



**Office Hours:**  
Wed 2-3pm  
**Location:**  
???

## 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 what, why and how of bioinformatics?**

## Bioinformatics Database

**Hands-on** exploration of several major databases and their associated tools.

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



## Bioinformatics (BIMM 143, Winter 2018)



Course Director

Prof. Barry J. Grant (Email: [bjgrant@ucsd.edu](mailto:bjgrant@ucsd.edu))

Instructional Assistant

Alexander Sharp (Email: [arsharp@ucsd.edu](mailto:arsharp@ucsd.edu))

Course Syllabus

[Winter 2018 \(PDF\)](#)

## Overview

Bioinformatics - the application of computational and analytical methods to biological problems - is a rapidly maturing field that is driving the collection, analysis, and interpretation of the avalanche of data in modern life sciences and medical research.

This upper division 4-unit course is designed for biology majors and provides an introduction to the principles and practical approaches of bioinformatics as applied to genes and proteins.

An integrated lecture/lab structure with hands-on exercises and small-scale projects emphasizes modern developments in genomics and proteomics. A detailed listing of

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

**BIMM 143**  
A hands-on introduction to the computer-based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD.

**Learning Goals**

**Overview**  
Bioinformatics - the application of computational and analytical methods to biological problems - is a rapidly maturing field that is driving the collection, analysis, and interpretation of the avalanche of data in modern life sciences and medical research.

This upper division 4-unit course is designed for biology majors and provides an introduction to the principles and practical approaches of bioinformatics as applied to genes and proteins.

An integrated lecture/lab structure with hands-on exercises and small-scale projects emphasizes modern developments in genomics and proteomics. A detailed listing of

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



## Learning Goals

At the end of this course students will:

- Understand the increasing necessity for computation in modern life sciences research.
- Be able to use and evaluate online bioinformatics resources including major biomolecular and genomic databases, search and analysis tools, genome browsers, structure viewers, and select quality control and analysis tools to solve problems in the biological sciences.
- Be able to use the 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.

## Specific Learning Goals

## 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!

**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.

**Learning Goals**

	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 exact, soft and local alignments.	5



## Course Structure

Derived from specific 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.

**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 Winter 2018
1	Tu, 01/09	<b>Welcome to BioInformatics</b> 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 computer setup
2	Th, 01/11	<b>Bioinformatics databases and key online resources</b> NCBI & EBI resources for the molecular domain of bioinformatics, Focus on GenBank, UniProt, Entrez and Gene Ontology, Hands on with BLAST, GenBank, OMIM, GENE, UniProt, Muscle, PFAM and PDB bioinformatics tools and databases
3	Tu, 01/16	<b>Sequence alignment fundamentals, algorithms and applications</b>

# Course Structure

Derived from specific learning goals

The screenshot shows the 'Lectures' section of the BIMM 143 course website. A red box highlights the 'Lectures' link in the sidebar. The main content area displays a table of lectures for Winter 2018:

#	Date	Topics for Winter 2018
1	Tu, 01/09	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, Student computer setup
2	Th, 01/11	Bioinformatics databases and key online resources NCBI & EBI resources for the molecular domain of bioinformatics. Focus on GenBank, UniProt, Entrez and Gene Ontology. Hands on with BLAST, GenBank, OMIM, GENE, UniProt, Muscle, PFAM and PDB bioinformatics tools and databases
3	Tu,	Sequence alignment fundamentals, algorithms and applications Sequence alignment, local and global alignments, dynamic programming

# Class Details

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

The screenshot shows the '1: Welcome to Foundations of Bioinformatics' section of the BIMM 143 course website. A red box highlights the 'Questions' link in the sidebar. The main content area includes:

- Topics:** Course introduction, Learning goals & expectations, Biology is an information science, History of Bioinformatics, Types of data, Application areas and introduction to upcoming course segments, Student 30-second introductions, Student computer setup.
- Goals:**
  - Understand course scope, expectations, logistics and ethics code.
  - Understand the increasing necessity for computation in modern life sciences research.
  - Get introduced to how bioinformatics is practiced.
  - Complete the pre-course questionnaire.
  - Setup your laptop computer for this course.
- Material:**
  - Pre class screen casts (also see below):
    - SC1: Welcome to BIMM-143
    - SC2: What is Bioinformatics?
    - SC3: How do we do Bioinformatics?
  - Lecture Slides: Large PDF, Small PDF
  - Handout: Class Syllabus

# Homework

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

The screenshot shows the 'Homework' section of the BIMM 143 course website. A red box highlights the 'Questions' link in the sidebar. The main content area includes:

- Homework:**
  - Questions
  - Readings:
    - PDF1: What is bioinformatics? An introduction and overview
    - PDF2: Advancements and Challenges in Computational Biology
    - Other: For Big Data Scientists, 'Janitor Work' Is Key Hurdle to Insights New York Times, 2014.
- Screen Casts:**
  - Welcome to "Foundations of Bioinformatics" (BGGN-2...)

# Homework

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

The screenshot shows the 'Homework' section of the BIMM 143 course website. A red box highlights the 'Questions' link in the sidebar. The main content area includes:

- Homework:**
  - Questions
  - Readings:
    - PDF1: What is bioinformatics? An introduction and overview
    - PDF2: Advancements and Challenges in Computational Biology
    - Other: For Big Data Scientists, 'Janitor Work' Is Key Hurdle to Insights New York Times, 2014.
- Screen Casts:**
  - Welcome to "Foundations of Bioinformatics" (BGGN-2...)

# Homework

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

The screenshot shows a Google Form titled "BIMM143 Lecture 1 Homework (W18)". It begins with a header "Please answer the following questions" and a note "Required". There is a field for "Email address" with the placeholder "Your email". Below it is a question: "Which of the following operating systems is most frequently used for bioinformatics tool development" with four options: Windows, iOS, Unix, and Perl. A red asterisk indicates this is required. Another question follows: "Which of the following databases contains primarily protein sequences" with the same four options. A small "1 point" is listed next to each question.

# Homework

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

This screenshot is identical to the one above, showing the same Google Form. However, it features a prominent red diagonal banner across the middle with the text "Homework is due before the next weeks class!".

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

UCSanDiego

**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
Use R to read and parse comma-separated (.csv) formatted files ready for subsequent analysis.	8, 9, 10, 11, 13, 15, 16
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
View and Interpret the structural models in the PDB.	10, 11
Explain the outputs from structure prediction algorithms and small molecule docking approaches.	11
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
Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.	13
Given 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 expressed genes and use online tools to interpret gene lists and annotate potential gene functions.	
Perform a GO analysis to identify the pathways relevant to a set of genes (e.g. identified by transcriptomic study or a proteomic experiment).	15
Use the KEGG pathway database to look up interaction pathways.	16
Use graph theory to represent biological data networks.	17
Understand the challenges in Integrating and Interpreting large heterogeneous high throughput data sets into their functional	18
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.	19

# BIMM-143 Learning Goals....

Delve deeper into “real-world” bioinformatics

UCSanDiego

**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

Explain the outputs from structure prediction algorithms and small molecule docking approaches.	11
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
Understand the process by which genomes are currently sequenced and the bioinformatics processing and analysis required for their interpretation.	13
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 expressed genes and use online tools to interpret gene lists and annotate potential gene functions.	15, 16
Perform a GO analysis to identify the pathways relevant to a set of genes (e.g. identified by transcriptomic study or a proteomic experiment).	16
Use the KEGG pathway database to look up interaction pathways.	17
Use graph theory to represent biological data networks.	17, 18
Understand the challenges in Integrating and Interpreting large heterogeneous high throughput data sets into their functional	19

## These support a major learning objective

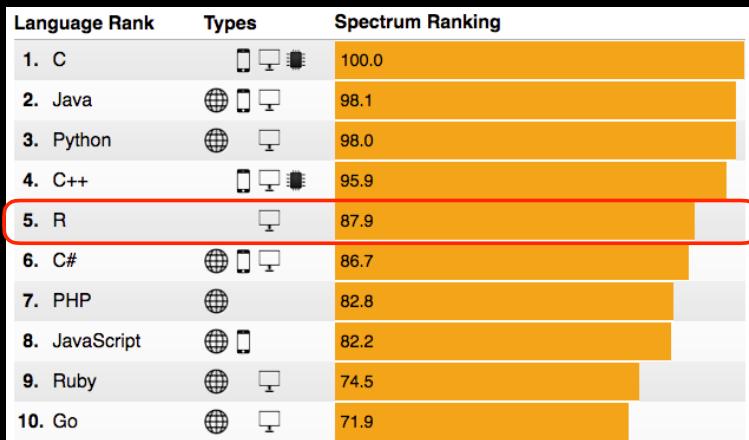
At the end of this course students will:

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

## Why use R?

Productivity  
Flexibility  
Designed for data analysis

## IEEE 2016 Top Programming Languages



<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 (Tobe Index)



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

Python

R



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

- R is the “lingua franca” of data science in industry and academia.
- Large user and developer community.
  - As of Jan 8th 2018 there are 12,039 add on **R packages** on **CRAN** and 1,473 on **Bioconductor** - more on these later!
- Virtually every statistical technique is either already built into R, or available as a free package.
- Unparalleled exploratory data analysis environment.

## Today's Menu

### Course Logistics

Website, screencasts, survey, ethics, assessment and grading.

### Learning Objectives

What you need to learn to succeed in this course.

### Course Structure

Major lecture topics and specific learning goals.

### Introduction to Bioinformatics

Introducing the *what, why and how* of bioinformatics?

### Computer Setup

Ensuring your laptop is all set for future sections of this course.

# 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>)

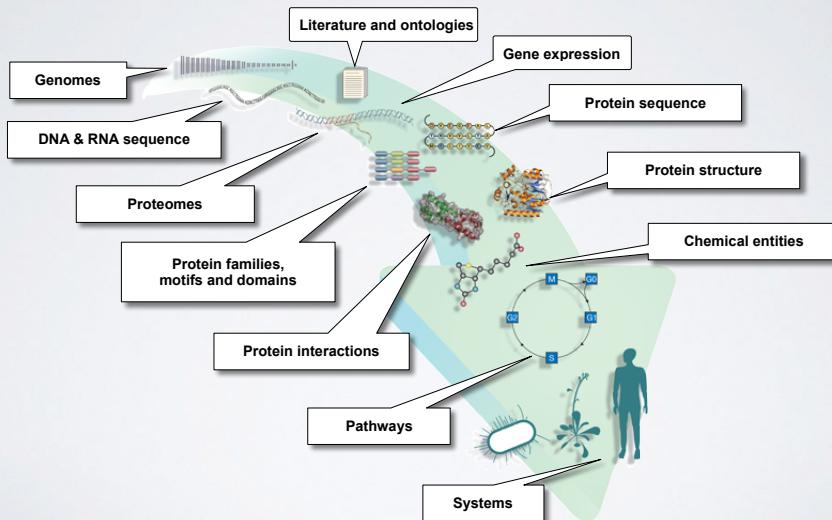
## MORE DEFINITIONS

- “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “**informatics**” techniques (derived from disciplines such as applied 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.

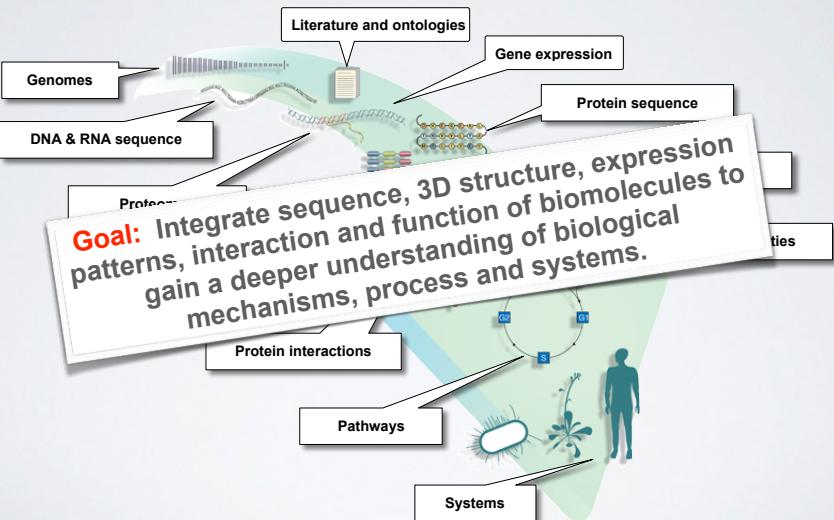
- “Bioinformatics is research, development, or application of **computational approaches** for expanding the use of **biological, medical, behavioral or health data**, including those to **acquire, store, organize and analyze** such data.”

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

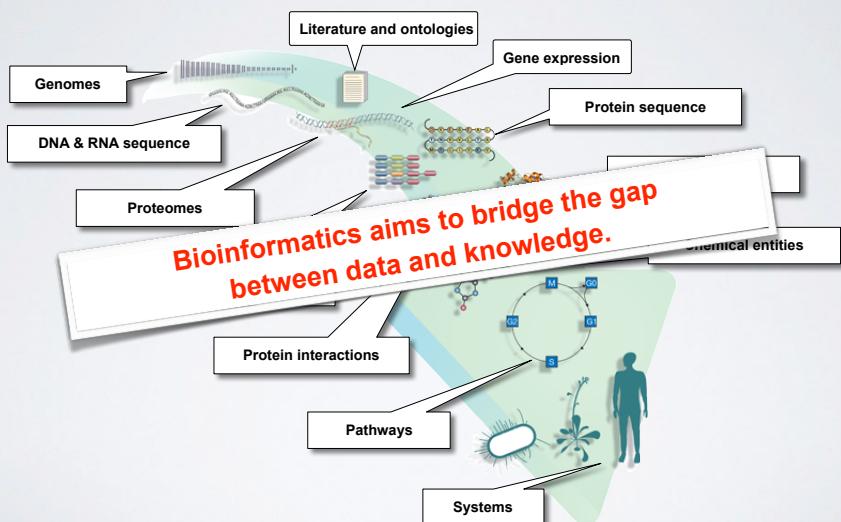
## Major types of Bioinformatics Data



## 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

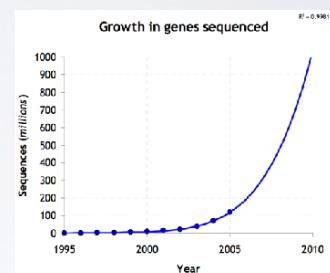
- DNA sequence determines protein sequence.
- Protein sequence determines protein structure.
- Protein structure determines protein function.
- Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.

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

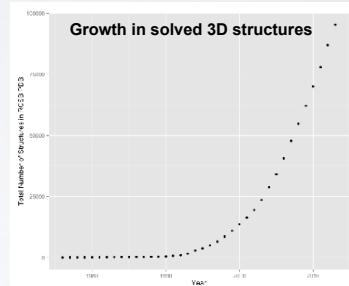


E.G. data from sequencing, structural genomics, proteomics, new high throughput assays, etc...

## Why do we need Bioinformatics?

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

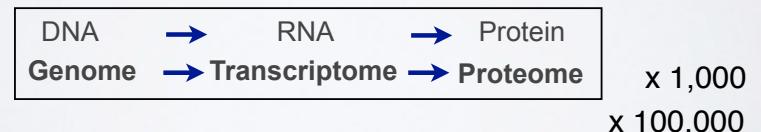
- Bioinformatics provides methods for the efficient:
  - **storage**
  - **annotation**
  - **search and retrieval**
  - **data integration**
  - **data mining and analysis**



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?

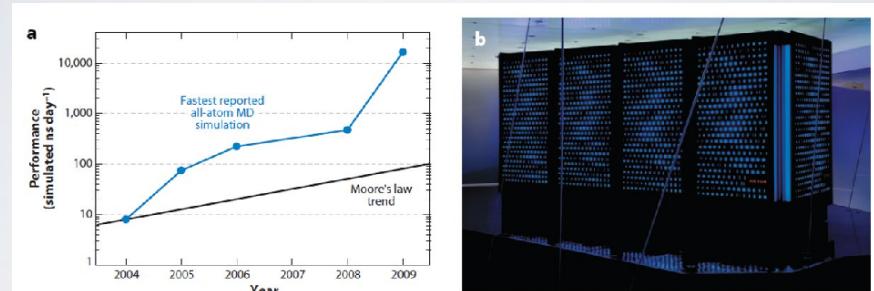
### Pre-packaged tools and databases

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

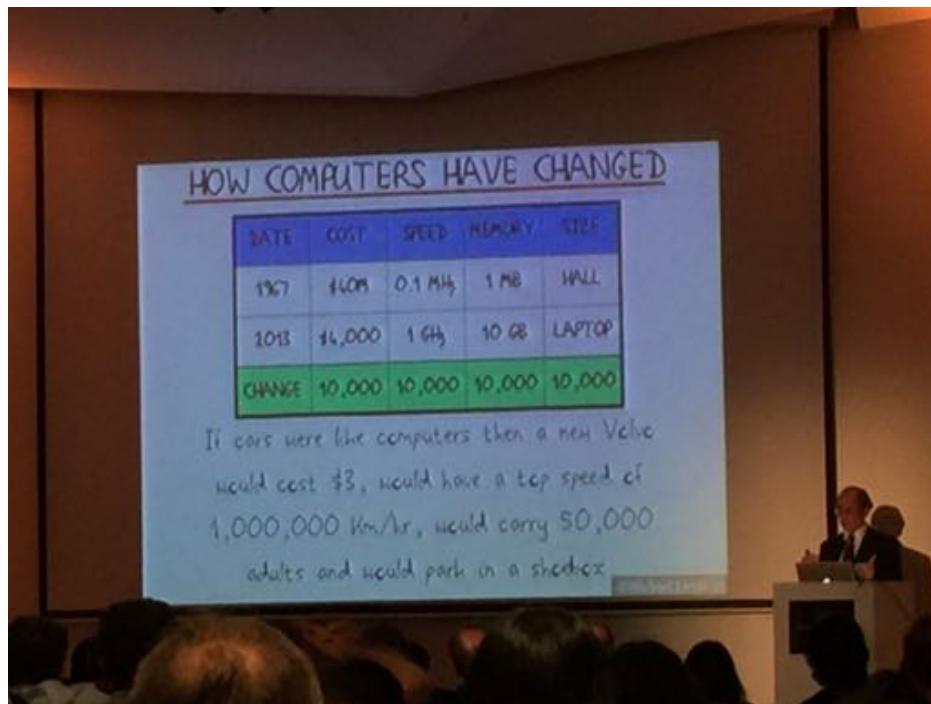
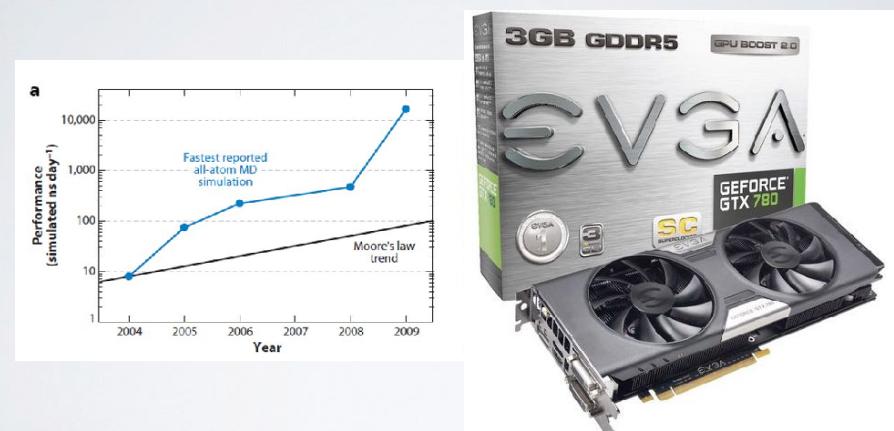
### Advanced tool application & development

- 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 his old but quite readable treatise, *Sequence Analysis in Molecular Biology; Treasure Trove or Trivial Pursuit*, provides a very appropriate conclusion:

- “Think about what you’re doing; use your knowledge of the molecular system involved to guide both your interpretation of results and your direction of inquiry; use as much information as possible; and do not blindly accept everything the computer offers you”.
- Key-Point: **Avoid the miss-use of ‘black boxes’!**

The screenshot shows the NCBI Protein BLAST search interface. It includes sections for General Parameters (Max target sequences: 500, Short queries checked, Expect threshold: 10, Word size: 3, Max matches in a query range: 0), Scoring Parameters (Matrix: BLOSUM62, Gap Costs: Existence: 11 Extension: 1, Compositional adjustments: Conditional compositional score), Filters and Masking (Filter: Low complexity regions, 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 "Even Blast has many settable parameters". Another callout box highlights "Related tools with different terminology" showing the BLOSUM matrix and gap penalties.

## Common problems with Bioinformatics

Confusing multitude of tools available

- Each with many options and settable parameters

Most tools and databases are written by and for nerds

- Same is true of documentation - if any exists!

Most are developed independently

Notable exceptions are found at the:

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

## Key Online Bioinformatics Resources: NCBI & EBI

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

The screenshot shows the NCBI homepage. It features a sidebar with links like "Resource List A-Z", "All Resources", "Data & Software", "Domain & Structures", "Diseases", "Genetics & Medicine", "Genomes & Maps", "Homology", "Microbiology", "Proteins", "Sequence Analysis", "Training", "Testing & Utilities", and "Validator". The main content area includes sections for "Welcome to NCBI", "Popular Resources" (Pubmed, Bookshelf, PubMed Health, Research, ReView, Rego, BioRxiv, BioMed Central, BioProtocols), "Get Started" (with links to "Analyze data using NCBI software", "Get NCBI data or software", "Search", "Learn how to access our services from NCBI databases", and "Submit your data to NCBI databases"), "3D Structures", "NCBI Announcements", and "New version of Genome Workbench". A callout box highlights the URL <http://www.ncbi.nlm.nih.gov>.

The screenshot shows the EBI homepage. It features a sidebar with links like "Home", "About", "Services", "Data", "Tools", "Events", "Publications", and "Contact". The main content area includes sections for "The European Bioinformatics Institute", "What is EBI?", "Research", "Data", "Tools", "Events", "Publications", and "Contact". A callout box highlights the URL <https://www.ebi.ac.uk>.

# National Center for Biotechnology Information (NCBI)

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



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

The screenshot shows the NCBI homepage with the "Popular Resources" section highlighted. Red arrows point to the links for PubMed, BLAST, Nucleotide, SNP, Gene, Protein, and PubChem. The "3D Structures" section is also visible at the bottom left.

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

The screenshot shows the NCBI homepage with the "Popular Resources" section on the right. It includes links to PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. At the bottom, there is a banner for "3D Structures" featuring a molecular model.

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

A large callout box on the right side of the page highlights "Notable NCBI databases include: GenBank, RefSeq, PubMed, dbSNP and the search tools ENTREZ and BLAST". The "Popular Resources" section is also visible on the right.

# Key Online Bioinformatics Resources: NCBI & EBI

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

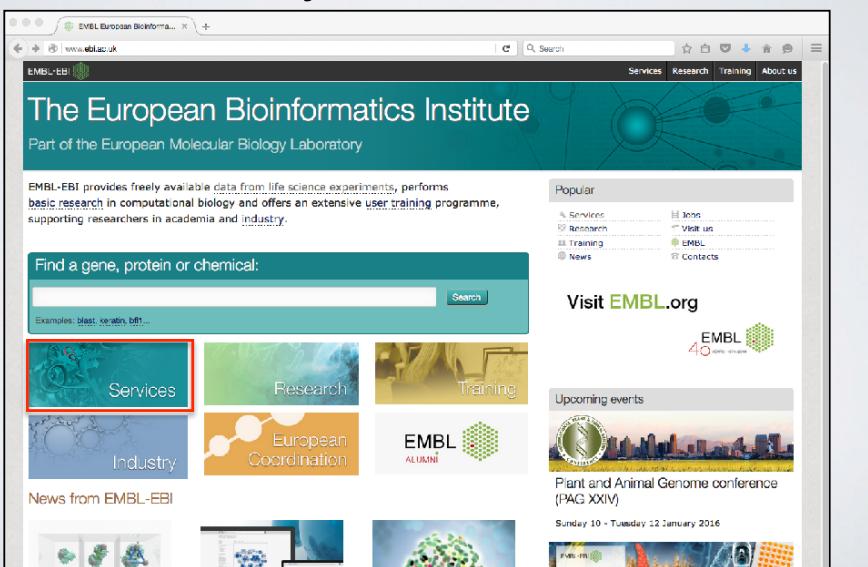


<http://www.ncbi.nlm.nih.gov>



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

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

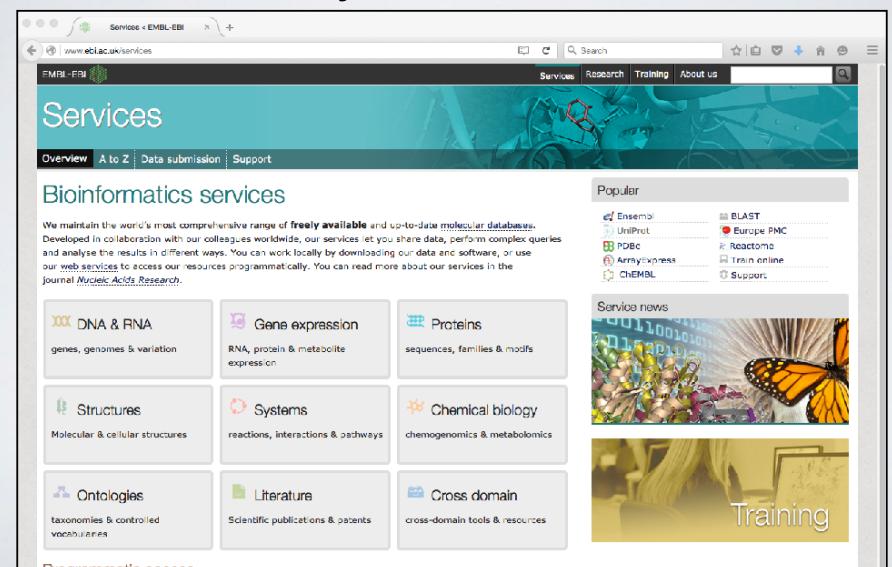


# European Bioinformatics Institute (EBI)

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



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



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

Services

**Bioinformatics services**

We maintain the world's most comprehensive range of **freely available** and up-to-date molecular databases. Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our web services to access our resources programmatically. You can read more about our services in the journal *Nucleic Acids Research*.

**Popular**

- Ensembl
- UniProt
- PDBe
- ArrayExpress
- ChEMBL

**DNA & RNA**  
genes, genomes & variation

**Gene expression**  
RNA, protein & metabolite expression

**Proteins**  
sequences, families & motifs

**Structures**  
Molecular & cellular structures

**Systems**  
reactions, interactions & pathways

**Chemical biology**  
chemogenomics & metabolomics

**Ontologies**  
taxonomies & controlled vocabularies

**Literature**  
Scientific publications & patents

**Cross domain**  
cross-domain tools & resources

**Training**

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

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

## Proteins

**Popular services**

- UniProt: The Universal Protein Resource**  
The gold-standard, comprehensive resource for protein sequence and functional annotation data.
- InterPro**  
A database for the classification of proteins into families, domains and conserved sites.
- PRIDE: The Proteomics Identifications Database**  
An archive of protein expression data determined by mass spectrometry.
- 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.
- InterProScan 5**  
InterProScan 5 searches sequences against InterPro's predictive protein signatures. Please note that InterProScan 4.8 has been retired.

**Quick links**

- Popular services in this category
- All services in this category
- Project websites in this category

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

The European Bioinformatics Institute  
Part of the European Molecular Biology Laboratory

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

Find a gene, protein or chemical:

Examples: blast, creatin, bfl1...

**Popular**

- Services
- Research
- Training
- News

Visit [EMBL.org](#)

Upcoming events

Plant and Animal Genome conference (PAG XXIV)  
Sunday 10 - Tuesday 12 January 2016

Services      Research      European Coordination      Training      EMBL ALUMNI

News from EMBL-EBI

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

Train online

Training > Train online Home > Course list > Glossary > Support & Feedback > Log In / Register

**Course content**

Using sequence similarity searching tools at EMBL-EBI: webinar

**Using sequence similarity search tools at EMBL-EBI: webinar**

Andrew Cooley and the EMBL-EBI support team

This webinar focuses on how to use tools like **BLAST** and PSI-Search to find homologous sequences in EMBL-EBI databases, including tips on which tool and database to use, input formats, how to change parameters and how to interpret the results pages.

**Popular**

- Train online
- Find us
- Funding

**Find us at...**

- Open days and career days
- Conference exhibitions
- EMBL courses and events
- Science careers events
- Schools for schools

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

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

## Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, BiolImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty\_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlilesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVbase, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klothe, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPep5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, MycDB, NDB, NRSub, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS- MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc ..... !!!

Next Class...

## MAJOR BIOINFORMATICS DATABASES AND ASSOCIATED ONLINE TOOLS

## Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, BiolImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty\_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genlilesne, GenLink, GENOTK, GenProtEC, GIFTS, GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB, HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD, HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVbase, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klothe, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPep5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, MycDB, NDB, NRSub, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS- MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc ..... !!!

*There are lots of Bioinformatics Databases  
For a annotated listing of major bioinformatics databases please see the online handout  
< Major Databases.pdf >*

# Side-note: Databases come in all shapes and sizes



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

## Primary, secondary & composite databases

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

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

## Today's Menu

<b>Course Logistics</b>	Website, screencasts, survey, ethics, assessment and grading.
<b>Learning Objectives</b>	What you need to learn to succeed in this course.
<b>Course Structure</b>	Major lecture topics and specific learning goals.
<b>Introduction to Bioinformatics</b>	Introducing the <i>what, why</i> and <i>how</i> of bioinformatics?
<b>Bioinformatics Database</b>	<b>Hands-on</b> exploration of several major databases and their associated tools.

## Your Turn!

[https://bioboot.github.io/bimm143\\_W18/lectures/#1](https://bioboot.github.io/bimm143_W18/lectures/#1)

A screenshot of a web browser displaying the first lecture page for BIMM143. The page features the UC San Diego logo and the course title 'BIMM 143'. Below this, there is an 'Overview' section with a list of links: 'Lectures' (which is highlighted with a red box), 'Computer Setup', 'Learning Goals', 'Assignments &amp; Grading', and 'Ethics Code'. To the right of the 'Overview' section, there are sections for 'Topics', 'Goals', and 'Material'. The 'Goals' section contains a bulleted list of learning objectives, and the 'Material' section lists various resources with download links. A red box also highlights the 'Feedback: Muddy Point Assessment' link under the 'Material' section.

**BIMM-143: INTRODUCTION TO BIOINFORMATICS (Lecture 1)****Bioinformatics Databases and Key Online Resources**[https://bioboot.github.io/bimm143\\_W18/lectures/#1](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.

&gt;example1

```
ATGGTGCAATCTGAATCCATTGAGAGAACGCTGCCGTAACTCTGCCTGGGGCAGGGTGAACTGTGCGTCTGAG  
TTGGTGTTGGAGGCCCTGGCTCTGGCTGGCTCTGGCGACGAGCTTGCTGGTGACCCAAAGCTGTCTTGGCTGTCTGG  
GGATCTGGTCGATGCGCTCGATTCAGTCATTAATGGCGAAACCTTCATGGTGCGCTACTGCAGAAAGTGCTGGT  
GCCTGCTGTTAATGGCCCTGGCTCATGGCTGCTGCTGCGAAACCTTCATGGTGCGCTACTGCAGAAAGTGCTGGT  
GTGACGACCTCTGGCACATCTGAATCTGGGAACTCTGGCTGGCTGGCCAACTCTGGCTGTGGTGCTGGCCAA  
TGACCTTGCGAAAGAATTCACCCCAACGTGCAGCTGCTGCTGCATGCACTGTGGTGCTGGCTGGCTGGCTGAA  
GCCCTGGCCCAACAGTGATCCTAGACTGGCTTCATGTGAAAGTGGTGGCTGGCTGGCTGGCTGAA
```

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:

[10:45 am]

[11:00 am]

— 11:10 am —

[11:35 am]

[12:05 pm]

— 12:15 am —

[12:45 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\\_S18/lectures/#1](https://bioboot.github.io/bimm143_S18/lectures/#1)

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

THANK YOU