

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/>

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**

Bioinformatics (BIMM 143, Winter 2018)

Course Director: Prof. Barry J. Grant (Email: bigrant@ucsd.edu)
Instructional Assistant: Alexander Sharp (Email: arsharp@ucsd.edu)
Course Syllabus: Winter 2018 (PDF) [\[link\]](#)

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 review 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!

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

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 Be able to query, search, compare and contrast the data contained in major bioinformatics databases and describe how these databases intersect (GenBank, GEHE, UniProt, PfAM, OMIM, PDB, UCSC, ENSEMBLE).	1, 2, 20
2	Z, 12, 13
3 Describe how nucleotide and protein sequence and structure data are represented (FASTA, FASTQ, GenBank, UniProt, PDB). 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.	3, 10
4 Calculate the alignment score between two nucleotide or protein sequences using a provided scoring matrix and be able to perform BLAST, PSI-BLAST, MMseqs2 and search against databases.	4, 5

Course Structure

Derived from specific learning goals

The screenshot shows a Microsoft Word document with the following content:

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	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, 01/16	Sequence alignment fundamentals, algorithms and applications

BIMM 143

A hands-on introduction to the computer based analysis of genomic and biomolecular data from the Division of Biological Sciences, UCSD. [\[link\]](#)

Overview

- [Lectures](#)
- [Computer Setup](#)
- [Learning Goals](#)
- [Assignments & Grading](#)
- [Ethics Code](#)

Course Structure

Derived from specific learning goals

Course Structure

Derived from specific learning goals

All Lectures are Tu/Th 9:00-12:00 pm in Warren Lecture Hall 215 (WLH 215) ([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	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, Rous on GenBank, UniProt, Entrez and Gene Ontology, Handes on with BLAST, GenBank, OMIM, GENE, UniProt, Pfam and PDB bioinformatics tools and databases
	Tu,	Sequence alignment fundamentals, algorithms and applications

The screenshot shows a Microsoft Word document window. The title bar reads "BIMM 143: Welcome to Foundations of Bioinformatics". The main content area contains the following text:

Class Details

Goals, Class material, Screencasts & Homework

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 BIMM143 course homepage. In the top right corner, there is a 'Homework' section with a red box highlighting the 'Questions' link. Below it, there are links to PDF documents and other resources. At the bottom, there is a 'Screen Casts' section featuring a video thumbnail of a man speaking.

Homework

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

The screenshot shows the 'BIMM143 Lecture 1 Homework (W18)' page. It has a blue header and asks users to answer questions. There are two questions with radio button options and a 'Feedback' link. The first question is about operating systems and the second is about protein databases.

Homework

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

The screenshot shows the same homework page as above, but with a large red diagonal banner across it that says 'Homework is due before the next weeks class!'. The rest of the page content is visible below the banner.

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

The screenshot shows the 'Learning Goals' section of the BIMM143 website. A red box highlights the 'Learning Goals' link in the sidebar. To the right is a table of learning objectives numbered 5 through 14, each with a brief description and associated points.

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.	14
	Given an RNA-Seq data file, find the set of significantly differentially	

BIMM-143 Learning Goals....

Delve deeper into “real-world” bioinformatics

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

Overview
Lectures
Computer Setup
Learning Goals
Assignments & Grading
Ethics Code

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 neighbourhood 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
13 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
14 Use the KEGG pathway database to look up interaction pathways. 17
15 Use graph theory to represent biological data networks. 17, 18
16 Understand the challenges in Integrating and Interpreting large heterogeneous high throughput data sets into their functional interpretation. 19

These support a major learning objective

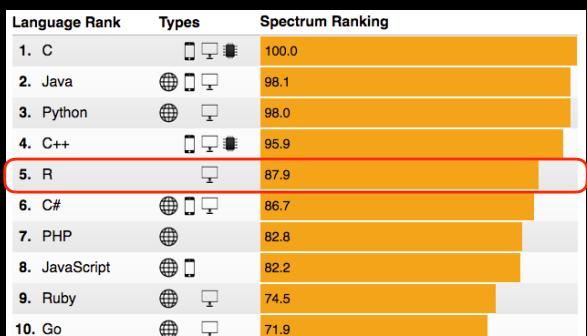
At the end of this course students will:

- 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?

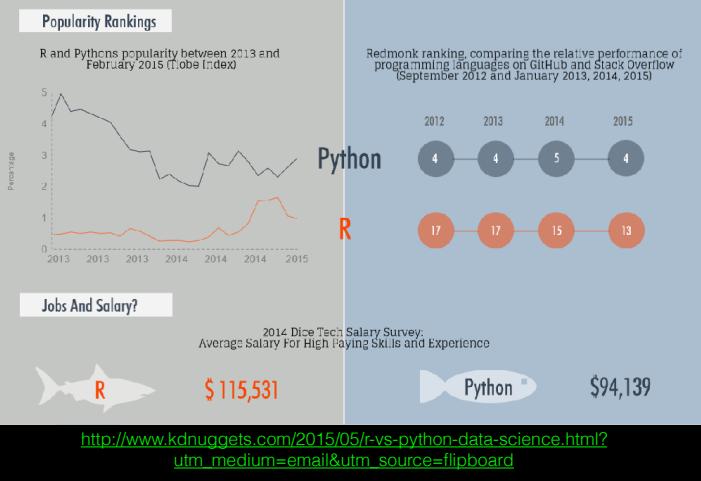
Productivity
Flexibility
Designed for data analysis

IEEE 2016 Top Programming Languages



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

R and Python: The Numbers



- 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 <i>what, why and 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

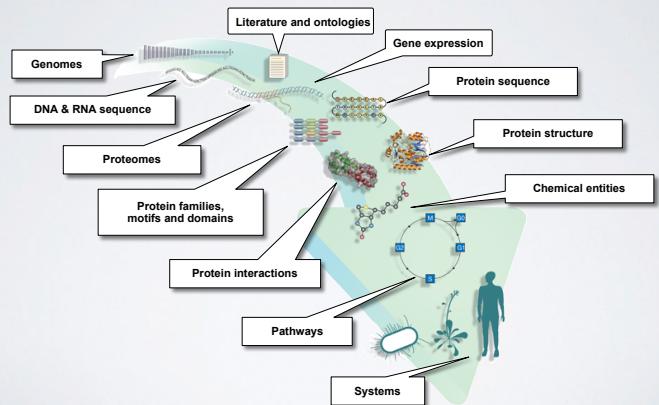
► "Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying "**informatics**" techniques (derived from disciplines such as applied maths, computer science, and statistics) to **understand** and **organize** the information associated with these molecules, on a **large-scale**. 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/l3gxrb6>)

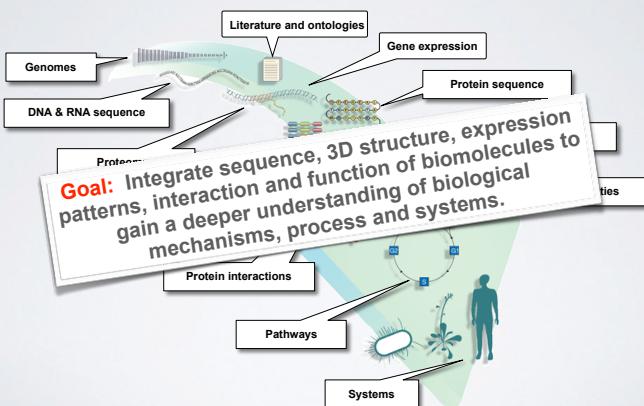
MORE DEFINITIONS

- “Bioinformatics is conceptualizing biology in terms of **macromolecules** and then applying “**informatics**” techniques (derived from disciplines such as applied mathematics, computer science, and statistics) to **understand** and **analyze** the information associated with these molecules, on a **large-scale**. Luscombe NM, et al. Methods 1999; 21:40-346.
- Key Point: Bioinformatics is Computer Aided Biology**
- “Bioinformatics is research, development, or application of computer 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>)

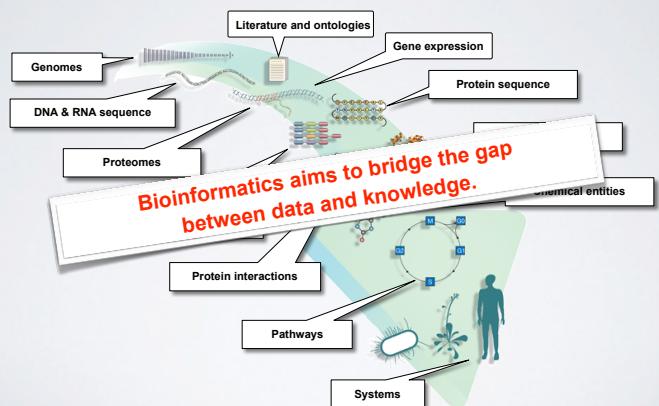
Major types of Bioinformatics Data



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

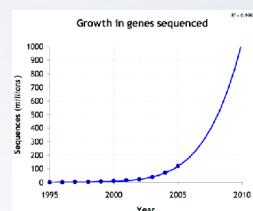
- DNA sequence determines protein sequence.
- Protein sequence determines protein structure.
- Protein structure determines protein function.
- Regulatory mechanisms (e.g. gene expression) determine the amount of a particular function in space and time.

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

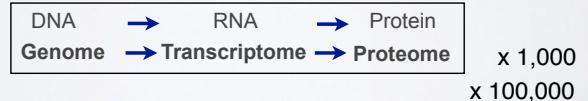
- Bioinformatics provides methods for the efficient:
 - **storage**
 - **annotation**
 - **search and retrieval**
 - **data integration**
 - **data mining and analysis**



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

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?

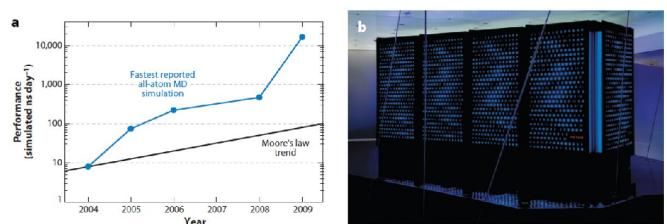
Pre-packaged tools and databases

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

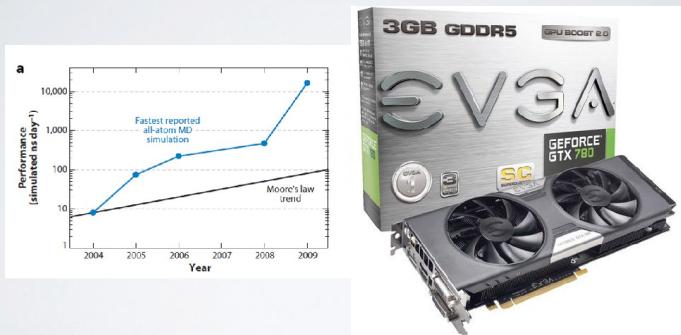
Advanced tool application & development

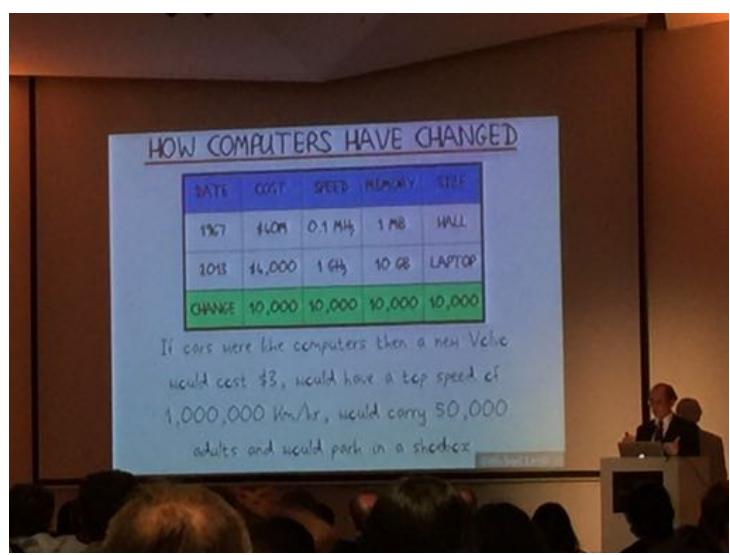
- Mostly on a UNIX environment
- Knowledge of programming languages frequently required (e.g. R, Python, Perl, C, Java, Fortran)
- May require specialized high performance computing...

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'!**

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)

Even Blast has many settable parameters

Related tools with different terminology

STEP 3 - Set your PROGRAM: FASTA

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score

Filters and Masking: Low complexity regions, Mask for lookup table only, Mask lower case letters

PSI/PHI/DELTA BLAST: Choose File, no file selected

PSI-BLAST: Threshold, Pseudocount

Upload PSSM: 0.005

Score Format: Default

Key Online Bioinformatics Resources: NCBI & EBI

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

Welcome to NCBI

NCBI is the National Center for Biotechnology Information, a major resource for biological information and tools for analyzing and understanding the genome.

Get Started

- Tools: Analyze data using NCBI software
- Databases: Search and browse our databases, including protein, nucleic acid, and metabolite databases.
- Resources: Find links to other resources, including journal articles, books, and software.

3D Structures

NCBI's 3D Structure database contains structures of proteins, nucleic acids, and small molecules. It includes crystallographic and NMR structures, and predicted conformations.

NCBI Assessments

NCBI's Assessments database contains annotations of gene products, including domain and motif predictions, and predicted secondary structure.

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

Welcome to EBI

The European Bioinformatics Institute (EBI) is a major resource for biological information and tools for analyzing and understanding the genome.

Get Started

• Tools: Analyze data using EBI software

• Databases: Search and browse our databases, including protein, nucleic acid, and metabolite databases.

• Resources: Find links to other resources, including journal articles, books, and software.

• Publications: Find links to publications in the EBI literature database.

• Events: Find links to events and meetings.

• News: Find links to news and updates.

• Help: Find links to help and support.

• Contact: Find links to contact information.

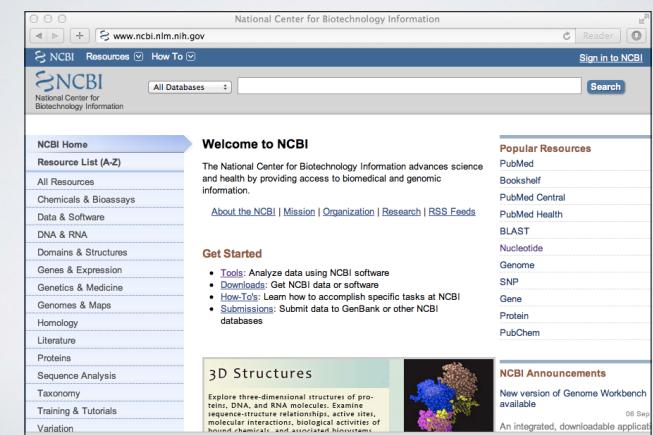
http://www.ebi.ac.uk

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

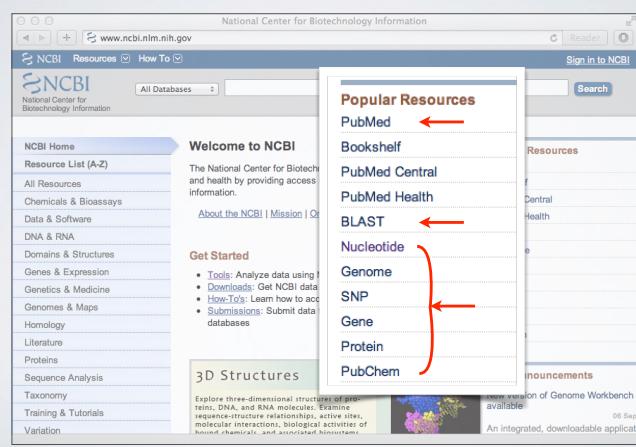


<http://www.ncbi.nlm.nih.gov>



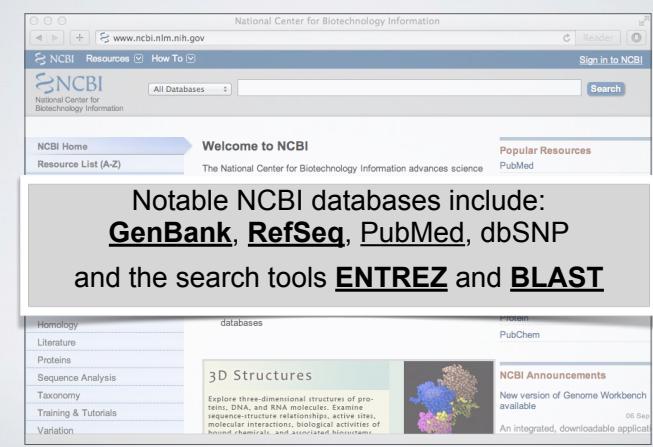
The screenshot shows the NCBI homepage with a navigation bar at the top. Below it is a sidebar with links to "Welcome to NCBI", "Popular Resources" (PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST), "Get Started" (with links to Tools, Downloads, How-Tos, Submissions), and "3D Structures". The main content area features a "3D Structures" section with a molecular visualization and a "NCBI Announcements" section.

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



This screenshot is similar to the one above but includes red arrows pointing to specific links in the "Popular Resources" section: PubMed, Bookshelf, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem.

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



This screenshot highlights the "Notable NCBI databases include: GenBank, RefSeq, PubMed, dbSNP" and "and the search tools ENTREZ and BLAST". It also shows the "3D Structures" and "NCBI Announcements" sections.

Key Online Bioinformatics Resources: NCBI & EBI

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



This screenshot shows the NCBI homepage with a "Welcome to NCBI" section and a "Popular Resources" sidebar.



This screenshot shows the EBI homepage with a "Welcome to EBI" section and a "Popular Resources" sidebar.

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

<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 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:

Services, Research, Training, Industry, European Coordination, EMBL ALUMNI, News from EMBL-EBI, News, Events, Newsletters, Mailing lists, Training, News, Contacts.

Visit EMBL.org

Upcoming events: Plant and Animal Genome conference (PAG XXVI), Sunday 10 - Tuesday 12 January 2016.

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*.

DNA & RNA, Gene expression, Proteins, Structures, Systems, Chemical biology, Ontologies, Literature, Cross domain.

Service news: EMBL-EBI, UniProt, PDB, ArrayExpress, ChEMBL.

Training

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*.

DNA & RNA, Gene expression, Proteins, Structures, Systems, Chemical biology, Ontologies, Literature, Cross domain.

Popular services: Ensembl, UniProt, PDBe, ArrayExpress, ChEMBL.

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
- InterPro
- PRIDE: The Proteomics Identification Database
- Pfam
- Clustal Omega
- HMMER – protein homology search
- InterProScan 5

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:

Services, Research, Training, Industry, European Coordination, EMBL ALUMNI, News from EMBL-EBI, News, Events, Newsletters, Mailing lists, Training, News, Contacts.

Visit EMBL.org

Upcoming events: Plant and Animal Genome conference (PAG XXVI), Sunday 10 - Tuesday 12 January 2016.

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

Train online

Using sequence similarity searching tools at EMBL-EBI: webinar

Course content: Using sequence similarity searching tools at EMBL-EBI.

Contributors: Andrew Cowley, using sequence similarity searching tools at EMBL-EBI.

Popular: Trainee, Find us...

Find us at...: Open days and career days, Conference invitations, EMBL courses and events, Extreme science events, Science 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**

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, 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, EC02DATABASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genllesne, 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, 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, 0-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS- MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc !!!

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, 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, DPLinteract, ECDC, ECGC, EC02DATABASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genllesne, 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, 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, 0-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D, SubtList, 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.

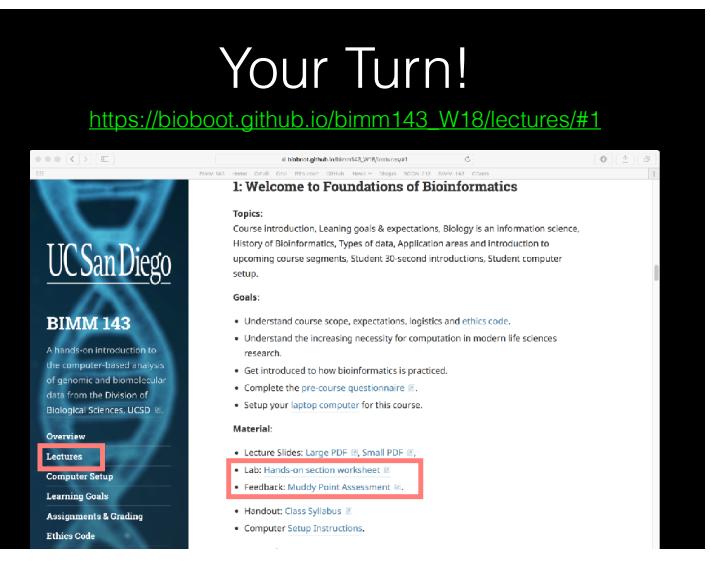
Primary, secondary & composite databases

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.

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.



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 | [10:45 am] |
| 2. GENE database @ NCBI | [11:00 am] |
| — BREAK — | — 11:10 am — |
| 3. UniProt & Muscle @ EBI | [11:35 am] |
| 4. PFAM, PDB & NGL | [12:05 pm] |
| — BREAK — | — 12:15 pm — |
| 5. Extension exercises | [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.