Slide 1

BIMM 143

Introduction to Bioinformatics

Barry Grant
UC San Diego

http://thegrantlab.org/bimm143

Slide 2

HELLO
my name is

BARRY

bjgrant@ucsd.edu

HELLO
~~my~~ HIS name is

ALEX

arsharp@ucsd.edu

Slide 3

Introduce Yourself!

Your preferred name,
Place you identify with,
Major area of study/research,
Favorite joke (optional)!

Slide 4

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 Bioinformatis	Introducing the *what*, *why* and *how* of bioinformatics?
Bioinformatics Database	**Hands-on** exploration of several major databases and their associated tools.

Slide 5

http://thegrantlab.org/bimm143/

Slide 6

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

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!

Course Structure
Derived from specific learning goals

Course Structure
Derived from specific learning goals

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

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

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

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

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 programing experience or high level math skills.

Provides a hook for increasing computational and data science competencies in the biosciences - valuable high demand translational skills!

Side Note: **Why stick with this course?**

Provides a hands-on practical introduction to major bioinformatics concepts and resources.

Covers modern hot topics and the intimate coupling of informatics with biology - highlighting the impact of computing advances and 'big data' on biology!

Designed for biology majors with no programing experience or high level math skills.

Provides a hook for increasing computational and data science competencies in the biosciences - valuable high demand translational skills!

BIMM-143 Learning Goals….
Data science R based learning goals

Slide 1

BIMM-143 Learning Goals….
Delve deeper into "real-world" bioinformatics

Slide 2

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.

Slide 3

Why use R?

Productivity
Flexibility
Designed for data analysis

Slide 4

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

Slide 5

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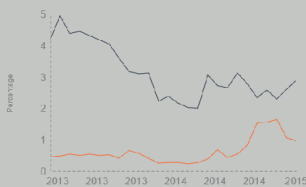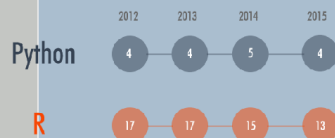
http://www.kdnuggets.com/2015/05/r-vs-python-data-science.html?utm_medium=email&utm_source=flipboard

Slide 6

- 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 leaning goals.
Introduction to Bioinformatis	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"** ~~niques~~ (derived from disciplines suc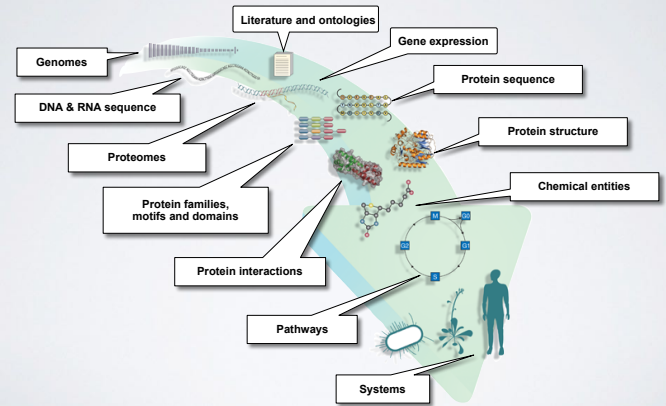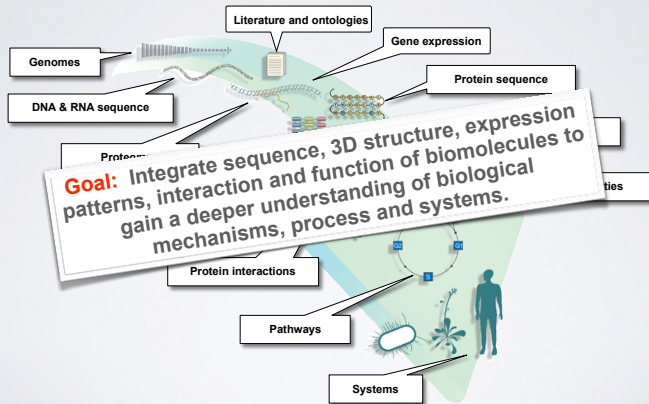h as applied math~~...~~ ~~uter~~ science, and statistics) to **understand** ~~a~~ ~~ze~~ the information associated with these ~~...~~ ~~e~~, on a **large-scale**. Luscombe NM, *et al.* Methods ~~...~~ 01;40:346.

Key Point: Bioinformatics is Computer Aided Biology

▸ "Bioinformati~~...~~ ~~arch, development, or application of~~ **compu**~~...~~ **approaches** for expanding the use of **bi**~~...~~ **medical**, **behavioral** or **health data**, including those to ~~a~~**quire**, **store**, **organize** and **analyze** such data." National Institutes of Health (NIH) (http://tinyurl.com/l3gxr6b)

Major types of Bioinformatics Data

Genomes • DNA & RNA sequence • Proteomes • Protein families, motifs and domains • Protein interactions • Pathways • Systems • Literature and ontologies • Gene expression • Protein sequence • Protein structure • Chemical entities

---

Major types of Bioinformatics Data

Goal: Integrate sequence, 3D structure, expression patterns, interaction and function of biomolecules to gain a deeper understanding of biological mechanisms, process and systems.

Major types of Bioinformatics Data

Bioinformatics aims to bridge the gap between data and knowledge.

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: <u>The key dogmas of molecular biology</u>
- *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 <u>*now*</u> 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**
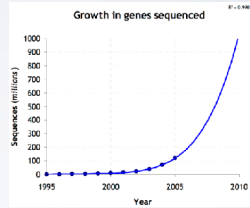 ‣ data **mining** and **analysis**

Growth in genes sequenced

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.

DNA	→	RNA	→	Protein
Genome	→	**Transcriptome**	→	**Proteome**

x 1,000
x 100,000

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 programing languages frequently required (*e.g.* **R**, Python, Perl, C, Java, Fortran)
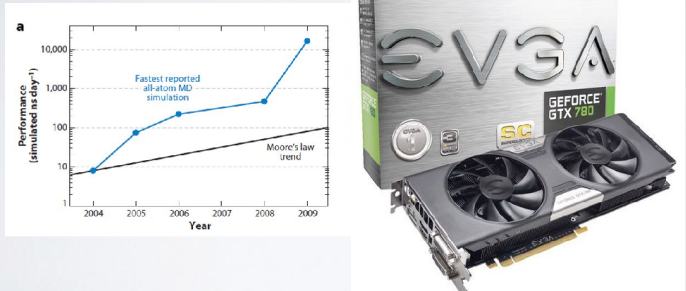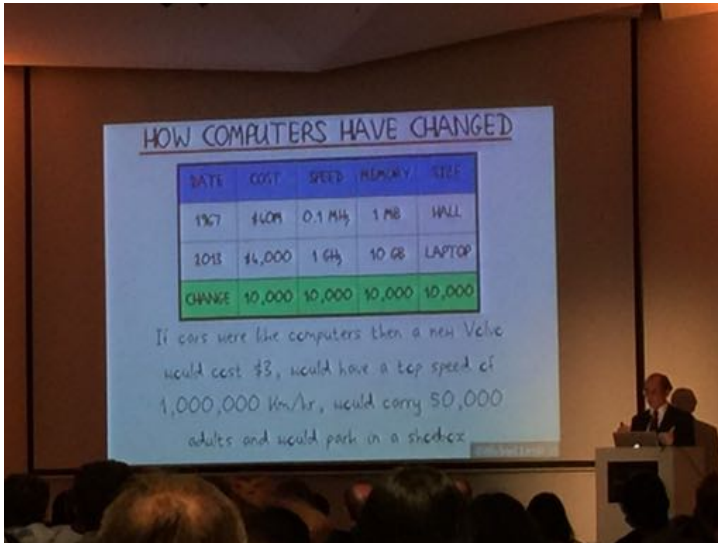- May require specialized high performance computing…

SIDE-NOTE: SUPERCOMPUTERS AND GPUS
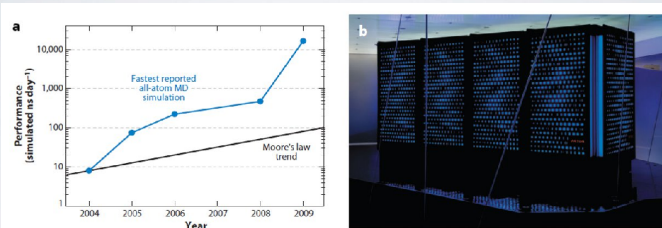
SIDE-NOTE: SUPERCOMPUTERS AND GPUS

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

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
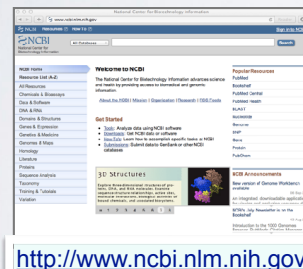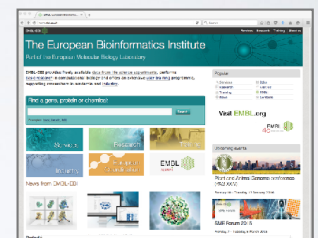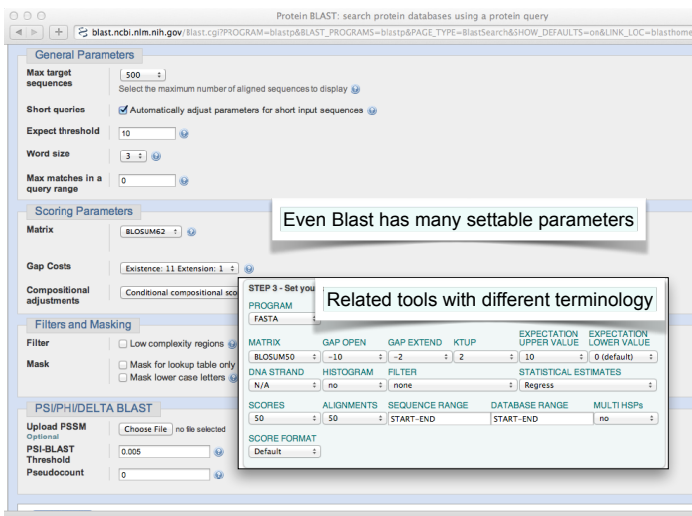
National Center for Biotechnology Information (NCBI)

- 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**
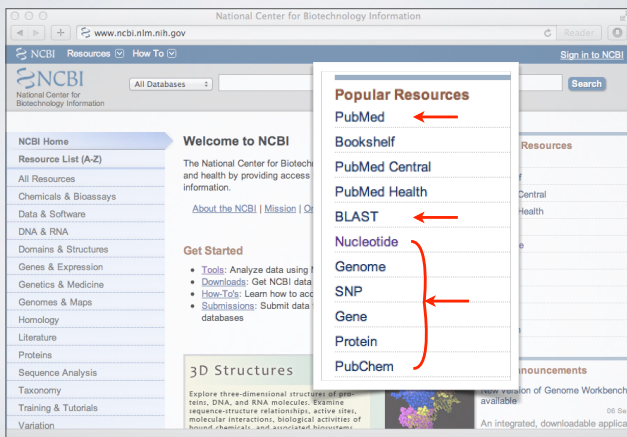 - Develop **software tools**
  - **Education** on and dissemination of biomedical information

  Bethesda, MD

- We will cover a number of core NCBI databases and software tools in the lecture

---

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

---

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

---

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

Notable NCBI databases include:
GenBank, **RefSeq**, PubMed, dbSNP
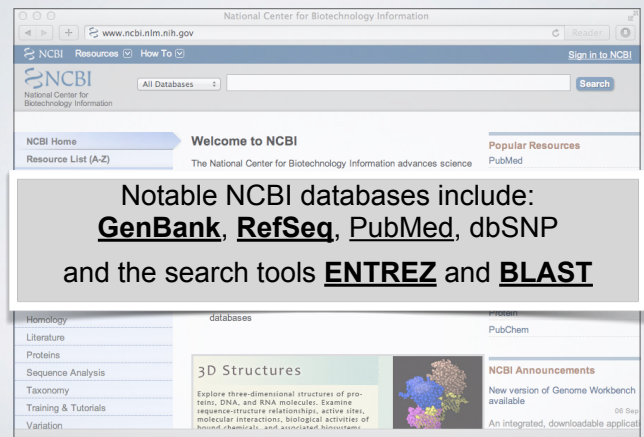and the search tools **ENTREZ** and **BLAST**

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
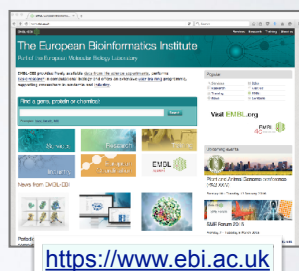
European Bioinformatics Institute (EBI)

- 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**

  Hinxton, UK

- We will briefly cover several EBI databases and tools that have advantages over those offered at NCBI

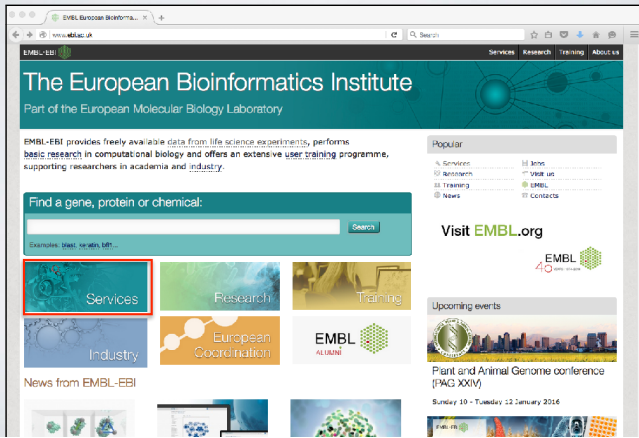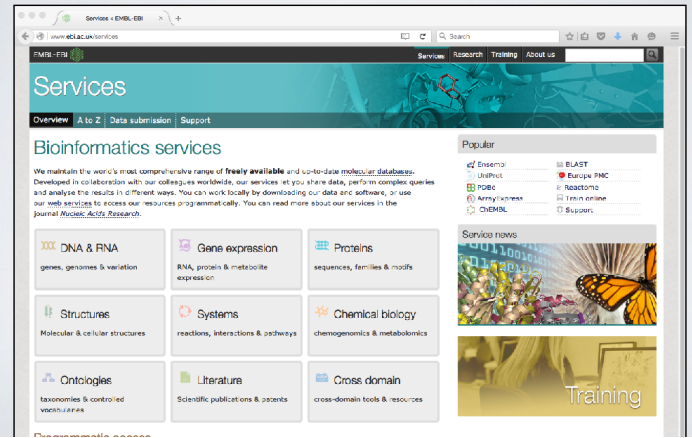The EBI maintains a number of high quality
curated **secondary databases** and associated tools

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
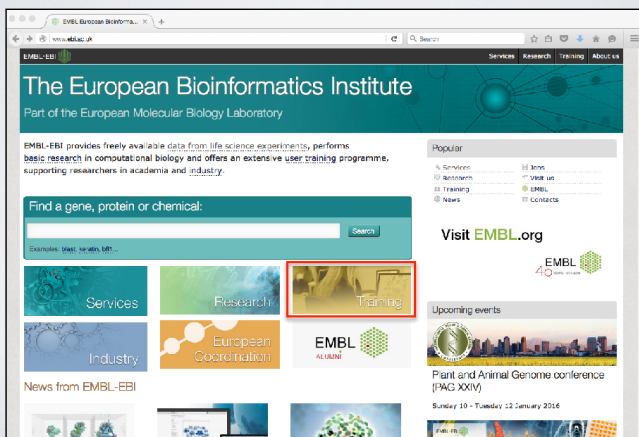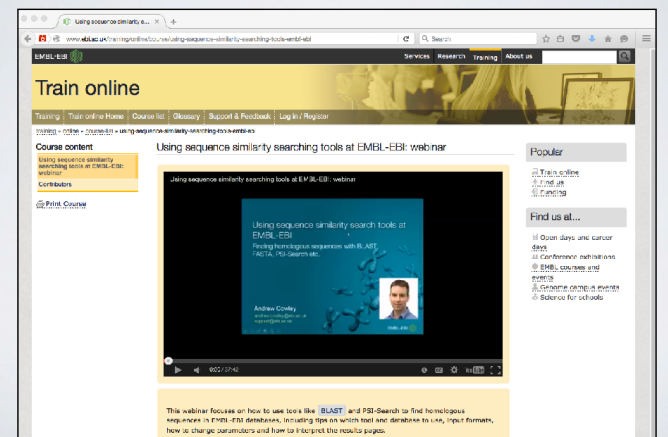

https://www.ebi.ac.uk
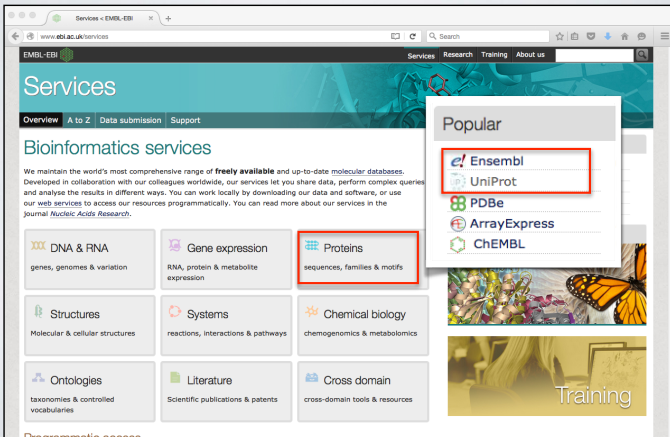
The EBI makes available a wider variety of **online tools** than NCBI
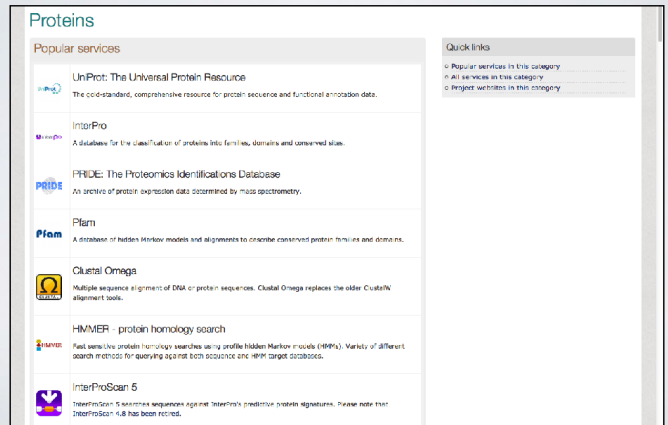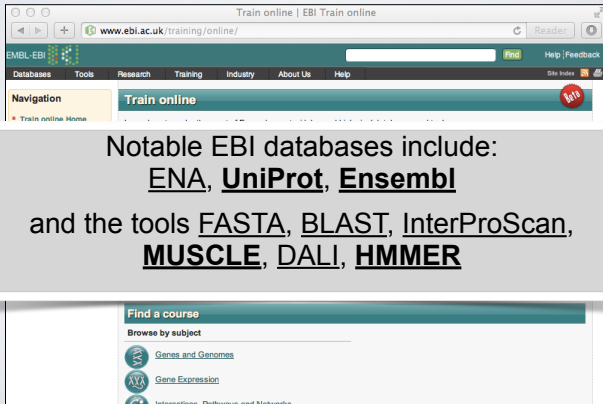

The EBI also provides a growing selection of **online tutorials**
on EBI databases and tools

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

Slide 1

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**

Slide 2

Next Class…

MAJOR BIOINFORMATICS DATABASES AND ASSOCIATED ONLINE TOOLS

Slide 3

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, BioImage, 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, ICTVDB, IL2RGbase, IMGT, Kabat, KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB, Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5 Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, MycDB, NDB, NRSub, 0-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB, PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD, PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE, PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE, SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase, SPAD, SRNA db, SRPDB, STACK, StyGene,Sub2D, SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS- MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB, TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE, VDRR, VectorDB, WDCM, WIT, WormPep, etc .... !!!!

Slide 4

Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, BioImage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOV... TKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY... AP, ChickGBASE, Colibri, COPE, CottonDB... bEST, dbSTS, DDBJ, DGP, DictyDb...
ECGC, EC02DBASE... THER, FlyBase... Link, ... HAEMB, ...vdb, HotMolecBase, ...RGbase, IMGT, Kabat, KDNA, ...Medline, Mendel, MEROPS, MGDB, MGI, MHC... OMAP, MJDB, MmtDB, Mol-R-Us, MPDB, MRR, MutBase, My... 0-lycBase, OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PD... 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 >

Slide 5

Side-note: Databases come in all shapes and sizes

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

Slide 6

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 Bioinformatis	Introducing the *what*, *why* and *how* of bioinformatics?
Bioinformatics Database	**Hands-on** exploration of several major databases and their associated tools.

Your Turn!

https://bioboot.github.io/bimm143_W18/lectures/#1

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 @ **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 am —
  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.