Assignment 2: Sequence Comparison

Due date: Wednesday, 2/3 10am

Part 1

Go to http://www.yeastgenome.org/. Search for the gene *Rap1*. Get the nucleotide sequence for the Rap1 coding region, and go to http://www.ncbi.nlm.nih.gov/. Find **BLAST**. Choose translated query vs. protein database (**BLASTX**). BLASTX will translate the sequence into a peptide and blast it against a protein database.

Paste the *Rap1* sequence in the space provided (in FASTA format) and select the "nr" database.

Click on **BLAST** to submit the job. How much time does it take to finish the job?

When you get the output, answer these questions:

Question 1
Why did you want to use the "nr" database as opposed to any of the other database options?

Question 2
How many hits did you get with an e-value < 1?

Question 3
In what species is the closest non-Saccharomyces relative? In that species, what is the score and % identity for Rap1's closest relative?

Next, run BLASTX with the BLOSUM80 scoring matrix instead of the default BLOSUM62 matrix. (The scoring matrix parameter can be changed by clicking on the Algorithm parameters links.)

Question 4
Did you get more or fewer hits than before? Why?

Question 5
Now what species is the closest non-Saccharomyces relative?

Question 6
Find the protein (*Rap1*) that *was* the closest relative according to BLOSUM62. What is the new % identity and score for *Rap1* when using BLOSUM80?

Now, BLAST again with BLOSUM62, but lower the gap existence penalty to 7.

Question 7
How many hits did you get? Why do you think this number changed the way it did?

Question 8
Why did the score of the closest ortholog change the way it did?

Question 9
If you lowered the word length, would you expect the search to take more or less time? Why?

Question 10
Isn't online BLAST really slow?

Hopefully, you have now realized two things. First, when we say "closest relative," the answer really depends on the scoring matrix and parameters we use. Second, using BLAST online is really slow.

Part 2

You are provided with sequencing reads from an enrichment sequencing experiment. Map the reads back to chr22 and identify what is enriched.

Use the chr22 fasta file from assignment 1 to create an index. Please copy the file “cp ~/assignment1/work/hs_ref_GRCh38.p2_chr22.fa ~/assignment2/work” and rename the file by using “mv ~/assignment2/work/hs_ref_GRCh38.p2_chr22.fa ~/assignment2/work/chr22.fa”.

Once created, use `reads.fq` to align the reads to chr22. This file is provided in `/home/assignments/assignment2/`

Using **bowtie2**, build an index for chr22. Bowtie2 is installed on the server, so just enter:

```
$ bowtie2-build <path to chr22.fa> <index filename prefix (minus trailing .X.bt2): eg. chr22_idx>
```

`bowtie2-build` will create various files under the name specified with different file extensions.

To learn about the various options available in bowtie to map sequencing reads, enter

```
$ bowtie2
```

Once you've built the index for chr22, align the reads to chr22 and output the reads that map uniquely.

```
$ bowtie2 <write your options here. Note that the reads in reads.fq are unpaired> 2> <report file>
```

The report file contains an alignment summary. The alignment summary is printed to the standard error stream. The bowtie2 manual talks about this in more detail:

> *When Bowtie 2 finishes running, it prints messages summarizing what happened. These messages are printed to the "standard error" ("stderr") filehandle. For datasets consisting of unpaired reads, the summary might look like this:*

20000 reads; of these:
20000 (100.00%) were unpaired; of these:
1247 (6.24%) aligned 0 times
18739 (93.69%) aligned exactly 1 time
14 (0.07%) aligned >1 times
93.77% overall alignment rate

Question 11
Report the command you used to align the reads. How many reads map uniquely to chr22? How many reads map to multiple locations? How many reads were unmappable?

Place your output file and your report file in the submissions folder.

Question 12
Using the script "nuc_count_FINAL.py", identify the frequency of each nucleodtide and di-nucleotide. What is enriched in this dataset? (Hint: Look at the relative enrichment of single and dinucleotides) Report the enrichment of all single and di-nucleotides. Describe how you calculated the enrichment. Mention what the dataset is enriched for and interpret it. What assay do you think the data came from?

Extra Credit

Use BLAST locally on the server

Did you know that BLAST is installed in our server? Now you can use the command line to run BLAST jobs! For extra credit, obtain the "unknown.fsa" file from Assignment2 folder and use BLASTn command to identify what species the sequences are from.

Question EC.1
What species is the the sequence from? Why do we use BLASTn instead of BLASTx?

What to turn in

- Output files
  - `Alignment file`
  - `Report file`
- Your `README.txt` with the answers to the questions and the commands you used to answer the questions
- Extra credit only
  - BLAST output file