Quantitative Cellular and Molecular Biology Laboratory
Computational Biology Department
Comp Bio 02-261
Spring 2019

Lab 3 – Microbiome Analysis Lab

February 8, 2019

Microbiome Sequencing

- Extract DNA
- Amplify 16S ribosomal RNA gene (found in all bacteria)
  - How can we do this for multiple species?
- Sequence copies of amplified 16S gene DNA

Programming Tasks

1. Implement function for matching of experimental read to known 16S gene
2. Implement alignment free sequence matching
3. Plot accuracy curves for different size k-mers (k=1,3,5,7,9,11)
4. Generate plots showing relative distributions of bacteria types in different microbiomes
5. Write function to BLAST unmatched sample sequences

What are you provided with?

- Sample sequencing reads organized by sample number
- All known 16S genes
  (n=20,486; average bp = 1350)
- Code with some helper functions implemented for you already.
  (n=332,649; average bp = 397)

1. Implement function for matching of experimental read to known 16S gene

- For a given sample sequence:
  - Determine 16S sequence with greatest local alignment

2. Implement alignment free sequence matching

- Alignment Free Sequence Matching
  - Matching two sequences based on the relative presence or absence of *k*-mers
  - *k*-mer = substring of length k
  - Example:
    - ACTGA -> 1-mer -> [A,C,T,G]
    - ACTGA -> 2-mer -> [AC, CT, TG, GA]
    - ACTGA -> 3-mer -> [ACT,CTG,TGA]
    - ACTGA -> 4-mer -> [ACTG,CTGA]

2. Implement alignment free sequence matching

- Simple Alignment Free Sequence Matching Algorithm
  1. Convert each sequence to k-mer sets
  2. Calculate Jaccard index of the pair of sets

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$\frac{\text{Count k-mers in both A and B}}{\text{Count k-mers in A and/or B}}$$

$$0 \leq J(A,B) \leq 1$$

- Determine similarity of Sequences A and B
- A = ACTGGA
- B = CGTGAG

2. Implement alignment free sequence matching

- Example: Determine similarity of Sequences A and B (k=2)
  - A = ACTGGA
  - B = CGTGAG

$$\frac{\text{Count k-mers in both A and B}}{\text{Count k-mers in A and/or B}}$$

2. Implement alignment free sequence matching

- Example: Determine similarity of Sequences A and B (k=2)
  - A = ACTGGA -> [AC, CT, TG, GG, GA]
  - B = CGTGAG -> [CG, GT, TG, GA, AG]

$$\frac{\text{Count k-mers in both A and B}}{\text{Count k-mers in A and/or B}}$$

2. Implement alignment free sequence matching

- Example: Determine similarity of Sequences A and B (k=2)
  - A = ACTGGA -> [AC, CT, TG, GG, GA]
  - B = CGTGAG -> [CG, GT, TG, GA, AG]

$$\frac{[TG, GA]}{[AC, CT, TG, GG, GA, CG, GT, AG]} = \frac{2}{8}$$

$$\frac{\text{Count k-mers in both A and B}}{\text{Count k-mers in A and/or B}}$$

Issues with this approach?

2. Implement alignment free sequence matching

- Example: Determine similarity of Sequences A and B (k=2)
  - A = ACTGGA -> [AC, CT, TG, GG, GA]
  - B = CGTGAG -> [CG, GT, TG, GA, AG]

$$\frac{[TG, GA]}{[AC, CT, TG, GG, GA, CG, GT, AG]} = \frac{2}{8}$$

$$\frac{\text{Count k-mers in both A and B}}{\text{Count k-mers in A and/or B}}$$

Issues with this approach?
What if sequences are different lengths?
(You solve this.)
How do we threshold?
(Task 3)

3. Plot accuracy curves for different size k-mers across different thresholds (k=1,3,5,7,9,11)

- Accuracy assessment:
  - How do we determine truth?
  - At what level of J(A,B) (or your similar function) do we consider the sequences matched?

3. Plot accuracy curves for different size k-mers across different thresholds (k=1,3,5,7,9,11)

- Accuracy assessment:
  - How do we determine truth?
    - “True” match is the 16s sequence with the best alignment score (>90% of highest possible score) when compared to our sample sequence.
  - At what level of J(A,B) (or your similar function) do we consider the sequences matched?
    - It depends…

Accuracy Assessment

Sample Sequence	Alignment Score	Alignment Best Match	K-mer Score	K-mer Best Match
1	0.96	16s_133	0.85	16s_133
2	0.80	16s_14	0.72	16s_124
3	0.97	16s_17	0.86	16s_17
4	0.83	16s_19	0.73	16s_19
5	0.95	16s_135	0.82	16s_1325
6	0.87	16s_12	0.80	16s_102

Accuracy Assessment

Sample Sequence	Alignment Score	Alignment Best Match	K-mer Score	K-mer Best Match
1	0.96	16s_133	0.85	16s_133
2	0.80	~~16s_14~~	0.72	16s_124
3	0.97	16s_17	0.86	16s_17
4	0.83	~~16s_19~~	0.73	16s_19
5	0.95	16s_135	0.82	16s_1325
6	0.87	~~16s_12~~	0.80	16s_102

Establish Truth:
Set Alignment
Threshold to 0.9.

Accuracy Assessment

Sample Sequence	Alignment Score	Alignment Best Match	K-mer Score	K-mer Best Match
1	0.96	16s_133	0.85	~~16s_133~~
2	0.80	~~16s_14~~	0.72	~~16s_124~~ ★
3	0.97	16s_17	0.86	~~16s_17~~
4	0.83	~~16s_19~~	0.73	~~16s_19~~ ★
5	0.95	16s_135	0.82	~~16s_1325~~
6	0.87	~~16s_12~~	0.80	~~16s_102~~ ★

Establish Truth:
Set Alignment
Threshold to 0.9.

Establish Predictions:
Set k-mer score threshold
to 0.87.

3/6
correct!

Accuracy Assessment

Sample Sequence	Alignment Score	Alignment Best Match	K-mer Score	K-mer Best Match
1	0.96	16s_133	0.85	16s_133 ★
2	0.80	~~16s_14~~	0.72	16s_124
3	0.97	16s_17	0.86	16s_17 ★
4	0.83	~~16s_19~~	0.73	16s_19
5	0.95	16s_135	0.82	16s_1325
6	0.87	~~16s_12~~	0.80	16s_102

Establish Truth:
Set Alignment
Threshold to 0.9.

Establish Predictions:
Set k-mer score threshold
to 0.71.

2/6
correct!

Accuracy Assessment

Sample Sequence	Alignment Score	Alignment Best Match	K-mer Score	K-mer Best Match
1	0.96	16s_133	0.85	16s_133 ★
2	0.80	~~16s_14~~	0.72	~~16s_124~~ ★
3	0.97	16s_17	0.86	16s_17 ★
4	0.83	~~16s_19~~	0.73	~~16s_19~~ ★
5	0.95	16s_135	0.82	16s_1325
6	0.87	~~16s_12~~	0.80	~~16s_102~~ ★

Establish Truth:
Set Alignment
Threshold to 0.9.

Establish Predictions:
Set k-mer score threshold
to 0.81.

5/6
correct!

Accuracy Assessment

Sample Sequence	Alignment Score	Alignment Best Match	K-mer Score	K-mer Best Match
1	0.96	16s_133	0.85	16s_133 ★
2	0.80	~~16s_14~~	0.72	~~16s_124~~ ★
3	0.97	16s_17	0.86	16s_17 ★
4	0.83	~~16s_19~~	0.73	~~16s_19~~ ★
5	0.95	16s_135	0.82	16s_1325
6	0.87	~~16s_12~~	0.80	~~16s_102~~ ★

Establish Truth: Set Alignment Threshold to 0.9.

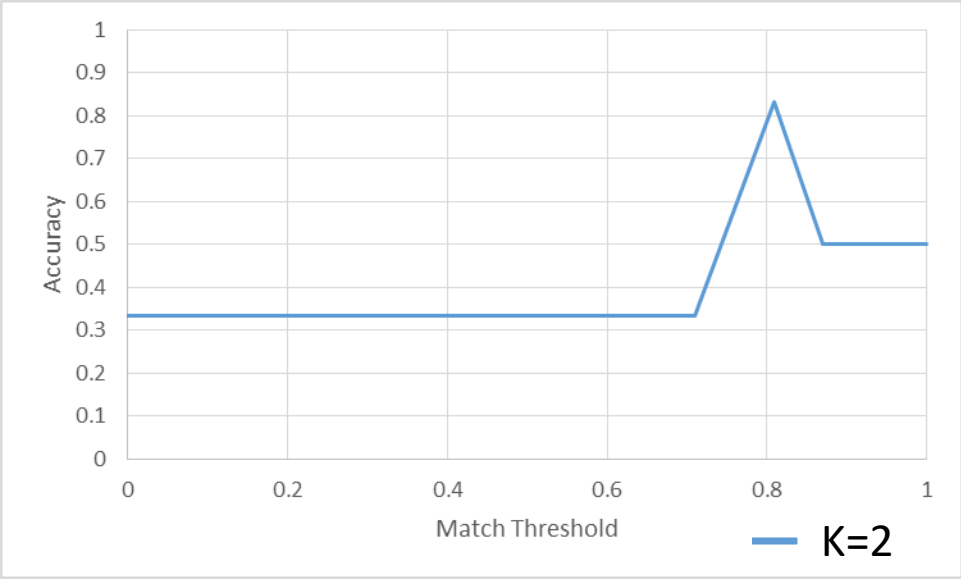
Different thresholds will yield different accuracies. We need to check a lot of thresholds to determine what is best.

Establish Predictions: Set k-mer score threshold to 0.81.

5/6 correct!

Parameter Selection Plot

- What is the best K and k-mer match score threshold?



If runtime is an issue, you may reduce the size of the libraries. Stick to at least 100 sequences in each library.

4. Generate plots showing relative distributions of bacteria types in different microbiomes

- For each sample location, calculate mixture fractions for each phylum.
  - Use GetPhylum(16s_seq_id) function to give you the phylum based on the ID.
  - Multiple samples for each location should be averaged after mixture fractions are calculated for final mixture.

- Plot these results in some interesting way.

5. Write function to BLAST unmatched sample sequences

- Some sequences are unmatched.
- What can we do about this?
  - BLAST -> Basic Local Alignment Search Tool
  - Align our sequence of interest against every an enormous set of publicly available data.
  - Mystery sequence:

CCGGTTAATCTCGTGCCAGCGACCGCGGTTACACGACAGACCCAAGACAATACCACCGGCGTAAAGCACG
ACTAAAACAGATATGTTACCACACTAGGGATAAAGCAAAACTGGGCTGTAAAAAGCCATAAGCCACACTA
AAAACTAAGCCCTAACATAAAACATCTTCGACTCGTGAAAGCAAGGACACAAACTAAGATTAGATACCTT
ACTATGCCCAGCCTTAACAAAACAATCAAATAACGAATTGTTCGCCAAATAACTACGAGTTAAAACTTAA
AATTTAAAAGACTTGACGGTACTTCACACCAACCTAGAGGAGCCTGTCTATTAACCGATAATCCACGATT
AACCCAACCCTTTCTAGCCCAACAGTCTATATACCGCCGTCGCCAGCTTACCTTGTGAAAGAAACAAAGT
AAGCCCAATAACATCACATTAATACGACAGGTCGAGGTGTAACCAATGAAAGGGGCCAAGATGGGCTACA
TTTTCTAATCCAGAAAAGCCTACAACGAATAAACTATGAAACTAGAAACTGAAGGCGGATTTAGCAGTAA
GCTAAGAATAGAATACTTAACCGAAATTAACGCAATGAAGTGCGCACACACCGCCCGTCATCCCTGTAAG
ACACATAAACTATTCATAATATCTTATCTTCTCCAAAGCAGGGCAAGTCGTAACATGGTAAGCGTACTGG

5. Write function to BLAST unmatched sample sequences

- Some sequences are unmatched.

12 unit:

- Write a function to BLAST unmatched sample sequences and return species information for best hit
  - In Anaconda Console: conda install -c conda-forge biopython
  - http://biopython.org/DIST/docs/tutorial/Tutorial.html

9 unit:

- Manually BLAST an unmatched sequence
  - https://blast.ncbi.nlm.nih.gov/Blast.cgi

What to turn in?

- Task 1: code
- Task 2: code
- Task 3: code and plot
- Task 4:
  - Code, plot, and caption
  - Look up (google, Wikipedia, etc.) one of the interesting phyla and write a paragraph about it and why you suspect it was discovered in these quantities
- Task 5:
  - 12 unit: code, paragraph describing top hit of one unmatched sequence
  - 9 unit: screenshot of BLAST result, paragraph describing top hit of one unmatched sequence (include sequence identifier)

Due: February 20