Biological databases

Lars Rønn Olsen
Assistant Professor, Technical University of Denmark
Learning objectives

After today, you will be able to:

• Distinguish between data, information, and knowledge
• Identify and use a selection of public, biological databases
• Leverage these databases for vaccine design
Data and information

**Data**: symbols with no added meaning and no significance beyond their existence. Example: unmapped reads

**Information**: meaning of the data, e.g. “what”, “where”, “when”, and “how many”. Example: genes, genomes (mapped reads, quality, origin)

**Knowledge**: justified belief of “how” the data and information relates to each other. Example: what do these genes do

**Wisdom**: when the acquired knowledge is applied to offer practical solutions. Example: leveraging knowledge for therapeutics, epidemiology, etc.

Data and information

Example: vaccine design workflow
- How to leverage databases and sequence analysis tools for vaccine design

Pathogen of interest: Dengue virus

General pathogen information:
- Wikipedia
- PubMed

Species information:
- NCBI Taxonomy

Genomic sequence data:
- NCBI GenBank

Gene information:
- GeneCards
- AmiGO

Gene expression profiles:
- NCBI GEO

Protein sequence data:
- NCBI protein
- SwissProt/UniProt

Selection of vaccine targets

Genomic sequence data:
- Whole genome sequencing
Example: vaccine design workflow
- How to leverage databases and sequence analysis tools for vaccine design

- Pathogen of interest: Dengue virus
  - Species information: NCBI Taxonomy

- Genomic sequence data: NCBI GenBank

- Gene information: GeneCards, AmiGO
  - Gene expression profiles: NCBI GEO
  - Protein sequence data: NCBI protein, SwissProt/UniProt

- Selection of vaccine targets

- Database
  - Tool
The free online encyclopedia Wikipedia is a GREAT way to start out when you need to look up information about a new gene, organism, tool, or technique.

HOWEVER: Wikipedia is NOT a reliable source of information, even if most entries are of a very good quality. Wikipedia entries have not been verified by experts and can potentially be wrong (everybody can go in and edit the text).

You CANNOT cite or quote Wikipedia as the source of your information – you will need to find the original primary source of the information or look it up in a scientific database.
Example: interested in a virus that you have not yet done any research on?

Get clued in on taxonomy, epidemiology, current treatments, genomic structure, etc.


When you know what to search for, then go to scientific databases.
Wikipedia
- How to use it, and how not to use it
NCBI is a collection of resources and databases from the US based National Center for Biotechnology Information. It is comprised of a number of useful resources, some of which we will introduce here today.


NCBI
- a collection of resources and databases

PubMed is a catalogue of articles published in medical and biotech research journals. Some are free (Open Access), some you have to pay to get access to.

HOWEVER: in 99% of the cases, you can still get the paper if you write the authors a nice email expressing interest in their work!


NCBI PubMed
- Biomedical literature from life science journals, and online books

NCBI PubMed
- Biomedical literature from life science journals, and online books
NCBI PubMed
- Biomedical literature from life science journals, and online books

Growth in PubMed

July 27, 2010

Twenty million papers in PubMed: a triumph or a tragedy?

A quick search on pubmed.gov today reveals that the freely available American database of biomedical literature has just passed the **20 million citations** mark*. Should we celebrate or commiserate passing this landmark figure? Is it a triumph or a tragedy that PubMed® is the size it is?
NCBI PubMed

- Biomedical literature from life science journals, and online books
NCBI PubMed
- Biomedical literature from life science journals, and online books

5-YEAR IMPACT FACTOR OF ISI JOURNAL PUBLICATIONS IN 2012 (1.2 MILLION)

Most findings are seldom cited and have low influence

“BIG LITERATURE”
You have to be specific when you search PubMed.

Searching PubMed is similar to searching with Google.

You can use logical operators such as “AND”, “NOT”, and “OR”.

For example: search for “(dengue virus) AND vaccine”
NCBI PubMed
- Biomedical literature from life science journals, and online books
Example: vaccine design workflow

- How to leverage databases and sequence analysis tools for vaccine design

Pathogen of interest:
Dengue virus

General pathogen information:
WikiPedia
PubMed

Species information:
NCBI Taxonomy

Genomic sequence data:
NCBI GenBank

Gene information:
GeneCards
AmiGO

Gene expression profiles:
NCBI GEO

Protein sequence data:
NCBI protein
SwissProt/UniProt

Selection of vaccine targets

Genomic sequence data:
Whole genome sequencing

Database
Tool
NCBI Taxonomy
- Curated classifications and nomenclatures for all organisms

Standardized, curated information about the taxonomy of all organisms

NCBI Taxonomy
- Curated classifications and nomenclatures for all organisms

Example: vaccine design workflow
- How to leverage databases and sequence analysis tools for vaccine design

Pathogen of interest: Dengue virus


Species information: NCBI Taxonomy

Genomic sequence data: NCBI GenBank

Gene information: GeneCards, AmiGO

Gene expression profiles: NCBI GEO

Protein sequence data: NCBI protein, SwissProt/UniProt

Selection of vaccine targets

Genomic sequence data: Whole genome sequencing

Database
Tool
NCBI GenBank
- An annotated collection of all publicly available genomic sequences

197,390,691 sequences in GenBank
NCBI GenBank

- An annotated collection of all publicly available DNA sequences

Example: vaccine design workflow

- How to leverage databases and sequence analysis tools for vaccine design

Pathogen of interest: Dengue virus


Species information: NCBI Taxonomy

Genomic sequence data: NCBI GenBank, Sequence Read Archive

Gene information: GeneCards, AmiGO

Gene expression profiles: NCBI GEO

Protein sequence data: NCBI protein, SwissProt/UniProt

Selection of vaccine targets

Database

Tool
Some information in GenBank, but mostly what the submitters add, and it is not always standardized, nor curated for quality.
Strain information
- Not all information is expert curated

COMMENT
This work was supported by the National Institute of Allergy and Infectious Diseases (NIAID), Genome Sequencing Centers for Infectious Diseases (GSCID) program.

FEATURES

Location/Qualifiers
1..1002
/organism="Influenza A virus (A/Peking duck/Heinersdorf/648-4/1986(H2N3))"
/mol_type="viral cRNA"
/strain="A/Peking duck/Heinersdorf/648-4/1986"
/serotype="H2N3"
/host="Peking duck"
/bio_material="CEIRS#150567#"
/db_xref="taxon:1166620"
/segment="7"
/lab_host="X/E1 passage(s)"
/country="Germany: Heinersdorf"
/collection_date="1986"

misc_feature
1..1002
/db_xref="IRD:NIGSP_CEIRS_SJC001_H2SJ_00050.MP"

gene
14..995
/gene="M2"

CDS
join(14..39,728..995)
/gene="M2"
Strain information
- Not all information is expert curated

“Peking duck”

misc_feature
/lab_host="X/E1 pas"
/country="Germany:"
/collection_date="1..1002"
/db_xref="IRD:NIGS"

gene
14..995
/gen="M2"
CDS
join(14..39,728..995)
/gen="M2"

“Pekin duck”
Strain information

- Not all information is expert curated

REFERENCE 2 (bases 1 to 11029)
Authors Deubel, V., Malkinson, M. and Banet, C.
Title Direct Submission
Journal Submitted (08-FEB-2002) CERVI, Institut Pasteur, 21 Avenue Tony Garnier, Lyon 69365, France
Features
Location/Qualifiers
source
1..11029
/origin="West Nile virus"
/mol_type="genomic RNA"
/strain="IS-98_STD1"
/host="sick stork"
/db_xref="taxon:11082"
/note="isolated in Israel in 1998"
CDS
/feature_id="AAPL87234.1"
/codon_start=1
/product="polyprotein precursor"
/protein_id="AAPL87234.1"
/translation="MSKKPGPQPGSRAVNLKRGMPRLVSLIGKRAMMSLIDGKGP:
RFVLLALLAFFRFTAIAIPRATVDRWGRGNQTTAMKHLLSFKKEALGTLTSAINRRSSK:
KKRGGKGTGIAVMIASVAVLSNFQGKVMTVNATDVTDVITIPATAAGKLCIVI:
AMDVGYMCDDTIYECPLASNGDPEDIDCWCTKSAVYVRYGRCTKTRHRSRRSRSL:"

Strain information
- Not all information is expert curated

FLUKB

FLUKB is a knowledge-based system that supports influenza vaccine discovery. It contains influenza sequence data, T-cell epitope, and B-cell epitope data, along with a set of tools for the analysis of targets of immune responses. Standardized nomenclature and dictionaries were developed and implemented to enable automated analysis queries and automation of updates. FLUKB was built using a modular framework allowing the implementation of analytical workflows. FLUKB supports the discovery of vaccine targets, analysis of viral diversity and of cross-reactivity and cross-neutralization of T-cells and neutralizing antibodies.

Search
- Keyword search
- BLAST search
- T-cell epitope search
- Neutralizing antibody list

Tools
- Sequence alignment
- Sequence variability analysis
- Block entropy calculation
- Block entropy calculation for T-cell epitope
- Strain mapper

Workflows
- Vaccine targets
- Cross-neutralization estimation
- B-cell epitope mapper

Crystal structure of Influenza hemagglutinin (H5) in complex with a broadly neutralizing antibody F10 (PDB ID: 3FKU)
Bonus: GeneCards
- Information about all known human genes

http://www.genecards.org/
Bonus: AmiGO
- Shorthand information about human gene function

http://amigo.geneontology.org/
Bonus: AmiGO
- Shorthand information about human gene function

http://amigo.geneontology.org/
Example: vaccine design workflow
- How to leverage databases and sequence analysis tools for vaccine design

Pathogen of interest: Dengue virus


Species information: NCBI Taxonomy

Genomic sequence data: NCBI GenBank, Sequence Read Archive

Gene information: GeneCards, AmiGO

Gene expression profiles: NCBI GEO

Protein sequence data: NCBI protein, SwissProt/UniProt

Selection of vaccine targets
Gene expression databases
- Expression of genes in different organisms

NCBI GEO (Gene Expression Omnibus) contains gene expression for a number of organisms - although primarily human.

It is, however, still relevant for vaccinology research!

Gene expression databases
- Expression of genes in different organisms

Gene expression databases
- Expression of genes in different organisms

Example: vaccine design workflow
- How to leverage databases and sequence analysis tools for vaccine design

Pathogen of interest: Dengue virus
Species information: NCBI Taxonomy
Genomic sequence data: NCBI GenBank, Sequence Read Archive
Gene information: GeneCards, AmiGO
Gene expression profiles: NCBI GEO
Protein sequence data: NCBI protein, SwissProt/UniProt
Selection of vaccine targets

Genomic sequence data: Whole genome sequencing

Database
Tool
Protein sequence databases
- NCBI has a resource for that, too!

Protein sequence databases
- UniProt/SwissProt – protein sequences and functional information

http://www.uniprot.org/
**Bonus: STRINGdb**

- Database with information about protein-protein interaction

http://string-db.org/
Bonus: Human Protein Atlas

- Database with protein expression in the human body

http://www.proteinatlas.org/
Want to explore further?

Have a look at Nucleic Acids Research’s database collection.

http://www.oxfordjournals.org/our_journals/nar/database/c/
Learning objectives

Now you should be able to:

• Distinguish between data, information, and knowledge
• Identify and use a selection of public, biological databases
• Find, explore, and download gene and protein sequences
Take home messages

• There are *a lot* of useful biological databases out there. We have not covered all of them today.

• If you are looking for a particular type of data, try searching on Google!

• If you want to explore databases, have a look at Nucleic Acids Research’s database collection (http://www.oxfordjournals.org/our_journals/nar/database/c/).

• Regarding articles: a lot of articles are not Open Access (you have to pay), but if you really want/need a paper, write the authors a nice email!