Assignment 2 Solutions

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

The nr database is “non-redundant”. This is advantageous for a number of reasons, almost all of which is due to the fact that this is a smaller database file than if you had just blindly combined Swiss-PROT, GenBank, and the model organism databases. The smaller size is good for computational reasons as BLAST can run faster. It’s good for statistical reasons because the smaller database means that you would expect to see fewer hits with a high score. It’s also good for biological reasons because in many cases two labs have sequenced the same gene and given it two different names. The nr database will hopefully prevent confusion caused by this.

2. How many hits did you get with an e-value < 1?
MTS 50: 69 hits
MTS 100: 124 hits
MTS 250/500/20000 : 150 hits

There has been some variability in the hits that are E values <1.
There seems to be a bug in BLAST when it comes to changing "Max Target Sequence" values.

http://blastedbio.blogspot.com/2015/12/blast-max-target-sequences-bug.html

"Hello,

Thank you for the report. We don't consider this a bug, but I agree that we should document this possibility better. This can happen because limits, including max target sequences, are applied in an early ungapped phase of the algorithm, as well as later. In some cases a final HSP will improve enough in the later gapped phase to rise to the top hits. In your case, relaxing the limit to 200 appears to have allowed hits that would have been excluded in the ungapped phase at 100 max target sequences to rise."

With this said, I noticed that changing max target sequence will change the overall number of hits that are E value <1. However, with in one value of max target sequence(MTS), the values should not be variable. For example, if MTS is set to 100, the hits should always be 124.

3. In what species is the closest non-S. cerevisiae relative?
Naumovozyma castelli OR T. delbruecki (Castelli was just moved into Saccharomyces genus)

What is the score and % identity for Rap1's closest relative?
Castelli 820 52%
delbruecki 853 and 52%

4. Did you get more or less hits than before? Why?
126. This matrix is better than blosum62 for differentiating between closely related sequences. The parameters, however, also changed, which is why there are not fewer hits with this matrix.

FULL Credit: This is likely because BLAST automatically adjusts its gap costs parameter to match the substitution matrix selected as a scoring parameter. More specifically, the BLOSUM62 substitution matrix is associated with a default gap cost of existence: 11 extension: 1of existence: 10 extension:dex were increased, one would expect to see a decrease in the number of hits obtained. This is because higher BLOSUM indices require closer (more stringent) matches, which would be harder to achieve for the same sequence when keeping all other scoring parameters constant.

5. Now what species is the closest non-Sacchromyces relative in?
Torulaspora delbrueckii

6. Find the protein that *was* the closest relative according to BLOSUM62. What is the new % identity and score?
778 889 72% Torulaspora delbrueckii **Correct Also**
MAX TOTAL
746 881 69% (or 85% in BLOSUM80) Naumovozyma castellii **CORRECT**

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

7. How many hits did you get? Why do you think this number changed the way it did?
124, which is one more than before. Some potential reasons:
1: The extension penalty was increased (from 1 to 2), so long gaps incur a higher penalty, resulting in a lower number of hits.
2: We have decreased the penalty for small gaps, so many more hits will score well. This increases the stat

8. Why did the score of the closest ortholog change the way it did?
Gap existence 7, gap extension 2: Naumovozyma castelli score: 732 identity: 52%
Gap existence 11 gap extension 1: Naumovozyma castelli score: 820 identity: 52%
The score decreased. There is probably a long gap in this alignment, so the difference is due to the increased gap extension penalty.

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

More, because there are more hits per word

10. Isn't online BLAST really slow?
Yes

PART#2

1.0 Report the command you used to align the reads.
bowtie2 -x <path and name of the index file: eg. chr22_idx> -U <unpaired reads file> -S <output file>

26662 reads; of these:
26662 (100.00%) were unpaired; of these:
9418 (35.32%) aligned 0 times

7113 (26.68%) aligned exactly 1 time
10131 (38.00%) aligned >1 times
64.68% overall alignment rate

1.1. How many reads map uniquely to chr21?
7113 (26.68%) aligned exactly 1 time
1.2. How many reads map to multiple locations?
10131 (38.00%) aligned >1 times
1.3 How many reads were unmappable?
9418 (35.32%) aligned 0 times

1.4 Place your output file and your report file in the submissions folder.
Sam file and report file are in the submission folder

2. What is enriched in this dataset?
(Hint: Look at the relative enrichment of single and di-nucleotides)

Report the enrichment of all single and di-nucleotides. Mention what the dataset is enriched for and interpret it. What assay do you think the data came from?

The dataset is enriched for CpGs, which are likely to come from MRE-seq or unmethylated DNA assays.

Original	Percent Reads	Percent Chr22	Enrichment
AA	0.09177332285779456	0.07950089892575024	1.154368367
AC	0.04823884228400537	0.0509160332337075	0.947419491
AG	0.05506703500012672	0.07485716383457762	0.735628124
AT	0.08062599792178828	0.060643355074755996	1.32951084
CA	0.06801682844615657	0.07630178506545118	0.891418574
CC	0.07035963200446055	0.06686140259781304	1.052320611
CG	0.06304027168816687	0.016261021945466296	3.876771823
CT	0.052398307017766176	0.07378648877838694	0.710134171
GA	0.08243353524089515	0.06280387944345324	1.312554829
GC	0.061652431761157714	0.054202057542232264	1.137455561
GG	0.08126669538991814	0.06805290437470775	1.194169391
GT	0.03717363205514864	0.05093501478392974	0.729824703

TA	0.03282510074258053	0.047313296460307026	0.693781732
TC	0.07245811896494919	0.061193241912920165	1.184086947
TG	0.0643449831462098	0.07682806540714607	0.837519243
TT	0.038325265478875735	0.07954339061939496	0.481815839

Nucleotides	Percent Reads	Percent Chr22	Enrichment
A	0.27542949277043843	0.26592178081189993	1.035753792
C	0.2533452225680747	0.23318764228383543	1.086443604
G	0.2629793035295624	0.23598397968303175	1.114394731
T	0.2082459811319	0.2649065972212	0.786110966

EC1: Canine genome