tools and resources.' The 'Learning Goals' link in the sidebar is highlighted with a red box.

At the end of this course students will:

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

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

Specific Learning Goals....

What I want you to know by course end!

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 as 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 between global and local alignment along with their major application areas.	4, 5
5 Calculate the alignment score between two nucleotide or protein sequences using a provided scoring matrix and be able to perform BLAST, PERBLAST, LMMER and needleman-wunsch local alignments.	5, 10

Course Structure

Derived from specific learning goals

Lectures

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.

#	Date	Topics for Spring 2018
1	Tu, 04/03	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	Th, 04/05	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
		Advanced sequence alignment and database searching

Course Structure

Derived from specific learning goals

Lectures

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.

#	Date	Topics for Spring 2018
1	Tu, 04/03	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	Th, 04/05	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
		Advanced sequence alignment and database searching

Class Details

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

The screenshot shows the 'Welcome to Foundations of Bioinformatics' page. It includes the UC San Diego logo, the course title 'BIMM 143', and a brief description of the course goals. The main content area lists 'Topics', 'Goals', and 'Material'. Under 'Goals', there are bullet points about understanding course scope, expectations, logistics, and ethics code. Under 'Material', there are links to pre-class screen casts (SC1, SC2, SC3), lecture slides, and a class syllabus.

Homework

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

The screenshot shows the 'Homework' section of the course website. It includes the UC San Diego logo, the course title 'BIMM 143', and a brief description of the course goals. The main content area lists 'Homework', 'Topics', and 'Readings'. Under 'Homework', there is a link to 'Questions'. Under 'Readings', there are links to PDFs and other resources. Below this, there is a 'Screen Casts' section featuring a video player showing a screencast titled 'Welcome to "Foundations of Bioinformatics" (BGGN-2...)'.

Homework

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

The screenshot shows the 'Homework' section of the course website. It includes the UC San Diego logo, the course title 'BIMM 143', and a brief description of the course goals. The main content area lists 'Homework', 'Topics', and 'Readings'. Under 'Homework', there is a link to 'Questions' which is highlighted with a red box. Under 'Readings', there are links to PDFs and other resources. Below this, there is a 'Screen Casts' section featuring a video player showing a screencast titled 'Welcome to "Foundations of Bioinformatics" (BGGN-2...)'.

Homework

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

The screenshot shows a Google Form titled 'BIMM143 Lecture 1 Homework'. The form asks users to answer the following questions. The first question is 'Which of the following operating systems is most frequently used for bioinformatics tool development?' with options for Windows, iOS, Unix, and Perl. The second question is 'Which of the following databases contains primarily protein sequences?' with options for MySQL, PostgreSQL, Oracle, and MongoDB. A note indicates that each question is worth 1 point.

Homework

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

BIMM143 Lecture 1 Homework
Please answer the following questions
* Required

Your email

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

Windows
 iOS
 Unix
 Perl

Which of the following databases contains primarily protein sequences 1 point

Random

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 programing 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 programing 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)
Assignments & Grading
Ethics Code

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 R to read and parse comma-separated (.csv) formatted files ready for subsequent analysis.	8, 9, 10, 11, 13, 15, 16
7 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
8 View and interpret the structural models in the PDB.	10, 11
9 Explain the outputs from structure prediction algorithms and small molecule docking approaches.	11
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.	13, 14, 15
11 Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.	13
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.	14
Given an RNA-Seq data file, find the set of significantly differentially	

BIMM-143 Learning Goals....

Delve deeper into “real-world” bioinformatics



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)

Assignments & Grading

Ethics Code

In bioinfohub/BIMM143_WtGoals/

Home	Goal	Blackett	Gibk	Notes	Disease	SGN-213	BIMM-143	QCodes
View and interpret the structural models in the PDB.	19, 21							
9 Explain the outputs from structure prediction algorithms and small molecule docking approaches.	11							
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.	13, 14, 15							
11 Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.	13							
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.	14							
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.	15, 16							
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).	16							
15 Use the KEGG pathway database to look up interaction pathways.	17							
16 Use graph theory to represent biological data networks.	17, 18							
17 Understand the challenges in Integrating and Interpreting large heterogeneous high throughput data sets into their functional	19							

Why use R?

Productivity Flexibility

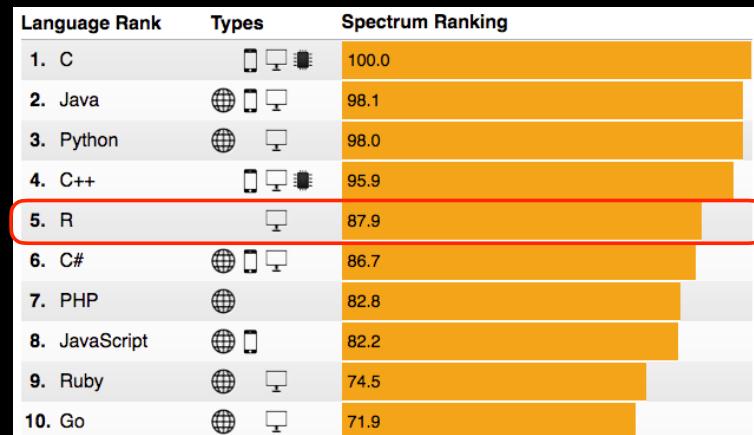
Designed for data analysis

These support a major learning objective

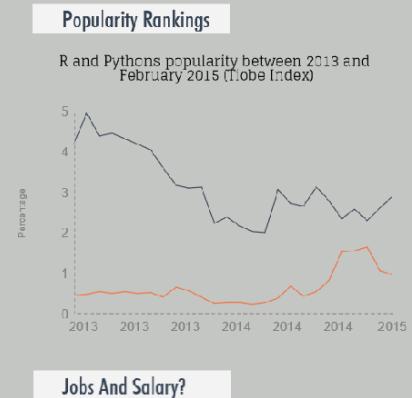
At the end of this course students will:

- 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



R and Python: The Numbers



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.
- Large user and developer community.
 - As of Aug 14th 2018 there are 12,907 add on **R packages** on **CRAN** and 1,560 on **Bioconductor** - more on these later!
- Virtually every statistical technique is either already built into R, or available as a free package.
- Unparalleled exploratory data analysis environment.

Today's Menu

Course Logistics	Website, screencasts, survey, ethics, assessment and grading.
Learning Objectives	What you need to learn to succeed in this course.
Course Structure	Major lecture topics and specific learning goals.
Introduction to Bioinformatics	Introducing the <i>what</i> , <i>why</i> and <i>how</i> of bioinformatics?
Computer Setup	Ensuring your laptop is all set for future sections of this course.

OUTLINE

Overview of bioinformatics

- The *what*, *why* and *how* of bioinformatics?
- Major bioinformatics research areas.
- Skepticism and common problems with bioinformatics.

Online databases and associated tools

- Primary, secondary and composite databases.
 - Nucleotide sequence databases (GenBank & RefSeq).
 - Protein sequence database (UniProt).
 - Composite databases (PFAM & OMIM).

Database usage vignette

- How-to productively navigate major databases.

Q. What is Bioinformatics?

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

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

Q. What is Bioinformatics?

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

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

... Bioinformatics is computer aided biology!

Q. What is Bioinformatics?

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

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

... Bioinformatics is computer aided biology!

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

MORE DEFINITIONS

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

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

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

MORE DEFINITIONS

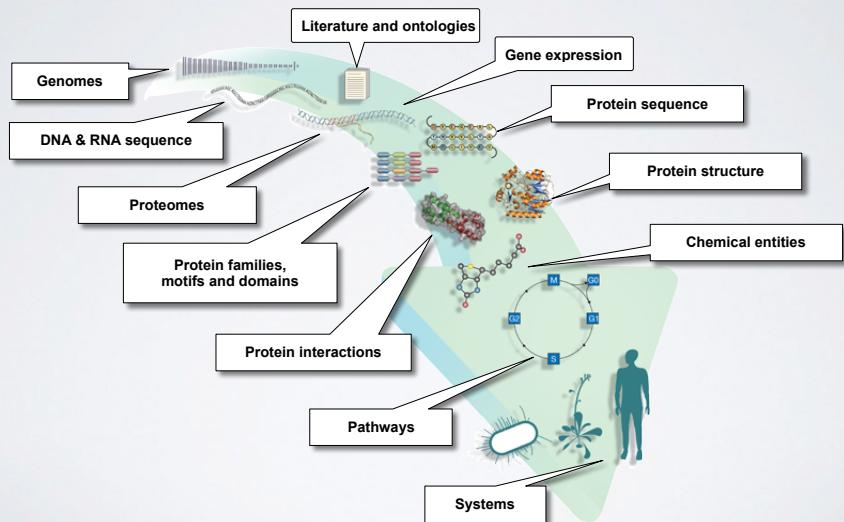
- “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “**informatics**” techniques (derived from disciplines such as applied math, computer science, and statistics) to **understand** and **analyze** the information associated with these molecules, on a large-scale.

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

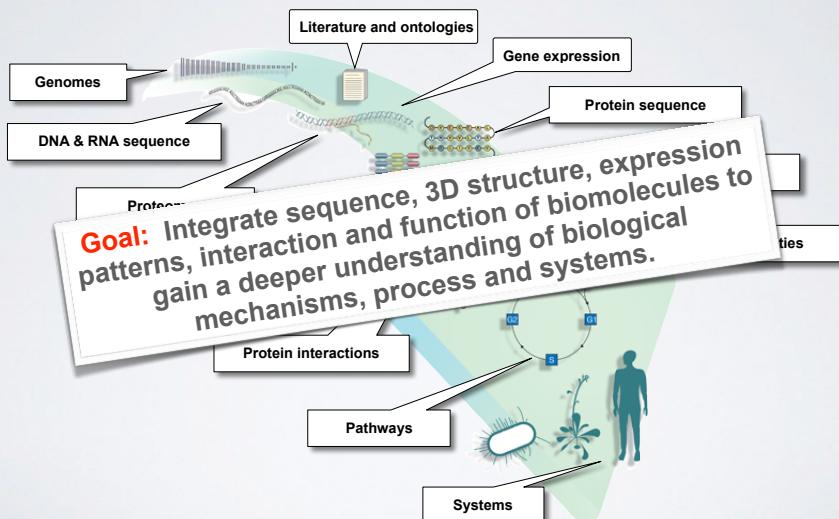
- “Bioinformatics is research, development, or application of **computational approaches** for expanding the use of **biological, medical, behavioral or health data**, including those to acquire, store, organize and analyze such data.”

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

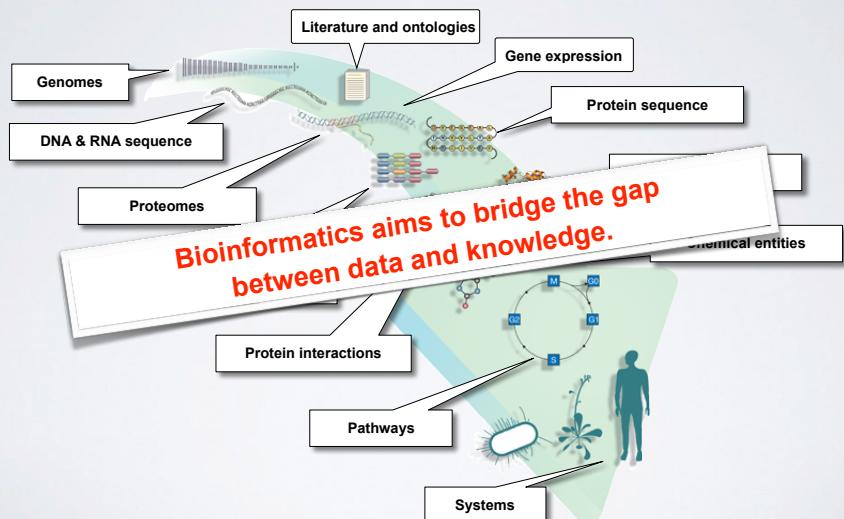
Major types of Bioinformatics Data



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

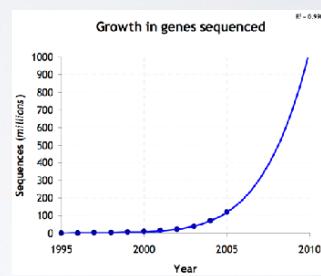
- DNA sequence determines protein sequence.
- Protein sequence determines protein structure.
- Protein structure determines protein function.
- Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function *in space and time*.

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

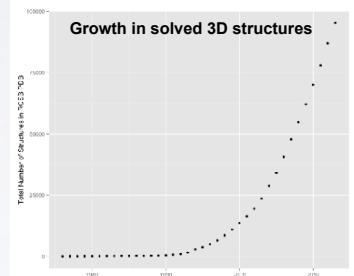


E.G. data from sequencing, structural genomics, proteomics, new high throughput assays, etc...

Why do we need Bioinformatics?

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

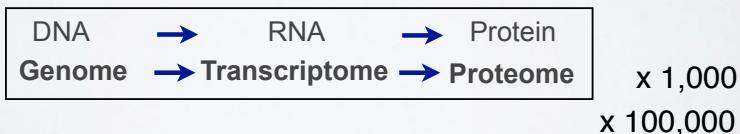
- Bioinformatics provides methods for the efficient:
 - storage
 - annotation
 - search and retrieval
 - data integration
 - data mining and analysis



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?

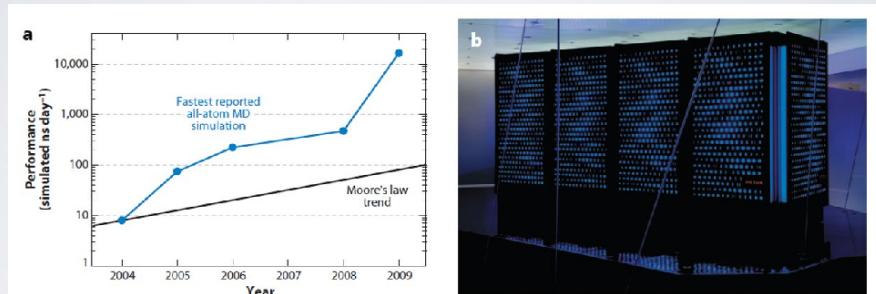
Pre-packaged tools and databases

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

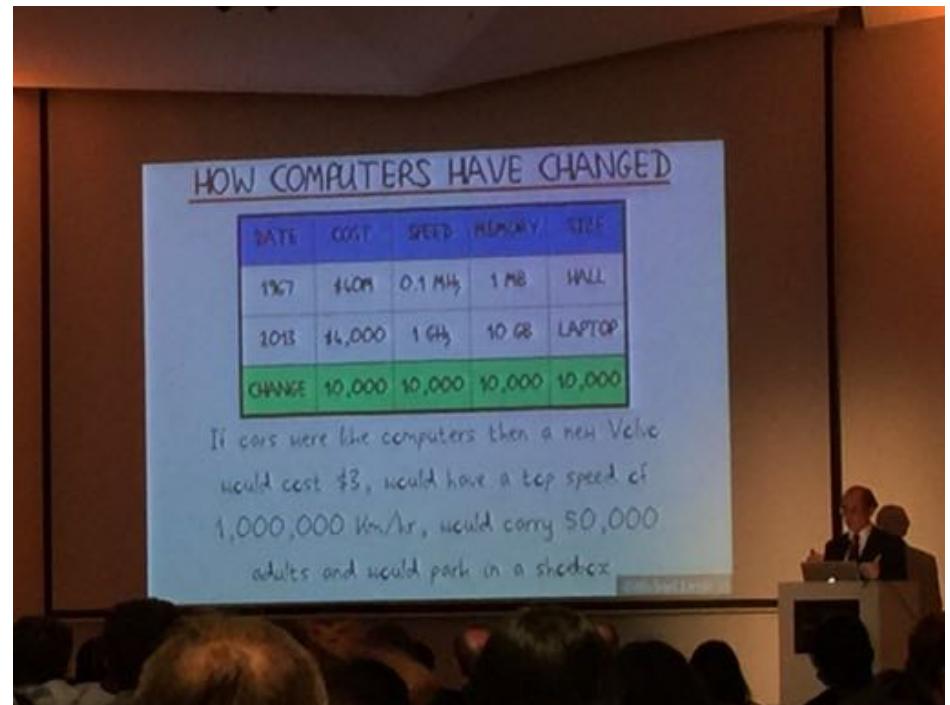
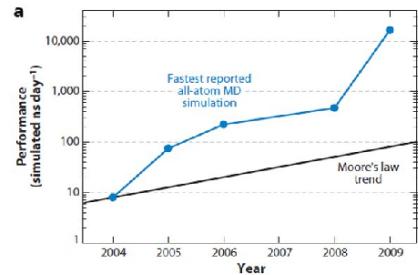
Advanced tool application & development

- 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, PSI-BLAST Threshold: 0.005, Pseudocount: 0). A callout box highlights the 'Even Blast has many settable parameters' note. Another callout box highlights the 'Related tools with different terminology' section, which shows a table for 'STEP 3 - Set your PROGRAM' with columns for MATRIX, GAP OPEN, GAP EXTEND, KTUP, EXPECTATION UPPER VALUE, and EXPECTATION LOWER VALUE, along with other tool-specific parameters like HISTOGRAM, FILTER, and STATISTICAL ESTIMATES.

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 Biology', 'Data & Schema', 'DNA & RNA', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Theory & Tutorials', and 'Visualize'. The main content area includes sections for '3D Structures' (with a 3D molecular model), 'NCBI Announcements' (with a 'New version of Genome PDB'), and 'Get Started' (with a 'Get started with NCBI software' link). Below these are 'Popular Resources' like PubMed, Bookshelf, and Reference Central.

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

The screenshot shows the European Bioinformatics Institute (EBI) homepage. It features a sidebar with links like 'Welcome to EBI', 'About the EBI', 'Get Started', 'Popular Resources' (Pubmed, Bookshelf, Reference Central), and 'Resources'. The main content area includes sections for 'Find a gene, protein or database', 'Vital EMBL-EBI', 'EMBL-EBI News', 'EMBL-EBI Events', and 'EMBL-EBI Resources'. There are also links to 'EMBL-CCB', 'EMBL-EBI Conference', and 'EMBL-EBI Forum'.

<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 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 small molecules, and interaction networks.



NCBI Announcements

New version of Genome Workbench available

An integrated, downloadable applicati

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

The screenshot shows the NCBI homepage with various links and sections. A red arrow points from the 'Popular Resources' section to the 'PubMed' link. Another red arrow points from the same section to the 'BLAST' link. A vertical red curly brace groups the 'SNP', 'Gene', 'Protein', and 'PubChem' links.

National Center for Biotechnology Information

NCBI Resources How To

All Databases

Popular Resources

PubMed

Bookshelf

PubMed Central

PubMed Health

BLAST

Nucleotide

Genome

SNP

Gene

Protein

PubChem

Welcome to NCBI

The National Center for Biotechnology and health by providing access information.

About the NCBI | Mission | Our

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

3D Structures

Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of human pharmaceuticals, and structural bioinformatics.

Reader

Sign in to NCBI

Search

Resources

Central

Health

Announcements

New version of Genome Workbench available

An Integrated, downloadable applicat

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

National Center for Biotechnology Information

NCBI Resources How To Sign in to NCBI

All Databases Search

NCBI Home Resource List (A-Z)

Welcome to NCBI The National Center for Biotechnology Information advances science

Popular Resources PubMed

Notable NCBI databases include: GenBank, RefSeq, PubMed, dbSNP

and the search tools ENTREZ and BLAST

Homology Literature Proteins Sequence Analysis Taxonomy Training & Tutorials Variation

databases

3D Structures

Explore three-dimensional structures of proteins, DNA, and RNA molecules. Examine sequence-structure relationships, active sites, molecular interactions, biological activities of ligands, pharmacological and environmental interactions.

Protein PubChem

NCBI Announcements

New version of Genome Workbench available

An integrated, downloadable applicati

Key Online Bioinformatics Resources: NCBI & EBI

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

The screenshot shows the NCBI homepage with a search bar at the top containing the query "NCBI". Below the search bar, there are several search results:

- NCBI** National Center for Biotechnology Information

On the left sidebar, there are links for "NCBI Home", "Resource List A-Z", "All Resources", "Chemicals & Bioassays", "Data & Tools", "DNA & RNA", "Domains & Structures", "Genes & Expression", "Genomes & Metabolism", "Genomics & Maps", "Homology", "Literature", "Phenotype", "Sequence Analysis", "Taxonomy", "Training & Tutorials", and "Variation".

The main content area features a "Welcome to NCBI" banner with the text: "The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information and tools". Below this is a section titled "About the NCBI" with links to "Mission | Organization | Research | NCBI Tools".

A "Get Started" section provides links to "Tools: Analyze data using NCBI software", "Databases: Search our databases", "Data: Download datasets", "Help: Learn how to access our specific tools at NCBI", and "Submit: Submit data to Genbank or other NCBI databases".

The right sidebar contains a "Popular Resources" section with links to "PubMed", "Bookshelf", "Published Central", "NCBI Bookshelf", "BLAST", "NCBI Resource", "Bioviewer", "BioSearch", "BioSearch2", "Protein", and "PubChem".

At the bottom, there is a "3D Structures" section with a link to "Folding@home: Distributed protein folding research", followed by a "NCBI Announcements" section with a link to "New version of Genomic Workbench", and a "NCBI News" section with a link to "NCBI News: New resources in the Bookshelf".

The screenshot shows the EMBL-EBI website. At the top, there's a search bar with placeholder text "Find a gene, protein or database". Below the search bar, a teal banner displays "News from EMBL-EBI" and "New: Plant and Animal Genomes conference 9-14 July 2016". The main header features the text "The European Bioinformatics Institute" and "Perform European Molecular Biology Laboratory". To the right of the header are links for "Search", "Research", "Helping", and "About". A sidebar on the left lists "Services", "Datasets", "Software", "Training", and "Events". On the right, there's a "People" section with a list of names and their titles, and a "Visit EMBL.org" button. The footer contains a "Publications" link and a "Feedback" link.

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

A screenshot of the EMBL-EBI Services homepage. The page features a dark blue header with the EMBL-EBI logo and navigation links for Services, Research, Training, and About us. Below the header is a banner for "The European Bioinformatics Institute Part of the European Molecular Biology Laboratory". The main content area includes a search bar, a "Popular" sidebar with links to various services like Ensembl, UniProt, and PDB, and a grid of cards for Services, Research, Training, European Coordination, and EMBL ALUMNI. A sidebar on the right provides links to Visit EMBL.org and upcoming events.

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

A screenshot of the EMBL-EBI Services homepage. The layout is similar to the previous screenshot, featuring a dark blue header, a banner for "Services", and a "Popular" sidebar. The main content area shows a grid of cards for Bioinformatics services, including DNA & RNA, Gene expression, Proteins, Structures, Systems, Chemical biology, Ontologies, Literature, and Cross domain. A "Service news" section features a butterfly image and a "Training" section.

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

A screenshot of the EMBL-EBI Services homepage, identical in layout to the one above it. The main difference is that the "Proteins" card in the "Bioinformatics services" grid is highlighted with a red border, drawing attention to it as a specific feature or tool being discussed.

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

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

This screenshot shows the 'Proteins' section of the EBI website. It lists several popular services under the heading 'Popular services': UniProt (The Universal Protein Resource), InterPro (A database for the classification of proteins into families, domains and conserved sites), PRIDE (The Proteomics Identifications Database), Pfam (A database of hidden Markov models and alignments to describe conserved protein families and domains), Clustal Omega (Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools), HMMER - protein homology search (Fast sensitive protein homology searches using profile hidden Markov models (HMMs). Variety of different search methods for querying against both sequence and HMM target databases), and InterProScan 5 (InterProScan 5 searches sequences against InterPro's predictive protein signatures. Please note that InterProScan 4.8 has been retired). A 'Quick links' sidebar on the right provides links to '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

This screenshot shows the main homepage of the European Bioinformatics Institute (EMBL-EBI). It features a large banner for 'The European Bioinformatics Institute, Part of the European Molecular Biology Laboratory'. Below the banner, there is a search bar and a 'Find a gene, protein or chemical:' input field with examples like 'blast', 'protein', 'bif1...'. To the right, there are several colored boxes representing different areas: 'Services' (blue), 'Research' (green), 'Industry' (blue), 'European Coordination' (orange), 'EMBL ALUMNI' (grey), and 'News from EMBL-EBI' (grey). A prominent yellow box labeled 'Training' is highlighted with a red border. On the far right, there is a sidebar for 'Visit EMBL.org' with links to 'Popular' services, 'Research', 'Training', and 'News', and a section for 'Upcoming events' with a thumbnail for 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 training course titled 'Using sequence similarity searching tools at EMBL-EBI: webinar'. The course content is described as 'Using sequence similarity searching tools at EMBL-EBI: webinar'. It includes a video player showing a presentation by Andrew Cowley, a 'Popular' sidebar with links to 'Train online', 'Course list', 'Glossary', 'Support & Feedback', and 'Log In / Register', and a 'Find us at...' sidebar with links to 'Open days and career days', 'Exhibition exhibitions', 'EMBL courses and training', 'Genome campus events', and 'Balance for schools'.

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

This screenshot shows the 'Train online' portal of the EBI. It features a navigation bar with links to 'Databases', 'Tools', 'Research', 'Training', 'Industry', 'About Us', and 'Help'. The main content area is titled 'Train online' and displays a box for 'Notable EBI databases include: ENA, UniProt, Ensembl' and 'and the tools FASTA, BLAST, InterProScan, MUSCLE, DALI, HMMER'. Below this, there is a 'Find a course' section with a 'Browse by subject' dropdown and options for 'Genes and Genomes' and 'Gene Expression'.

Next Class...

MAJOR BIOINFORMATICS DATABASES AND ASSOCIATED ONLINE TOOLS

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, Picty_cDB, DIP, DOGS, DOMO, DPD, DPLinteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEKB, 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-Us, 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 >*

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, Picty_cDB, DIP, DOGS, DOMO, DPD, DPLinteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEKB, 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-Us, MPDB, MRR, MutBase, MycDB, NDB, NRSub, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, Subtilist, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS- MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc !!!

Side-note: Databases come in all shapes and sizes



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

Primary, secondary & composite databases

Bioinformatics databases can be usefully classified into *primary*, *secondary* and *composite* according to their data source.

- **Primary databases** (or archival databases) consist of data derived experimentally.
 - **GenBank:** NCBI's primary nucleotide sequence database.
 - **PDB:** Protein X-ray crystal and NMR structures.
 - **Secondary databases** (or derived databases) contain information derived from a primary database.
 - **RefSeq:** non redundant set of curated reference sequences primarily from GenBank
 - **PFAM:** protein sequence families primarily from UniProt and PDB
 - **Composite databases** (or metadatabases) join a variety of different primary and secondary database sources.
 - **OMIM:** catalog of human genes, genetic disorders and related literature
 - **GENE:** molecular data and literature related to genes with extensive links to other databases.

Your Turn!

https://bioboot.github.io/bimm143_F18/lectures/#1



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 (highlighted)

Computer Setup

Learning Goals

Assignments & Grading

Ethics Code

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

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.

BIMM-143: INTRODUCTION TO BIOINFORMATICS (Lecture 1)

Bioinformatics Databases and Key Online Resources
https://bioboot.github.io/bimm143_W18/lectures/#1

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
ATGGTGTGCACTTGACTCCCTGTGGAGAAGTCGTCGCGTTACTGGCCCTGTGGGCCAAAGTGAACTGTGATGAG
TGTGGTGTGGACGCCCTGGGGCAAGCTGGCCTGGTGTCCTACCCCTTGAGGACCAAGAGTTCTTGACTGCCTCTGGT
GGATCTGTCACACTCTGGTACATGAGTATGGCAACCTCTGAAGTCAGGCTCATGCGAAAGAACCTGGCTGGT
GCCCTTGTGCTGGGGCTGGCTGGCTACCTCTGGGACACCCTGAGGACCTTGACCAACTCTGGTGTGAGCTGACTGGCTGG
GTGACACAGCTGGCACGCTGGACCTGGACCTGGCTGCTGGCRAAGCTGCTGIVGCTNGVGTGCTGCCA
TCACCTTGGCCAARAGTAATCACCACCAAGTGCAGGCNCGCTATCAGAAAATGGTGCGCTGCTGTCATT
GCCCTGGCCCACAGTATCACTAAGCTGCGCTTCTGCTGTCACATT
```

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: [10:35 am]
2. GENE database @ NCBI	[10:55 am]
— BREAK —	— 11:05 am —
3. UniProt & Muscle @ EBI	[11:30 am]
4. PFAM, PDB & NGL	[12:00 pm]
— BREAK —	— 12:10 am —
5. Extension exercises	[12:40 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.
- There are a large number of primary, secondary and tertiary bioinformatics databases.
- The NCBI and EBI are major online bioinformatics service providers.
- Introduced Gene, UniProt, PDB databases as well as a number of 'boutique' databases including PFAM and OMIM.

HOMEWORK

https://bioboot.github.io/bimm143_F18/lectures/#1

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

THANK YOU