Basic Sequence Analysis

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

After today, you will be able to:

• Understand how BLAST works

• Use BLAST for sequence similarity search

• Understand the theory behind \textit{de novo} assembly of genomes from sequence reads

• Understand the theory behind short read alignment

• Understand how multiple sequence alignment works

• Use multiple sequence alignment to examine sequence variability

• Use public web services to predict vaccine targets in pathogens
Example: vaccine design workflow

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
BLAST
- Searching databases for sequences

BLAST (Basic Local Alignment Search Tool) is a tool to query a database for sequences similar to an input sequence.

Imagine you have sequenced a gene from an unknown sample, and you would like to know what it is. You can use BLAST in NCBI to compare your sequence to ALL the 197,390,691 sequences in GenBank!

BLAST
- Searching databases for sequences

Example: you would like to know what the following sequence is:

X) AATGCCG

You have the following three sequences in your database:

A) CGTGTGATC
B) AATGCCG
C) GCTGTGAC

BLAST
- Searching databases for sequences

Example: you would like to know what the following sequence is:

AATGCCG

You have the following three sequences in your database:

A) CGTGTGATC
B) AATCCCG
C) GCTGTGAC
X) AATGCCG

(Tip: use the font “courier new”)

Example: you would like to know what the following sequence is:

X) AATGCCCG

You have the following three sequences in your database:

A) CGTGTGATC
B) AATCCCG
C) GCTGTGAC

BLAST
- Searching databases for sequences

Example: you would like to know what the following sequence is:

X) AATGCCG

You have the following three sequences in your database:

A) AATCCCG
B) AATCCCG
C) AATCC

X) AATGCCG  X) AATG-CCG  X) AATGCCG
A) AATCCCG  B) AATCCCG  C) AAT-CC-

Which match is best?

BLAST
- Searching databases for sequences

Example: vaccine design workflow

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

Database

Tool

Genomic sequence data:
Whole genome sequencing
Sequencing read alignment
- Tools for next generation sequencing

Sequencing without reference:
  De novo assembly

Sequencing with reference:
  Short read alignment

Gene or genome sequence

Base calls

Quality control/trimming

Raw sequencing output

Database
  Tool

Sequence read archive

GenBank/RefSeq
De novo assembly

- Making sense of sequencing reads without a reference gene or genome

ATGACGTTT
AAATTCCCC
TTTCTGAAA
AAATTCCCC
CCCCTGGCC

A) ATGACGTTT
B) AAATTCCCC
C) TTTCTGAAA
D) CCCCTGGCC

ATGACGTTTCTGAAATTCCCCCTGGCC
De novo assembly
- Making sense of sequencing reads without a reference gene or genome

A) ATGACGCCC
B) CCCTTCCCC
C) CCCCTGCC
D) CCCCTGGCC

ATGACGCCC CCCCTGCC
CCCCTGCCC CCCCTGGCC
ATGACGCCCTGCCCTCCCCTGGCC

ATGACGCCC CCCCTGCCC
CCCCTTCCCC CCCCTGGCC
ATGACGTTTCTCCCTCCCTGGCC
**De novo assembly**
- Making sense of sequencing reads without a reference gene or genome

B) CCC TTCC

A) ATGACC CCC

D) CCCCTGGCC

E) TTCCCTGGCC

C) CCCCTGCCC

???
Sequencing read alignment
- Tools for next generation sequencing

Sequencing without reference:
*De novo* assembly

Sequencing with reference:
Short read alignment

Gene or genome sequence

Raw sequencing output

Base calls

Quality control/trimming

Database

Tool

Sequence read archive

GeneBank/RefSeq
Short read alignment
- Aligning reads to a reference gene or genome

<table>
<thead>
<tr>
<th>Reference Reads</th>
<th>ATGACGTCAGCTGT TTGGCGACATCG TCCGATCAGT CGATTTA TT CGATAA TCGCTC TCTTTAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATGACG</td>
<td>TTGGCG</td>
</tr>
<tr>
<td>CTTGCG</td>
<td>AGTCGA</td>
</tr>
<tr>
<td>TCTCGG</td>
<td>TCTCGG</td>
</tr>
<tr>
<td>TGTGCC</td>
<td>TGTGCC</td>
</tr>
<tr>
<td>CTGACT</td>
<td>CTGACT</td>
</tr>
<tr>
<td>TCTTTT</td>
<td>TCTTTT</td>
</tr>
</tbody>
</table>

Read Depth “x”
Make your sequences available
- Other researchers can benefit immensely from your work!

Your sequence reads can be deposited in the Sequence Read Archive.
Make your sequences available
- Other researchers can benefit immensely from your work!

Your assembled/mapped sequences can be deposited in GenBank.

---

GenBank Overview

What is GenBank?
GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research, 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the ftp site. The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for previous GenBank releases are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release.

An annotated sample GenBank record for a Saccharomyces cerevisiae gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank
There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with Entrez Nucleotide, which is divided into three divisions: CoreNucleotide (the main collection), dbEST (Expressed Sequence Tags), and dbGSS (Genome Survey Sequences).
- Search and align GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently; see BLAST info for more information about the numerous BLAST databases.
- Search, link, and download sequences programatically using NCBI e-utils.
Example: vaccine design workflow

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

Database
Tool
Predicting T cell epitopes
- Tools for vaccine target discovery

Hypothesis:

1. Protein sequences
2. Sequence variability analysis
   Multiple sequence alignment
3. Prediction of epitopes
4. Selection of epitopes
5. Epitopes for vaccine

Immune Epitope Database

Database
Tool
Multiple sequence alignment
- Aligning sequences to determine variability

Multiple sequence alignment is a type of algorithm that allows you to compare 2 or more sequences simultaneously.

This is highly useful, for example when analyzing the variability of a certain protein in a pathogen.

There are many different algorithms for this purpose – one of the most famous being ClustalW from 1997.

In fact, this tool has been cited 53,288 times (number 10 in the paper mountain) and is still cited heavily.

HOWEVER! The main author of ClustalW, Des Higgins, has asked people to stop using and citing it as there are upgrades and other better tools available today!
Multiple sequence alignment
- Aligning sequences to determine variability

How does it work? Very similar to BLAST, except all sequences are considered:

AATCCCGA  AATCCC-GA
AATCCCCGT  AATCCCCGT
AATCCT  AATCC---T
Multiple sequence alignment
  - Aligning sequences to determine variability

In practice, you copy or download all your sequences of interest in fasta format.

The fasta format looks like this:

> This line is the header. You can write whatever you want here
ATCAGACTGTGCTGATCG...

You then paste or upload the sequences to a multiple sequence alignment web server
(or install it locally if you have very large datasets) and run the alignment. One is
webPRANK which is available through EBI (European Bioinformatics Institute).

http://www.ebi.ac.uk/goldman-srv/webprank/
Multiple sequence alignment
- Aligning sequences to determine variability

http://www.ebi.ac.uk/goldman-srv/webprank/
Multiple sequence alignment
- Aligning sequences to determine variability

http://www.ebi.ac.uk/goldman-srv/webprank/
Predicting T cell epitopes
- Tools for vaccine target discovery
Prediction of T cell epitopes

- What are T cell epitopes?

Virus proteins are cleaved to short peptides

Some of these peptides bind to the human leukocyte antigen protein (HLA)

When virus peptides are presented on the surface of the infected cell, T cells kill the infected cells

The HLA protein come in different flavors
Finding T cell epitopes
- Traditional approach to finding immunogenic regions in pathogens
Prediction of T cell epitopes
- Computational approach to finding immunogenic regions in pathogens

Prediction of T cell epitopes
- How to predict peptide binding to HLA

There are a lot of algorithms to predict epitopes in pathogens and cancer cells.

Among the best performing is NetMHC.
Prediction of T cell epitopes
- How to predict peptide binding to HLA
Prediction of T cell epitopes
- How to predict peptide binding to HLA

Epitopes
- MRCVGVGNR
- RCVGVGNRD
- CVGVGNRDF
- VGVGNRDFV
- GVGNRDFVE
- VGNRDFVEG
- GNRDFVEGL
- NRDFVEGLS
- DFVEGLSGA
- FVEGLSGAT
- VEGLSGATW
- EGLSGATWV
- GLSGATWVD

Not epitopes
- MRCVGVGNR
- RCVGVGNRD
- CVGVGNRDF
- VGVGNRDFV
- GVGNRDFVE
- VGNRDFVEG
- GNRDFVEGL
- NRDFVEGLS
- DFVEGLSGA
- FVEGLSGAT
- VEGLSGATW
- EGLSGATWV
- GLSGATWVD

http://www.cbs.dtu.dk/services/NetMHC/
Prediction of T cell epitopes
- How to predict peptide binding to HLA

MRCVGVGNRFVEGLSGATWVDVVLFQCLESIEGKAVQHENLKYTVIITVHTGDQHQVG

Potential epitopes
GVGNRDFVE
VGNRDFVEG
FVEGLSGAT
VEGLSGATW

http://www.cbs.dtu.dk/services/NetMHC/
Predicting T cell epitopes
- Tools for vaccine target discovery

- Protein sequences
- Sequence variability analysis
  - Multiple sequence alignment
- Prediction of epitopes
- Selection of epitopes
- Epitopes for vaccine
- Immune Epitope Database
- HLA allele
  - HLA allele frequency

Database
Tool
HLA databases - Databases with information about the human leukocyte antigen

HLA comes in different flavors (alleles).

Different alleles bind different peptides.

Humans have six different HLA alleles (three from each parent).

Different HLA alleles are prevalent in different populations – see the HLA allele frequency database

http://www.allelefreqencies.net/

You can also explore the sequences of the HLAs in the HLA allele database

https://www.ebi.ac.uk/ipd/imgt/hla/
HLA databases
- Databases with information about the human leukocyte antigen

http://www.allelefrequencies.net/
**HLA databases**

- Databases with information about the human leukocyte antigen

http://www.allelefrequencies.net/
Predicting T cell epitopes
- Tools for vaccine target discovery

Protein sequences

Sequence variability analysis
Multiple sequence alignment

Prediction of epitopes

Selection of epitopes

Epitopes for vaccine

HLA allele
HLA allele frequency

Database
Tool

Immune Epitope Database
Selection of epitopes
- Choosing HLA binders

Combining variability analysis (multiple sequence alignment) and HLA binding predictions lets you pick the best epitopes for your vaccine.

This can be done manually or using, for example, the BlockCons tool.
Selection of epitopes
- Choosing HLA binders

http://met-hilab.cbs.dtu.dk/blockcons/
Selection of epitopes
- Choosing HLA binders
Selection of epitopes
- Choosing HLA binders

After you have selected your potential epitopes, you should check whether they are present in human proteins, as vaccination with these may lead to lack of efficacy or potentially autoimmunity.

How would you do this?

Answer: use BLAST in GenBank, and search for the peptides in human proteins.
Predicting T cell epitopes
- Tools for vaccine target discovery

- Protein sequences
- Sequence variability analysis
- Multiple sequence alignment
- Prediction of epitopes
- Selection of epitopes
- Epitopes for vaccine

- HLA allele
- HLA allele frequency

- Immune Epitope Database

- Database
- Tool
The Immune Epitope Database contains epitopes that other researchers have reported in the literature.

Use it to see if anyone has worked experimentally with your predicted peptides before. If someone has already tested whether it gives rise to an immune response, you can use this to inform your decision to use them or not.
Immune Epitope Database (IEDB)
- Databases with epitopes in an array of pathogen species

http://www.iedb.org/
Prediction of T cell epitopes
- Computational approach to finding immunogenic regions in pathogens

16 years of reverse vaccinology
- What have we achieved?

1,779 reports of novel or improved prediction algorithms

4,622 reports of novel vaccine targets

142 active clinical trials of various forms of DNA vaccines

2 approved DNA vaccines
...protecting horses against West Nile virus
and dogs against melanoma.
There are a lot of tools out there

Use Google to search for tools for your specific questions.

There are also online for a where you can ask questions if you cannot find the answer. For example BioStars:

https://www.biostars.org/

You can also explore tools in the tools registry at:

https://bio.tools/
Take home messages

• Exploring and re-analyzing published data is incredibly useful - when you sequence something, contribute to the research community and upload you raw and processed sequence data!