# DIY Bioinformatics: BLAST & friends

Richard Tillett
NV INBRE Bioinformatics Core
Feb 24, 2017

## Today's goal

- My goal today is to help you answer one question…

## "What am I going to do with this!?"


## We'll make some sense of it (starting with BLAST)!

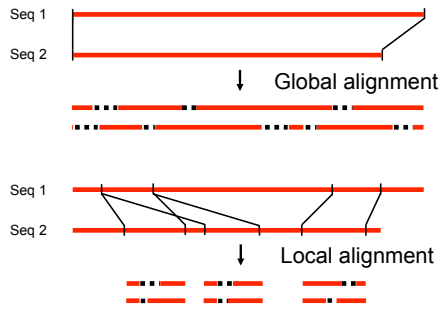

## How will we do make sense of it?

- **<u>Using the Basic Local Alignment Search Tool (BLAST)</u>**
-
- <u>What is BLAST?</u>
- BLAST is a set of tools used to identify imperfect matches (similarity) between a query nucleotide/protein sequence with sequences stored in a database. For a given query sequence, BLAST reports sequences with aligned regions of similarity.

## BLAST

- Basic Local Alignment Search Tool
- based on Smith-Waterman (SW) local alignment
  - SW alignment employs a type of math called dynamic programming
- BLAST uses a heuristic (shortcut) called <u>word method</u>

## Global vs Local Alignment




*Tao Tao, NLM, NIH, Bethesda, MD*

## BLAST basics

- BLAST needs two things:
  - A query
    - Nucleotide or protein sequence
    - Typically in FASTA format text
  - A database to search
    - On web-based blasts, this will be a dropdown menu

## FASTA

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITN
LFSAIPYIGTNLVEWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSN
NPLGLTSDSDKIPFHPYYTIKDFLGLLILILLLLLALLSPDMLGDPDNHMPADPLNTPLHI
KPEWYFLFAYAILRSVPNKLGGVLALFLSIVILGLMPFLHTSKHRSMMLRPLSQALFWTL
TMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGXIENY
>SEQUENCE_2
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAAKKADRLA
AEGLVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRLKDPNKP
EHKIPQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNIIPGKMNSFIADNSQLDS
KLTLMGQFYVMDDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKTEDFAAEVAA
QL
>SEQUENCE_3
SATVSEINSETDFVAKNDQFIALTKDTTAHIQSNSLQSVEELHSSTINGVKFEEYLKS
QIATIGENLVVRRFATLKAGANGVVNGYIHTNGRVGVVIAAACDSAEVASKSRDLLRQ
ICMH
```

- A very basic specification consisting of:
  - >Definition line
  - One or more lines of sequence
  - Nucleotide OR protein
  - File named as .fasta, .fa, .fna, .faa, .txt

Schlauch BCH709

## Why use BLAST?

"What is my sequence and what does it do?"
- You obtained DNA sequence from an experiment and need to know more about it
- Comparing the novel sequence to known protein sequences to hypothesize gene function
- Comparing a partial sequence to whole genome sequences

- Finding homologs to interesting genes in your favorite species
- Specialized BLASTs
  - Make gene-specific primers for PCR (Primer-BLAST)
  - Screen a sequence for vectors (VecScreen)
  - Comparing two sequences (bl2seq)

## Types of BLASTing

- BLASTN: DNA query vs DNA database
- BLASTP: protein query vs protein database
- BLASTX: DNA query translated into all six reading frames to produce translated protein sequences, which are the used to query a protein seq database
- TBLASTN: protein query vs a DNA sequence database with seqs translated into all 6 reading frames
- TBLASTX: DNA query sequence translated into all 6 reading frames against DNA seq db translated into all 6 reading frames

Schlauch BCH709

## Database Similarity Searching

general goal

• submit a sequence of interest (**query**)

• find most similar sequences from XX databases

• the most similar sequences ***may*** have the most similar function to query sequence

Schlauch BCH709

## E-Values

rules of thumb

$E < 10^{-50}$ → high confidence of homologous relationships

$10^{-50} < E < .01$ → result of homology

$.01 < E < 10$ → no significance, perhaps remote homology

$E > 10$ → two sequences are randomly related

Nota Bene: as E is directly dependent on the length of the database, as the database(s) increase in length, E will increase, which likely will yield less hits



## How BLAST Works

1. Make lookup table of "words" for query
2. Scan database for hits (word match/seed)
3. Extend alignment both directions
   - Ungapped extensions of hits (initial HSPs)
   - Gapped extensions (no traceback)
   - Gapped extensions (traceback + alignment details)



## Nucleotide Words

Make a lookup table based on the word size.

11-mer

ATGCTGCTAGTCGATGACGTAGCTACCGATAT

ATGCTGCTAGT
ATGCTGCTAGTCGATGACGTAGCTACC
TGCTGCTAGTC
GCTGCTAGTCG
CTGCTAGTCGA
TGCTAGTCGATGACGTAGCTACCGATA
. . .



## Word Hits & Extensions

Nucleotide: one exact match

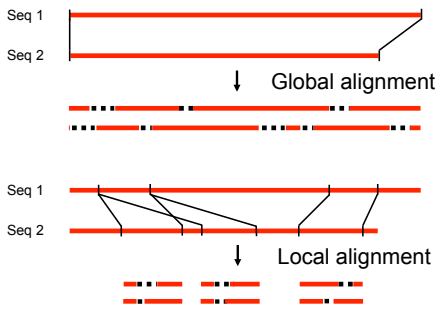
ATGCTGCTAGTCGATGACGTAGCTA
⇦ TAGTCGATGA ⇨

Protein: two matches within 40 residues

PHAIEKCYTGCTLAQEADDTA
⇦ IDK ⇨ ⇦ EAD ⇨



## Global vs Local Alignment






## BLAST Parameters

- Max target sequences: how many possible alignments you want to BLAST to return
- Automatically adjust parameters for short sequences: defaults: BLOSUM62, for short sequences, PAM30 shorter sequences = 50 bases,
- Expect threshold: how many matches from the database should be due to chance alone?
- Word size: the length of the word that the BLAST algorithm first uses between the query and database sequences (and then extends left and right) defaults: 3 or 6 AA's; 11 bases, smaller values are more sensitive

## BLAST Parameters

- Max matches in a query range: you can limit the number of matches to a particular part of a sequence
- Match/Mismatch Scores: select from pull-down menu
- Gap Costs: choose from specifics or use a linear function
- Filter: you can filter out low-complexity regions or known repeat regions based on organism
- Mask: you can highlight parts of your sequence to be filtered by the filter above



## BLAST Parameters

- before aligning, you may want to exclude regions of low complexity ~ regions containing short repeats
- they make up ~15% of all protein sequence data
- may artificially increase the similarity score
- alignment software allows you to mask these regions
  - hard masking: excludes the regions completely
  - soft masking: excludes the regions in first step but uses them for the extension step



## Test drive time!

Go to
https://github.com/rltillett/diy-blast/

& click "diy_blast.md"