Laboratory 7: BLAST and e-PCR

This week you will locate sequences using BLAST (Altschul et al. 1990, 1997) and re-PCR (Schuler 1997). These tools have very different approaches to finding things: BLAST attempts to match the entire query sequence to a set of references and returns the best matching reference portions whereas re-PCR simulates PCR by finding priming sequences in the correct orientation and range from one another and returns the coordinates of the priming sites. Although re-PCR was designed as a PCR simulation tool (e.g. to test primers *in silico*), it can be very useful for finding and extracting sequences from poorly curated sources.

Tasks

(1) Fully setup BLAST and create a database from GenBank.

(a) Read the on–line BLAST manual (http://www.ncbi.nlm.nih.gov/books/NBK1763/).

(b) Download the taxonomy name lookup table from NCBI by typing `wget ftp://ftp.ncbi.nlm.nih.gov/blast/db/taxdb.tar.gz` in the terminal.

(c) Decompress the taxonomy table by typing `tar xvzf taxdb.tar.gz` in the terminal.

(d) Download the taxonomy by accession lookup table from NCBI by typing `wget https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/nucl_gb.accession2taxid.gz` in the terminal.

(e) Search and download non–angiosperm vacular plant plastid sequences by typing `esearch -db nuccore -query 'refseq[filter] AND "complete genome" AND plastid[filter] AND Tracheophyta NOT Magnoliophyta' | efetch -format fasta > sequences.fasta` in the terminal.

(f) Confirm that 'sequences.fasta' has 479 sequences using grep. Answer question (1).

(g) Create a taxonomy mapping table for the downloaded sequences.

(1) First, install the bloom utility by typing `sudo apt install golang-github-dcso-bloom-cli` in the terminal. Enter your password if/when prompted.

(2) Next, create an empty compressed Bloom (1970) filter by typing `bloom -gz create -p 0.0000001 -n $(grep -c '^>' sequences.fasta) accessions.bloom.gz` in the terminal. Answer question (2).

(3) Then, add the GenBank accessions from the downloaded sequences to the bloom filter by typing `grep '^>' sequences.fasta | awk '{print $1}' | tr -d '>' | bloom -gz insert accessions.bloom.gz` in the terminal.

(4) Finally, extract the corresponding NCBI taxonomy identifiers by typing `gzip -cdk nucl_gb.accession2taxid.gz | bloom -gz -d $'\t' -f 1 -s check accessions.bloom.gz | awk -F'\t' 'BEGIN{OFS="\t"}{print $2,$3}' > accession2taxon` in the terminal. Answer question (3).

(5) The 'nucl_gb.accession2taxid.gz' file may now be deleted if disk space is at a premium.

(h) Format a BLAST database by typing `makeblastdb -dbtype nucl -in sequences.fasta -input_type fasta -parse_seqids -taxid_map accession2taxon -hash_index -out plastid -blastdb_version 5` in the terminal. Eleven files should have been created. Answer question (4).

(2) Extract a *Cycas taitungensis ycf1* query sequence from the BLAST database by typing `blastdbcmd -db plastid -dbtype nucl -entry NC_009618.1 -strand minus -range 129861-135188 > query.fasta` in the terminal.

(3) Carry out a nucleotide BLAST search with default settings by typing `blastn -query query.fasta -task blastn -db plastid -outfmt '6 sscinames sseqid evalue bitscore score length pident qstart qend sstart send sseq' -num_threads $(nproc) -max_target_seqs 5000 -out default-nbn.txt` in the terminal. Answer question (5).

(4) Have a look at the output ('default-nbn.txt') using the text editor of your choice. Count the number of unique accessions hit by typing `awk -F'\t' '{print $2}' default-nbn.txt | sort -u | wc -l` in the terminal. Answer question (6).

(5) Install the 'datamash' descriptive statistics tool by typing `sudo apt install datamash` in the terminal. Provide your password when prompted.

(6) Calculate the median and Inter–Quartile Range (IQR) of sequence length by typing `awk -F'\t' '{print $6}' default-nbn.txt | datamash q1 1 median 1 q3 1 iqr 1` in the terminal. Enter the output in the table below and answer question (7).

(7) The default settings are clearly not working perfectly—many of the hits are not full length and all of the reference data should contain a *ycf1* sequence. Try switching to another search strategy by typing `blastn -query query.fasta -task dc-megablast -db plastid -outfmt '6 sscinames sseqid evalue bitscore score length pident qstart qend sstart send sseq' -num_threads $(nproc) -max_target_seqs 5000 -out default-nbm.txt` in the terminal. Calculate the median and IQR of sequence length with datamash and enter the output in the table below. Answer question (8).

(8) Alternatively, try searching using amino acid translation by typing `tblastx -query_gencode 11 -db_gencode 11 -query query.fasta -db plastid -outfmt '6 sscinames sseqid evalue bitscore score length pident qstart qend sstart send sseq' -num_threads $(nproc) -max_target_seqs 5000 -out default-tbx.txt` in the terminal. When calculating the discriptive statistics, be sure to multiply the aminio acid sequence length by 3 to make it comparable to the nucleotide sequence length.

(9) The scoring matrix (default == BLOSUM62) can be used to change the search results. Using the base query from task (8) try varying the scoring matrix using the '-matrix' flag. Fill in the table below. Answer question (9).

query	program	BLOSUM	sequences	quartile 1	median	quartile 3	IQR
DNA	blastn	—					
DNA	dc-megablast	—					
AA	tblastx	45					
AA	tblastx	50					
AA	tblastx	62					
AA	tblastx	80					
AA	tblastx	90					

(10) Install and setup e-PCR.

(a) Type `sudo apt install ncbi-epcr` in the terminal. Provide your password when prompted.

(b) Read the re-PCR man page (type `man re-PCR` in the terminal).

(c) Format the e-PCR database by typing `famap -b plastid.mmap -t N sequences.fasta` in the terminal.

(d) Create the e-PCR hash by typing `fahash -b plastid.hash -w 3 -f 2 plastid.mmap` in the terminal (settings follow Little 2014). Answer question (10).

(e) Create an e-PCR primer file by typing `echo -e 'rbcL1/rbcLA\tTTGGCAGCATTYCGAGTAACTCC \tCCTTTRTAACGATCAAGRC' > rbcL-primer.txt` in the terminal. This primer set should bind to an internal portion of many *rbcL* sequences Palmieri et al. (2009).

(11) Start re-PCR by typing `re-PCR -S plastid.hash -n 5 -g 0 -d 100-300 -o rbcL5.rePCR rbcL-primer.txt` in the terminal. Answer question (11).

(12) Rerun the re-PCR search using '-n 9'. Be sure not to overwrite 'rbcL5.rePCR'. Answer question (12).

(13) Extract the re-PCR identified sequences from 'rbcL5.rePCR' by typing `grep -v '^#' rbcL5.rePCR | perl -F'\t' -lane '{$F[2]=~s/\-/minus/;$F[2]=~s/\+/plus/;print("blastdbcmd -db plastid -dbtype nucl -entry ".$F[1]." -strand ".$F[2]." -range ".$F[3]."-".$F[4])}' | bash > rbcL5.rePCR.fasta` in the terminal. Answer question (13).

(14) Add second primer set Hofreiter et al. (2000); Poinar et al. (1998) to 'rbcL-primer.txt' by typing `echo -e 'Z1aF/h2aR\tATGTCACCACCAACAGAGACTAAAGC\tCGTCCTTTGTAACGATCAAG' >> rbcL-primer.txt` in the terminal.

(15) Search using both primer sets by typing `re-PCR -S plastid.hash -n 5 -g 0 -d 100-300 -o rbcL2.rePCR rbcL-primer.txt` in the terminal. Answer question (14).

Questions (https://forms.gle/gjt5PhDayb1FfxY89)

(1) For task (1)(f):

(a) How many species are represented by the downloaded sequences?

(b) What command string did you use to determine this?

(2) For task (1)(g)(2), what does each of the bloom options do?

(3) For task (1)(g)(4), what does each of the bloom options do?

(4) For task (1)(h), what does each of the makeblastdb options do?

(5) For task (3), what does each of the blastn options do?

(6) For task (4), how many unique sequences (not accessions) were found?

(7) For task (6):

(a) What is the median and IQR of the sequence length?

(b) Are these numbers larger or smaller than you would expect?

(c) Why?

(8) Does task (3) or (7) produce more complete results? Why?

(9) Which BLAST search works better (i.e. gets more, longer sequences)? Defend your answer using the BLAST output.

(10) For task (10)(d):

  (a) What does the '-w' option do?
  (b) How would changing the -w value change your search results?
  (c) What does the '-f' option do?
  (d) How would changing the -f value change your search results?

(11) For task, (11):

  (a) What does each of the re-PCR options do?
  (b) Are the first five sequences retrieved consistent with the GenBank *rbcL* annotations (hint: use the re-PCR output to find the genome location in the complete GenBank record)?

(12) For task (12):

  (a) Do the additional sequences retrieved appear to be *rbcL* sequences?
  (b) When you BLAST the retrieved sequences, are appropriate hits located?

(13) For task (13):

  (a) What does each step in the command string do?
  (b) Do the extracted sequences appear to be *rcbL* sequences?
  (c) How can you tell?
  (d) Does re-PCR offer a viable method of search?
  (e) When would you use it in place of BLAST?

(14) For task (15):

  (a) Which primer amplifies better?
  (b) How did you determine this?
  (c) Are the conditions simulated with re-PCR realistic?
  (d) Why or why not?

Literature cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers & D. J. Lipman. 1990. Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller & D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389–3402.

Bloom, B. H. 1970. Space/time trade–offs in hash coding with allowable errors. Communications of the ACM 13: 422–426.

Hofreiter, M., H. N. Poinar, W. G. Spaulding, K. Bauer, P. S. Martin, G. Possnert & S. Pääbo. 2000. A molecular analysis of ground sloth diet through the last glaciation. Molecular Ecology 9: 1975–1984.

Little, D. P. 2014. A DNA mini–barcode for land plants. Molecular Ecology Resources 14: 437–446.

Palmieri, L., E. Bozza & L. Giongo. 2009. Soft fruit traceability in food matrices using real–time PCR. Nutrients 1: 316–328.

Poinar, H. N., M. Hofreiter, W. G. Spaulding, P. S. Martin, B. A. Stankiewicz, H. Bland, R. P. Evershed, G. Possnert & S. Pääbo. 1998. Molecular coproscopy: dung and diet of the extinct ground sloth *Nothrotheriops shastensis*. Science 281: 402–406.

Schuler, G. D. 1997. Sequence mapping by electronic PCR. Genome Research 7: 541–550.

Due at the start of class March 14.