

**BGGN 213**  
**Foundations of Bioinformatics**  
 Barry Grant  
 UC San Diego  
<http://thegrantlab.org/bgg213>

## Recap From Last Time:

- Bioinformatics is computer aided biology.
  - 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 (see [handout!](#)).
- The **NCBI** and **EBI** are major online bioinformatics service providers.
- Introduced **Gene**, **UniProt** and **PDB** databases as well as a number of 'boutique' databases including **PFAM** and **OMIM**.
- Introduced the notion of *controlled vocabularies* and *ontologies* via exploring **GO** annotations.
- Also covered: Course structure; Introductions, Software setup and **Database Vignette...**

## Example Vignette Questions:

- What chromosome location and what genes are in the vicinity of a given query gene? **NCBI GENE**
- What can you find out about molecular functions, biological processes, and prominent cellular locations? **EBI GO**
- What amino acid positions in the protein are responsible for ligand binding? **EBI UniProt**
- What variants of this gene are associated with gastric cancer and other human diseases? **NCBI OMIM**
- Are high resolution protein structures available to examine the details of these mutations? How might we explain their potential molecular effects? **RCSB PDB**
- What is known about the protein family, its species distribution, number in humans and residue-wise conservation? **EBI PFAM**

## TODAYS MENU

- More hands-on exploration of these databases and their associated tools (searching with a propose!)
- Major hands-on sections include:
  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. Optional extension exercises [~20 mins]
- Please do answer the last review question (**Q19**).
- We encourage [discussion](#) and [exploration!](#)

## SideNote: Bioinformatics Databases

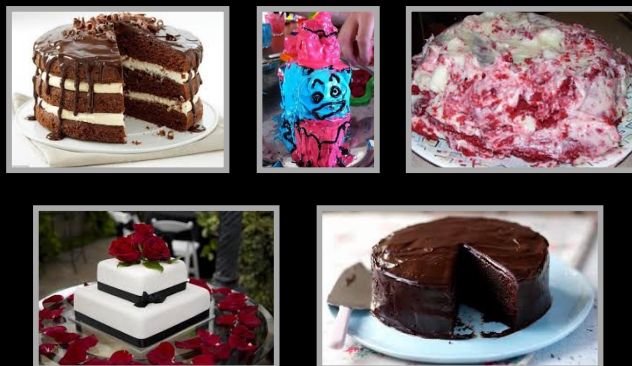
AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, Biolmage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP, ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG, CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb, Picty\_cDB, DIP, DOGS, DOMO, DPD, DPInteract, ECDC, ECGC, EC02DBASE, EcoCyc, EcoGene, EMBL, EMD db, ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView, GCRDB, GDB, GENATLAS, Genbank, GeneCards, Genilesne, 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, 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 .....!!!!

## SideNote: Bioinformatics Databases

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb, ARR, AsDb, BBDB, BCGD, Beanref, Biolmage, BioMagResBank, BIOMDB, BLOCKS, BovGBASE, BOVMAP, BSORF, BTKbase, CANSITE, CarbBank, CARBHYD, CATH, CAZY, 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, Genilesne, 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, 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.

## Finding Bioinformatics Databases

The screenshot shows the Oxford Journals website with the following content:

- Navigation: Home, Gmail, Gcal, Bitbucket, GitHub, News, Disqus, BGGN-213
- Logo: OXFORD ACADEMIC Journals
- Location: You are here: NAR Journal Home > Database Summary Paper Categories
- Section: NAR Database Summary Paper Category List
- Navigation links:
  - Compilation Paper
  - Category List
  - Alphabetical List
  - Category/Paper List
  - Search Summary Papers
- Database categories:
  - Nucleotide Sequence Databases
  - RNA sequence databases
  - Protein sequence databases
  - Structure Databases
  - Genomics Databases (non-vertebrate)
  - Metabolic and Signaling Pathways
  - Human and other Vertebrate Genomes
  - Human Genes and Diseases
    - CancerResource
    - Protein Mutant Database
    - General human genetics databases
    - General polymorphism databases
    - Cancer gene databases
    - Gene-, system- or disease-specific databases
  - Microarray Data and other Gene Expression Databases
  - Proteomics Resources
  - Other Molecular Biology Databases
  - Organelle databases
  - Plant databases
  - Immunological databases
  - Cell biology
- URL: <http://www.oxfordjournals.org/nar/database/c/>

# GENBANK & REFSEQ: NCBI'S NUCLEOTIDE SEQUENCE DATABASES

## What is GenBank?

- GenBank is NCBI's primary nucleotide only sequence database
  - ▶ Archival in nature - reflects the state of knowledge at time of submission
  - ▶ Subjective - reflects the submitter point of view
  - ▶ Redundant - can have many copies of the same nucleotide sequence
  - ▶ GenBank is actually three collaborating international databases from Europe, US and Japan



## GenBank sequence record

**Homo sapiens kinesin family member 5A (KIF5A), mRNA**  
NCBI Reference Sequence: NM\_004984.2

FASTA Graphics

Go to: [ ]

LOCUS NM\_004984 3897 bp mRNA linear PRI 10-JAN-2014

DEFINITION Homo sapiens kinesin family member 5A (KIF5A), mRNA.

**ACCESSION NM\_004984.2**

VERSION NM\_004984.2 GI:45446748

KEYWORDS RefSeq.

SOURCE Homo sapiens (human)

ORGANISM Homo sapiens  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 3897)  
AUTHORS Kawaguchi, K.

TITLE Role of kinesin-1 in the pathogenesis of SPC10. A rare form of her

JOURNAL Neu

PUBMED 221

REMARK Gen

spa

Rev

2

Pro

Doi

TITLE alpha-Synuclein oligomers impair neuronal microtubule-kinesin interplay

JOURNAL J. Biol. Chem. 288 (30), 21742-21754 (2013)

Peptide hormone metabolism

MHC class II antigen presentation

**GenBank flat file format has defined fields including unique identifiers such as the ACCESSION number.**

**This same general format is used for other sequence database records too.**

## Side node: Database accession numbers

Database **accession numbers** are strings of letters and numbers used as **identifying labels** for sequences and other data within databases

- ▶ Examples (all for retinol-binding protein, RBP4):

X02775 NT_030059	GenBank genomic DNA sequence Genomic contig	DNA
N91759.1 NM_006744	An expressed sequence tag (1 of 170) RefSeq DNA sequence (from a transcript)	RNA
NP_007635 AAC02945 Q28369 1KT7	RefSeq protein GenBank protein UniProtKB/SwissProt protein Protein Data Bank structure record	Protein
PMID: 12205585	PubMed IDs identify articles at NCBI/NIH	Literature

# GenBank sequence record

GenBank sequence record for Homo sapiens kinesin family member 5A (KIF5A), mRNA. The page displays the FASTA format sequence and associated metadata.

# GenBank sequence record

GenBank sequence record for Homo sapiens kinesin family member 5A (KIF5A), mRNA. A red arrow points to the 'FASTA' link in the top navigation bar, with a text box stating: "Can set different display formats here".

# FASTA sequence record

FASTA sequence record for Homo sapiens kinesin family member 5A (KIF5A), mRNA. The page displays the FASTA format sequence and associated metadata.

FASTA sequence files consist of records where each record begins with a ">" and header information on that same line. Each subsequent line of the record is sequence information.

This format is commonly used by sequence analysis programs.

# GenBank 'graphics' sequence record

GenBank 'graphics' sequence record for Homo sapiens kinesin family member 5A (KIF5A), mRNA. The page displays the graphical sequence and associated metadata.

FASTA sequence files consist of records where each record begins with a ">" and header information on that same line. Each subsequent line of the record is sequence information.

This format is commonly used by sequence analysis programs.

# GenBank sequence record, cont.

# GenBank sequence record, cont.

# GenBank sequence record, cont.

# RefSeq: NCBI's Derivative Sequence Database

- RefSeq entries are hand curated best representation of a transcript or protein (in their judgement)
- Non-redundant for a given species although alternate transcript forms will be included if there is good evidence

- Experimentally verified transcripts and proteins accession numbers begin with "NM\_" or "NP\_"
- Model transcripts and proteins based on bioinformatics predictions with little experimental support accession numbers begin with "XM\_" or "XP\_"
- RefSeq also contains contigs and chromosome records

# UNIPROT: THE PREMIER PROTEIN SEQUENCE DATABASE

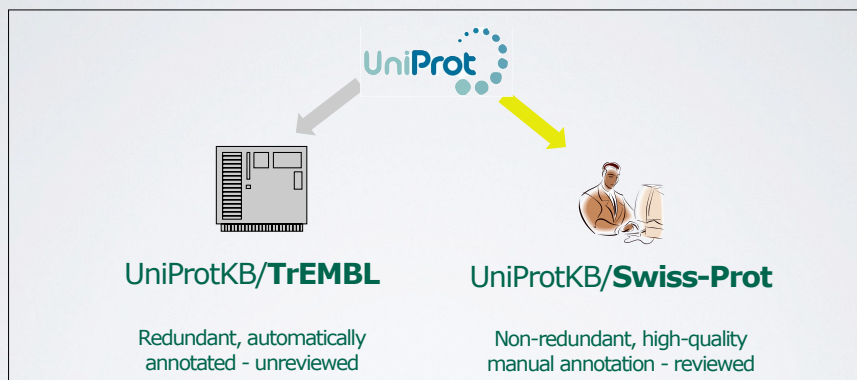
## UniProt: Protein sequence database

UniProt is a comprehensive, high-quality resource of protein sequence and functional information

• UniProt comprises four databases:

- 1. UniProtKB** (Knowledgebase)  
Containing **Swiss-Prot** and **TrEMBL** components (these correspond to hand curated and automatically annotated entries respectively)
- 2. UniRef** (Reference Clusters)  
Filtered version of UniProtKB at various levels of sequence identity  
e.g. **UniRef90** contains sequences with a maximum of 90% sequence identity to each other
- 3. UniParc** (Archive) with database cross-references to source.
- 4. UniMES** (Metagenomic and Environmental Sequences)

## The two sides of UniProtKB



★ Unreviewed, UniProtKB/TrEMBL **Q9N0H9** (Q9N0H9\_EQUAS)

★ Reviewed, UniProtKB/Swiss-Prot **P38398** (BRCA1\_HUMAN)

Indicators of which part of UniProt an entry belongs to include the color of the stars and the ID

## The main information added to a UniProt/Swiss-Prot entry

**Sequence**

**References**

**Literature Annotations**

**Ontologies**

**Protein names**

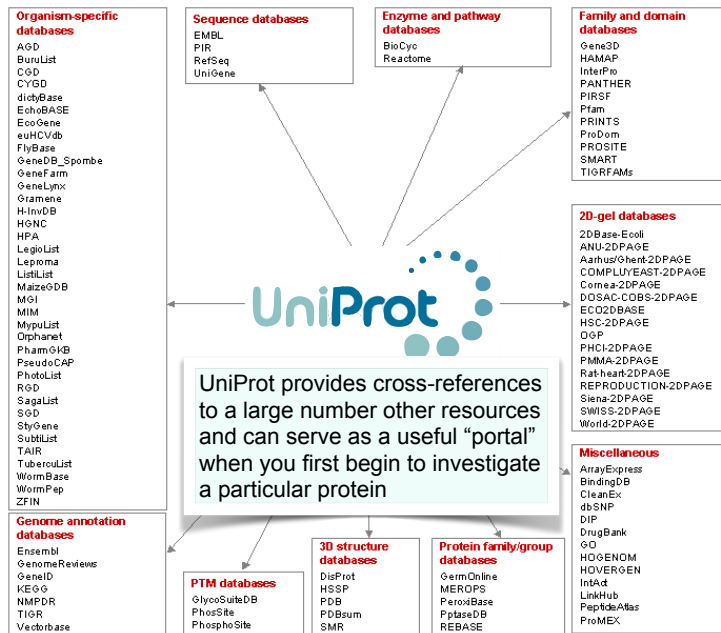
**Nomenclature**

**Gene names**

**Protein quaking**  
Also known as: Mpk1

**Sequence features**

**UniProt/Swiss-Prot**

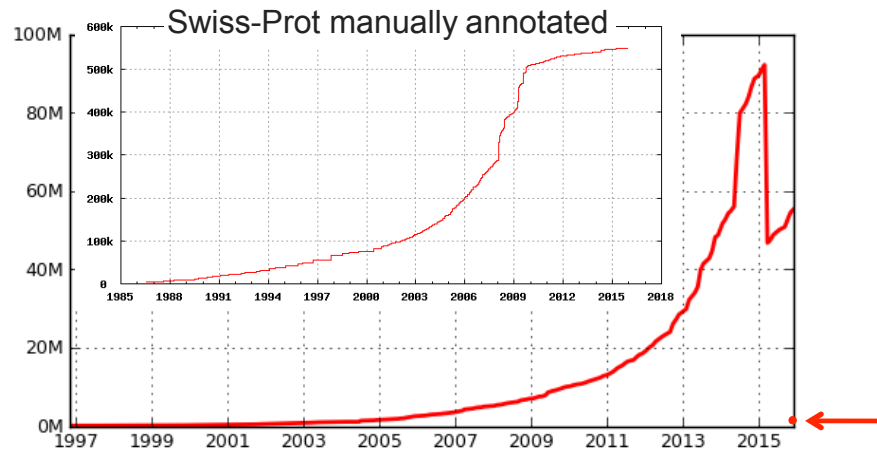


25

## UniProt/Swiss-Prot vs UniProt/TrEMBL

- UniProtKB/Swiss-Prot is a **non-redundant** database with one entry per protein
- UniProtKB/TrEMBL is a **redundant** database with one entry per translated ENA entry (ENA is the EBI's equivalent of GenBank)
  - ▶ Therefore TrEMBL can contain multiple entries for the same protein
  - ▶ Multiple UniProtKB/TrEMBL entries for the same protein can arise due to:
    - Erroneous gene model predictions
    - Sequence errors (Frame shifts)
    - Polymorphisms
    - Alternative start sites
    - Isoforms
    - OR because the same sequence was submitted by different people

## Side note: Automatic Annotation (a.k.a. sharing the wealth)



27

## Your Turn!

[https://bioboot.github.io/bggn213\\_f17/lectures/#2](https://bioboot.github.io/bggn213_f17/lectures/#2)

UC San Diego

**BGGN 213**

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

- Be able to describe how nucleotide and protein sequence and structure data are represented (FASTA, FASTQ, GenBank, UniProt, PDB).
- Be familiar with online tools at the EBI and NCBI including Muscle and BLAST.
- The goals of the hands-on session is to introduce a range of core bioinformatics databases and associated online services whilst actively investigating the molecular basis of several common human disease.

**Material:**

- Lecture Slides: Large PDF, Small PDF,
- Handout: Major Bioinformatics Databases
- **Hands-on section worksheet**
- Muddy point assessment

**Homework:**

BGGN-213: FOUNDATIONS OF BIOINFORMATICS (Lecture 2)

Bioinformatics Databases and Key Online Resources

[https://bioboot.github.io/bgg213\\_f17/lectures/#2](https://bioboot.github.io/bgg213_f17/lectures/#2)

Dr. Barry Grant  
Oct 2017

**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

```
ATGGTGCATCTGACTCCTGTGGGAAGTCTGCCCTTACTGCCCTGTGGGGCAAGTGAACGTGGATGAAG
TTGGTGGTGGAGCCCTGGGCAGGCTGCTGGTGGCTACCTTGGACCCAGAGGTTCTTTGAGTCTTTGG
GGATCTGTCCACTCCTGATCCAGTTATGGGCAACCTTAAGGTGAAGGCTCATGGCAAGAAAGTCTCGGT
GCTTTAGTATGGCTGGCTCACCTGGACAACTCAAGGGCACCTTTGCCACACTGAGTGGAGTGCACCT
GTGACAACTGGACGTGGATCTTGAACCTTCAGGCTCTGGGCAAGTGGCTGGTCTGTGGTGGCCCA
TCACCTTGGCAAGAAATTCACCCNCCAGTCCAGGCTGCCTATCGAAAGTGGTGGCTGGTGGCTAAT
GCCTGGCCCAAGTATCACTAAGCTCGCTTCTTGTGCTGCCAATTT
```

The only information you are given is the above sequence so you must begin your investigation with a sequence search - for this example we will use NCBI's **BLAST** service at: <http://blast.ncbi.nlm.nih.gov/>

Note that there are several different "basic BLAST" programs available at NCBI (including nucleotide BLAST, protein BLAST, and BLASTx).

## YOUR TURN!

- There are five major hands-on sections including:

1. BLAST, GenBank and OMIM @ **NCBI** [~35 mins]
2. GENE database @ **NCBI** [~15 mins]  
— BREAK —
3. UniProt & Muscle @ **EBI** [~25 mins]
4. PFAM, PDB & NGL [~30 mins]  
— BREAK —
5. Extension exercises [~30 mins]

- ▶ Please do answer the last review question (**Q19**).
- ▶ We encourage discussion and exploration!

## YOUR TURN!

- There are five major hands-on sections including:

- |  |                          |
|--|--------------------------|
| 1. BLAST, GenBank and OMIM @ <b>NCBI</b> | End times:<br>[ 9:45 am] |
| 2. GENE database @ <b>NCBI</b>           | [10:00 am]               |
| — BREAK —                                | — 10:10 am —             |
| 3. UniProt & Muscle @ <b>EBI</b>         | [10:35 am]               |
| 4. PFAM, PDB & NGL                       | [11:05 am]               |
| — BREAK —                                | — 11:15 am —             |
| 5. Extension exercises                   | [11:45 am]               |

- ▶ Please do answer the last review question (**Q19**).
- ▶ We encourage discussion and exploration!

## HOMEWORK

[https://bioboot.github.io/bgg213\\_f17/lectures/#2](https://bioboot.github.io/bgg213_f17/lectures/#2)

- Please do answer the last review question from today (**Q19**)
- Complete the **lecture 1 homework questions** for Thur.
- Check out the "Background Reading" material online.

THANK YOU