

**BIMM 143**  
**Introduction to Bioinformatics**  
 Barry Grant  
 UC San Diego  
<http://thegrantlab.org/bimm143>

**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

<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>	<b>Introducing the <i>what, why</i> and <i>how</i> of bioinformatics?</b>
<b>Bioinformatics Database</b>	<b>Hands-on</b> exploration of several major databases and their associated tools.

http://thegrantlab.org/bimm143/

UC San Diego

## Bioinformatics (BIMM 143, Fall 2018)

**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**

**Assignments & Grading**

**Ethics Code**

**Course Director**  
Prof. Barry J. Grant (Email: bjgrant@ucsd.edu)

**Instructional Assistant**  
Chao Shi (Email: bioshichao@gmail.com)

**Course Syllabus**  
Fall 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.

http://thegrantlab.org/bimm143/

UC San Diego

## Bioinformatics (BIMM 143, Fall 2018)

**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**

**Assignments & Grading**

**Ethics Code**

**Course Director**  
Prof. Barry J. Grant (Email: bjgrant@ucsd.edu)

**Instructional Assistant**  
Chao Shi (Email: bioshichao@gmail.com)

**Course Syllabus**  
Fall 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.

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

UC San Diego

## Bioinformatics (BIMM 143, Fall 2018)

**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**

**Assignments & Grading**

**Ethics Code**

### 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!

**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, BLAST, BLAST, BLAST, and BLAST searches using local databases.	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	<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, Hands on with major Bioinformatics databases and key online NCBI and EBI resources
2	Th, 04/05	<b>Sequence alignment fundamentals, algorithms and applications</b> 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
		<b>Advanced sequence alignment and database searching</b>

## 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	<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, Hands on with major Bioinformatics databases and key online NCBI and EBI resources
2	Th, 04/05	<b>Sequence alignment fundamentals, algorithms and applications</b> 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
		<b>Advanced sequence alignment and database searching</b>

# Class Details

## Goals, Class material, Screencasts & Homework

**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**

**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:**

- 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

**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**

**Assignments & Grading**

**Ethics Code**

### 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)

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

# Homework

## Goals, Class material, Screencasts & Homework

**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**

**Assignments & Grading**

**Ethics Code**

### 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)

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

# Homework

## Goals, Class material, Screencasts & Homework

**BIMM143 Lecture 1 Homework**

Please answer the following questions

\* Required

Email address \*

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

GenBank



# 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

GenBank

## 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

5	Calculate the alignment score between two nucleotide or protein sequences using a provided scoring matrix and be able to perform BLAST, PSI-BLAST, HMMER and protein structure based database searches and interpret the results in terms of the biological significance of an e-value.	5, 10
6	Use 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. Given an RNA-Seq data file, find the set of significantly differentially	14

# BIMM-143 Learning Goals....

Delve deeper into “real-world” bioinformatics

Goal Number	Goal Description	Page Reference
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

## 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

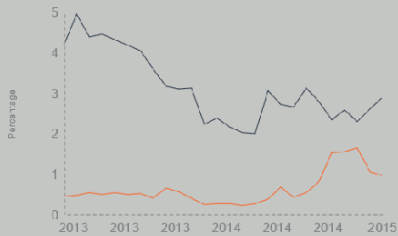
Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

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

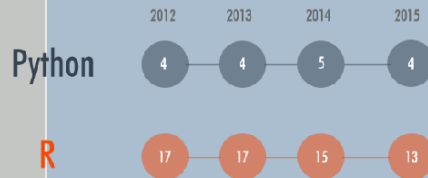
## R and Python: The Numbers

### Popularity Rankings

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



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



### Jobs And Salary?

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



\$115,531



\$94,139

[http://www.kdnuggets.com/2015/05/r-vs-python-data-science.html?utm\\_medium=email&utm\\_source=flipboard](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 leaning goals.
Introduction to Bioinformatics	Introducing the <i>what, 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> )

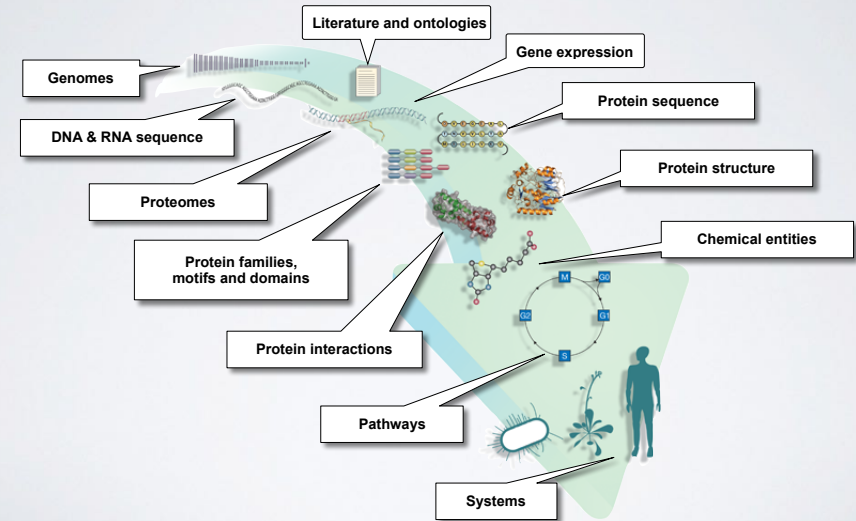
## MORE DEFINITIONS

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

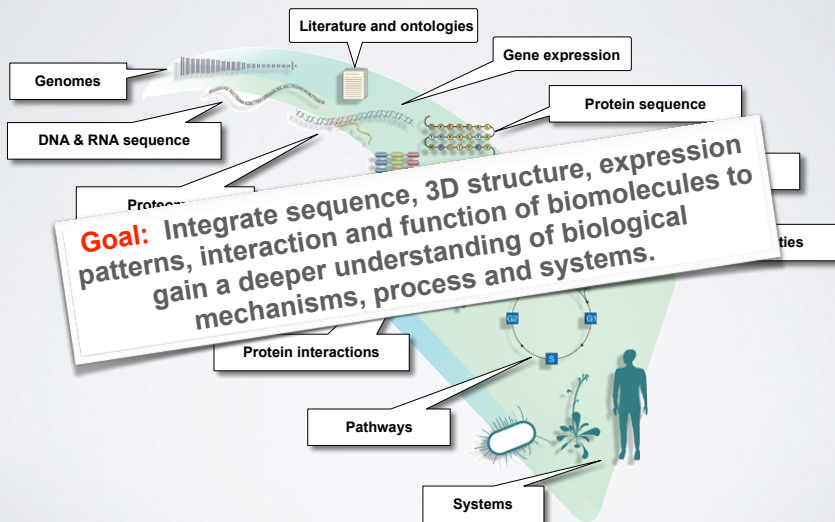
- ▶ “Bioinformatics is the research, development, or application of **computational approaches** for expanding the use of biological, **medical, behavioral or health data**, including those to **acquire, store, organize and analyze** such data.”  
National Institutes of Health (NIH) ( <http://tinyurl.com/l3gxr6b> )

**Key Point: Bioinformatics is Computer Aided Biology**

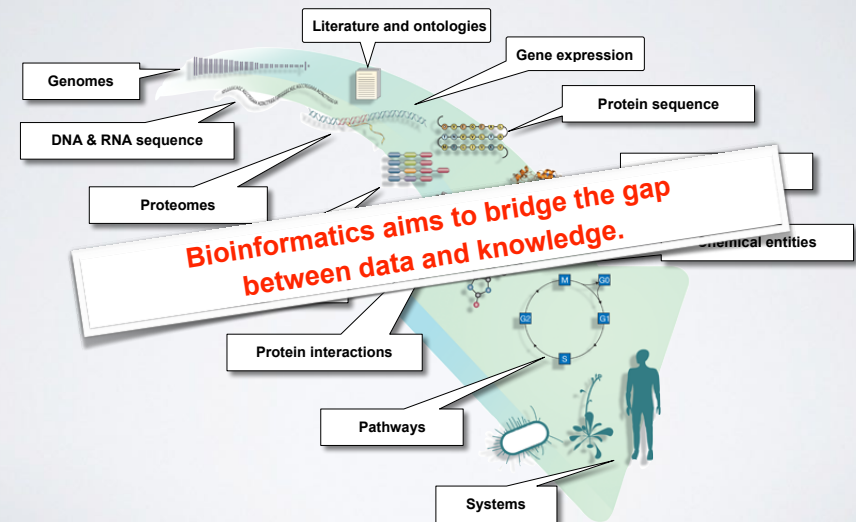
## Major types of Bioinformatics Data



## 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

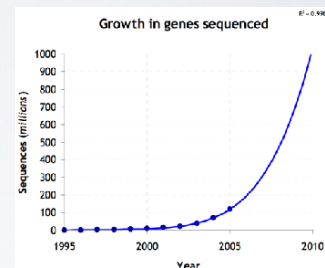
- DNA sequence determines protein sequence.
- Protein sequence determines protein structure.
- Protein structure determines protein function.
- Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.

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

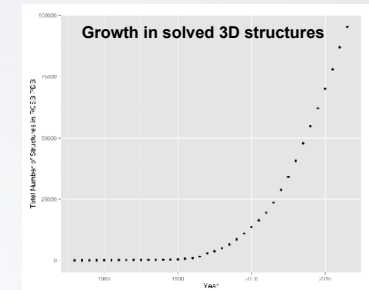


E.G. data from sequencing, structural genomics, proteomics, new high throughput assays, etc...

## Why do we need Bioinformatics?

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

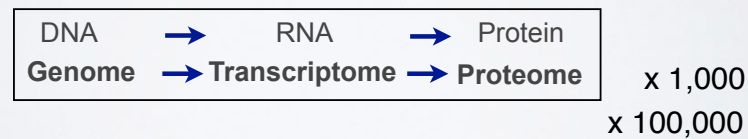
- Bioinformatics provides methods for the efficient:
  - ▶ storage
  - ▶ annotation
  - ▶ search and retrieval
  - ▶ data integration
  - ▶ data mining and analysis



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?

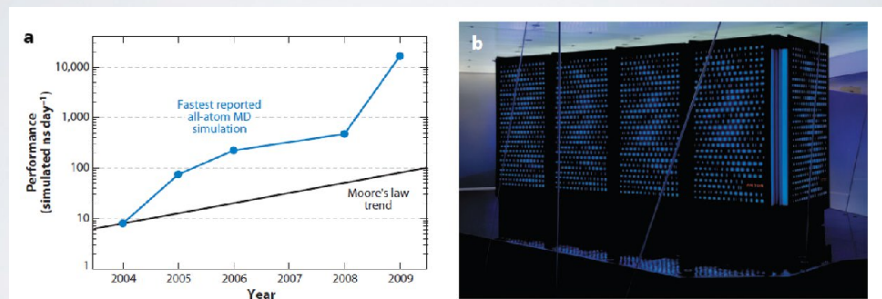
### Pre-packaged tools and databases

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

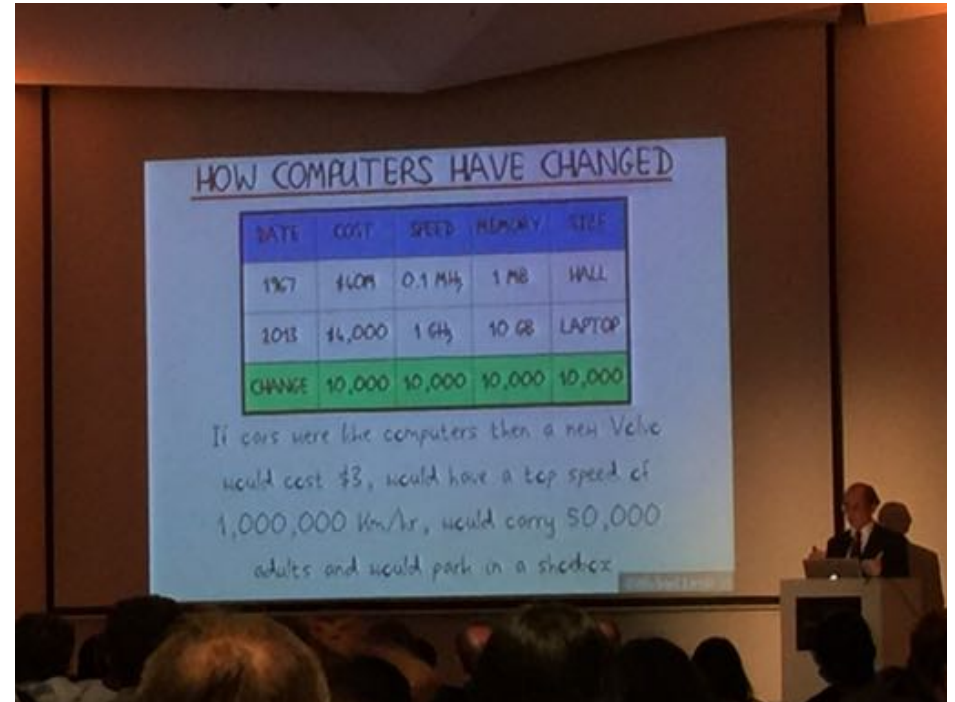
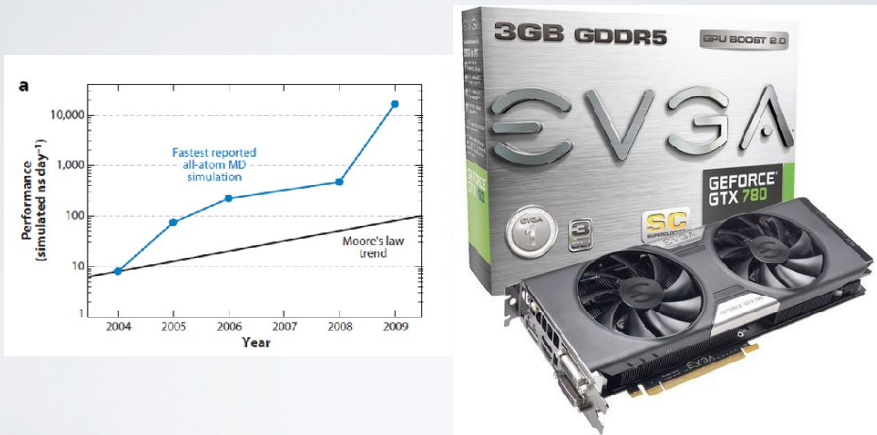
### Advanced tool application & development

- 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'!**





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

The screenshot shows the NCBI homepage with a navigation menu on the left and a 'Popular Resources' sidebar on the right. The sidebar lists: PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem. The main content area includes a 'Welcome to NCBI' message, a 'Get Started' section with links to Tools, Downloads, How-To's, and Submissions, and a '3D Structures' section.

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

This screenshot is similar to the previous one but with a 'Popular Resources' dropdown menu open. Red arrows point to 'PubMed', 'BLAST', 'Nucleotide', and 'Gene' in the dropdown list. The main content area is partially obscured by the dropdown.

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

The screenshot shows the NCBI homepage with a white text box overlaid in the center. The text box contains the following text: 'Notable NCBI databases include: **GenBank**, **RefSeq**, **PubMed**, **dbSNP** and the search tools **ENTREZ** and **BLAST**'. The background shows the same homepage layout as the previous screenshots.

## Key Online Bioinformatics Resources: NCBI & EBI

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

A small thumbnail version of the NCBI homepage screenshot, showing the navigation menu and 'Popular Resources' sidebar.

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

A small thumbnail version of the EBI homepage, showing the header 'The European Bioinformatics Institute' and various navigation options.

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

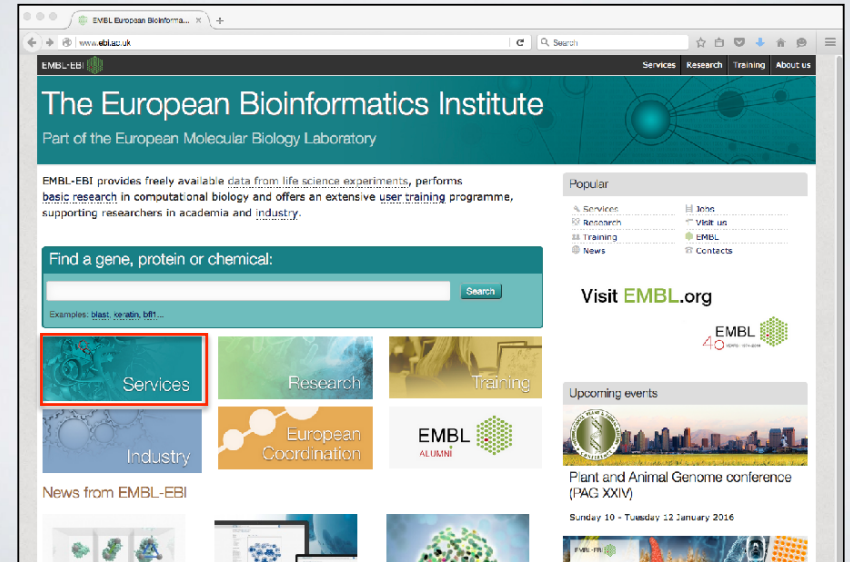


# European Bioinformatics Institute (EBI)

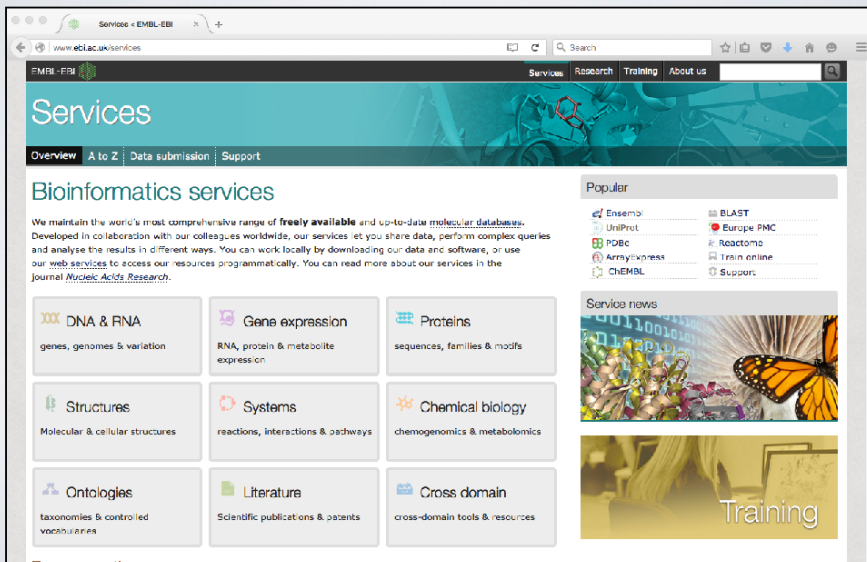
- Created in 1997 as a part of the European Molecular Biology Laboratory (EMBL)
- EBI's mission includes:
  - providing freely available **data** and **bioinformatics services**
  - and providing advanced **bioinformatics training**
- We will briefly cover several EBI databases and tools that have advantages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



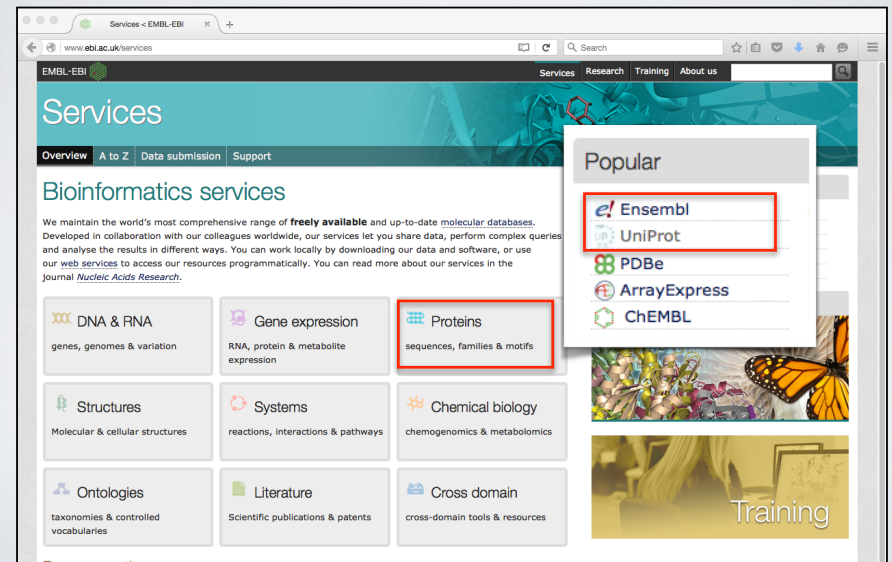
The EBI maintains a number of high quality curated **secondary databases** and associated tools



The 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



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

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

The screenshot shows the 'Proteins' section of the EBI website. It features a list of 'Popular services' on the left and 'Quick links' on the right. The services listed include UniProt, InterPro, PRIDE, Pfam, Clustal Omega, HMMER, and InterProScan 5. Each service is accompanied by a small icon and a brief description of its function.

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

The screenshot shows the EBI homepage. The 'Training' link in the main navigation bar is highlighted with a red box. The page includes a search bar, a 'Find a gene, protein or chemical' section, and various service tiles for Services, Research, Training, Industry, and European Coordination. A 'Visit EMBL.org' banner and 'Upcoming events' section are also visible.

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

The screenshot shows an online training webinar page titled 'Using sequence similarity searching tools at EMBL-EBI: webinar'. The page features a video player with a thumbnail of the presenter, Andrew Cowley. The course content is detailed, and there are navigation options for 'Train online', 'Find a course', and 'Support & Feedback'.

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

The screenshot shows the EBI online training page. A text overlay in the center lists notable EBI databases and tools. Below the overlay, the 'Find a course' section is visible, showing a 'Browse by subject' menu with options like 'Genes and Genomes', 'Gene Expression', and 'Interactions, Databases and Networks'.

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

## Next Class...

# MAJOR BIOINFORMATICS DATABASES AND ASSOCIATED ONLINE TOOLS

## Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Bearref, 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, Pcty\_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, HIBD, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-U, MPDB, MRR, MutBase, MycDB, NDB, NRSUB, O-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS- MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc ..... !!!!

## Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Bearref, 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, Pcty\_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, HIBD, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB, HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5, Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-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 an annotated listing of major bioinformatics databases please see the online handout  
< [Major Databases.pdf](#) >

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



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



# 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

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

# Your Turn!

[https://bioboot.github.io/bimm143\\_F18/lectures/#1](https://bioboot.github.io/bimm143_F18/lectures/#1)

The screenshot shows a web browser window with the URL [https://bioboot.github.io/bimm143\\_F18/lectures/#1](https://bioboot.github.io/bimm143_F18/lectures/#1). The page title is "1: Welcome to Foundations of Bioinformatics". The left sidebar has a menu with "Lectures" highlighted in a red box. The main content area includes "Topics", "Goals", and "Material" sections. The "Material" section lists "Lecture Slides: Large PDF" and "Lab: Hands-on section worksheet" (both highlighted in red boxes), and "Feedback: Muddy Point Assessment".

## 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.

>example1

```
ATGGTGCACTCTGACTCTCTGTGGAGAGTCTCCGCTTACTGCCCTGTGGGCAAGGTGACGTGGATGAG  
TTGGTGGTGGAGCCCTGGGAGAGCTGGCTGGTGGTACCTCTGGACACAGAGTCTTTGAGTCTTGG  
GGACTCTTCCACTCTGATGCACTATGGGCAACCTTAGGTGAGGCTCATGGCAAGAAGTCTGGT  
GCCTTAGTATGATGCTGGCTACCTGGACACCTCAGGGCCACTTGGCCACTGAGTGGCTGCACT  
GTGACAGCTGCACCTGGATCCTGAGAAGTCTAGGCTCTGGCCAAAGCTGCTGTGTGTGGCCCA  
TCACCTTGGCAAGAATTACCCACCACTGAGGCTGCCATCAGAAAGTGTGGCTGGTGGCTAAT  
GCCCTGGCCACAGATCACTAAGCTGGCTTCTTGTGTGCTCAATTT
```

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:

- |  |              |
|--|--------------|
|  | End times:   |
| 1. BLAST, GenBank and OMIM @ <b>NCBI</b> | [10:35 am]   |
| 2. GENE database @ <b>NCBI</b>           | [10:55 am]   |
| — BREAK —                                | — 11:05 am — |
| 3. UniProt & Muscle @ <b>EBI</b>         | [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](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