"Sequence comparison, tricks of the trade"

Computational Metagenomics Workshop, University of Mauritius, 2014

Christos A. Ouzounis (CERTH) and Vasilis J. Promponas (UCY)

1. Dot plots for visual exploration

Use dotlet at SIB: http://myhits.isb-sib.ch/cgi-bin/dotlet on the following sequences

a. Synthetic sequence

Create one or use the examples below:

Indicative example (a true sequence).

```
>gi|398025370|ref|XP_003865846.1| hypothetical protein[Leishmania donovani]
MLATLTDITNVQASLPQPPKGKIAAANVAHVSPATVRQSVAQHNPYKWQCLPAGTYSAAEMMADMEDACT
LPVQELFQEKEDRAPPPPGGPPVCNTLVQFKCHQAEYVSSMVVDKGQYVVVQGDRGIDIGVVIRINTNSQ
KTYVERTGPVGSILRHATQREVDYWATDLKDAEAMAAAFCQAKARSHGLPMEIAYAEYQFDKKKLTFYYE
ARTRIDFVQLLKELYREYGCRIWMEKLRNHD
```

And an artificial derivative.

```
>gi|GGGG
MLATLTDITNVQASLPQPPKGKIAAANVAHVSPATVRQSVAQHNPYKWQCLPAGTYSAAEMMADMEDACT
LPVQELFQEKEDRAPPPPGGPPVCNTLGGGGGGGGGGGGGGGGGGGGGGVQFKCHQAEYVSSMVVDKG
QYVVVQGDRGIDIGVVIRINTNSQKTYVERTGPVGSILRHATQREVDYWATDLKDAEAMAAAFCQAKARS
HGGGGGGGGGGGGGGGGGGGLPMEIAYAEYQFDKKKLTFYYEARTRIDFVQLLKELYREYGCRIWMEKL
RNHD
```

b. Real sequence (Retrieve from public databases)

E.g., use http://www.ncbi.nlm.nih.gov/protein, or another database of your choice. If you are missing protein names use as keywords WD40, HEAT or ARM repeats, Collagen.

c. Homopolymeric run (create one)

For example, on the command line run (with appropriate arguments)

```
$perl –e 'for(1..$ARGV[0]){$a.=$ARGV[1]} print $a, "\n"'
```

Important note: the first student to get to obtain a valid homopolymeric run using this perl one liner wins a "tea" session at Flying Dodo tonight.

d. Real sequence data

Use one of the files huntingtin.fas, obscurin.fas in the archive.

e. Pairwise alignments

Use data from 1d to perform heuristic pairwise local comparison with Blast2seq http://blast.ncbi.nlm.nih.gov/ and exact dynamic programming-based pairwise alignment with the Smith-Waterman algorithm implementation at https://www.ebi.ac.uk/Tools/psa/.

Note: you should perform all comparisons using the exact same parameters (Matrix: BLOSUM62, Gap Open Penalty: 10, Gap Extension Penalty: 1). Then compare the results and discuss.

2. Compositional bias and sequence database search

a. BLAST a homopolymeric sequence vs NR (http://blast.ncbi.nlm.nih.gov/)

Pick a different homopolymer each (assigned alphabetically in class).

Note: Observe hits/patterns and discuss with your "neighbors".

b. Select one of the top hits identified in 2a and perform a BLAST search against the NR database

Record your results.
Note: if you retrieved no results borrow an ID from your neighbor.

c. Perform compositional bias detection and masking using CAST (http://athina.biol.uoa.gr/CAST/) for the same query with 2b

Discuss the results.

Try varying substitution matrices for this run and comment on the differences.

d. Use the filtered sequence in step 2c to repeat the search in 2b after turning all BLAST filters off

Compare your results with 2b.

3. Perform Multiple Sequence Alignment

Use CLUSTALO (https://www.ebi.ac.uk/Tools/msa/clustalo/) on ten hits selected from the results in 2d.

Optionally: Visualize using JalView (http://www.jalview.org/)

4. Some Interesting Links (further reading)

a. Repeat detection tools

Repeats DB: http://repeatsdb.bio.unipd.it/

Tuebingen Sequence Analysis Tools (includes HHpredID): http://toolkit.tuebingen.mpg.de/sections/seqanal

LRR finder: http://www.lrrfinder.com

Repro: http://www.ibi.vu.nl/programs/reprowww/

TRUST: http://www.ibi.vu.nl/programs/trustwww/

ProRepeat: http://prorepeat.bioinformatics.nl:443/dev/f?p=131:1:853145912003765

Radar: http://www.ebi.ac.uk/Tools/pfa/radar/

Internal Repeats Finder (www non-functional, source available): http://nihserver.mbi.ucla.edu/Repeats/

REPETITA: http://protein.bio.unipd.it/repetita/

b. Low complexity - Compositional Bias Detection

XNU source: http://blast.advbiocomp.com/pub/xnu/

SEG source: ftp://ftp.ncbi.nih.gov/pub/seg/

CAST: http://athina.biol.uoa.gr/CAST/ (soon available at http://troodos.biol.ucy.ac.cy)

LCR-eXXXplorer server: http://repeat.biol.ucy.ac.cy/mgb2/gbrowse/swissprot/

SubSeqer: http://www.compsysbio.org/subseqer/index.php

SIMPLE: http://www.biochem.ucl.ac.uk/bsm/SIMPLE/

GBA: http://bioinformatics.cise.ufl.edu/GBA/GBA.htm