Barcoding Bioinformatics
Part II

Jason Williams
Cold Spring Harbor Laboratory, DNA Learning Center
williams@cshl.edu
@JasonWilliamsNY
DNALC Live

This is an experiment; give us feedback on what you would like to see!
DNALC Live

• Provide genetics, molecular biology, and bioinformatics learning resources

• Laboratory and computer demos, short online courses for middle school, high school, and the general public

• Interviews with scientists, help for teachers

• At-home activities, social media contests, and more
DNALC Website and Social Media

dnalc.cshl.edu

dnalc.cshl.edu/dnalc-live
DNALC Website and Social Media

youtube.com/DNALearningCenter

facebook.com/cshldnalc

@dnalc

@dna_learning_center
Barcoding Bioinformatics
Part II
Who is this course for?

- **Audience(s):** US AP Biology (high school grades 10-12) AND Intro undergraduate biology

- **Format:** 3 sessions (1 per week); ~45 minutes each

- **Exercises:** Follow along with our online bioinformatics tool DNA Subway

- **Learning resources:** Slides and packet available (teachers can also request the teacher edition)
Course Learning Goals

• Learn how DNA can be used to identify unknown organisms

• Understand how we obtain DNA Sequence and access its quality

• Use BLAST* to compare an unknown DNA Sequence to known sequences

• Compare DNA Sequences using phylogenetics

*AP Bio (Lab 3 – Comparing DNA Sequences)
Lab Setup

• We will be using DNA Subway – You can get a free account at cyverse.org (optional)
Barcoding Bioinformatics
Part II
(Sequence cleaning and BLAST)
Steps for today’s session

• Recap on our experimental dataset
• Review of sequence quality
• Sequence cleaning and pairing
• Introduction to BLAST
Recap of the dataset
Steps to DNA Barcoding

Organism is sampled → DNA is extracted → “Barcode” amplified

Sequenced DNA is compared with DNA in a barcode database
Example barcoding experiment

Mary Acheampong, Bobby Glover, and Marisa VanBrakle

Mentor: Allison Granberry
Hostos-Lincoln Academy of Science,
The Bronx

2012 UBP Grand Prize Winners
Example barcoding experiment

<table>
<thead>
<tr>
<th>Sample Letter</th>
<th>Form</th>
<th>DNA Expected</th>
<th>DNA Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Capsule</td>
<td>Ginkgo biloba</td>
<td>Rice: Oryza sativa</td>
</tr>
<tr>
<td>B</td>
<td>Capsule</td>
<td>Ginkgo biloba</td>
<td>Rice: Oryza sativa</td>
</tr>
<tr>
<td>C</td>
<td>Capsule</td>
<td>Ginkgo biloba</td>
<td>Rice: Oryza sativa</td>
</tr>
<tr>
<td>D</td>
<td>Tablet</td>
<td>Ginkgo biloba</td>
<td>No sequence available</td>
</tr>
<tr>
<td>E</td>
<td>Capsule</td>
<td>Ginkgo biloba</td>
<td>Rice: Oryza sativa</td>
</tr>
<tr>
<td>F</td>
<td>Liquid</td>
<td>Ginkgo biloba</td>
<td>No sequence available</td>
</tr>
<tr>
<td>G</td>
<td>Capsule</td>
<td>Ginkgo biloba</td>
<td>No sequence available</td>
</tr>
<tr>
<td>H</td>
<td>Tea</td>
<td>Ginkgo biloba</td>
<td>Other rib DNA present but not Melia pipera</td>
</tr>
<tr>
<td>L</td>
<td>Capsule</td>
<td>Ginkgo biloba</td>
<td>Rice: Oryza</td>
</tr>
<tr>
<td>Aedes adult</td>
<td>Anopheles adult</td>
<td>Culex adult</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------</td>
<td>-------------</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Aedes larva</th>
<th>Anopheles larva</th>
<th>Culex larva</th>
</tr>
</thead>
</table>

Photograph by Michele M. Cutwa, University of Florida.
Why does this matter?

**Aedes:**
- Chikungunya
- Dengue fever
- Lymphatic filariasis
- Rift Valley fever
- Yellow fever
- Zika

**Anopheles:**
- Malaria
- Lymphatic filariasis

**Culex:**
- Japanese encephalitis
- Lymphatic filariasis
- West Nile fever
Experimental components/design

Materials

• We have DNA from unknown mosquito samples
• We can obtain DNA from known samples
Experimental components/design

Materials

• We have DNA from unknown mosquito samples
• We can obtain DNA from known samples

Hypothesis

• We can use computational methods (BLAST/phylogenetic analysis) to infer the species
Experimental components/design

Materials

- We have DNA from unknown mosquito samples
- We can obtain DNA from known samples

Hypothesis

- We can use computational methods (BLAST/phylogenetic analysis) to infer the species

Controls

- We have sensitivity controls (sequence quality, BLAST parameters)
- We have outgroup sequences (non-mosquito, negative controls) and known samples (positive controls)
Review of sequencing and quality
DNA Sequencing

Cycle Sequencing

To this mix, we also add a second type of nucleotide; one that has a slightly different chemical formula. These "dideoxynucleotides (ddNTP)" can be recognized by a DNA sequencer.

https://dnalc.cshl.edu/resources/animations/cycseq.html
Chromatogram/Electropherogram
Some sequence examples...

High Quality Sequence

Acceptable Quality Sequence

Low Quality Sequence (multiple base calls per position)

Low Quality Sequence (no base calls)
## Phred scores...

<table>
<thead>
<tr>
<th>Phred Score</th>
<th>Error (bases miscalled)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1,000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10,000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100,000</td>
<td>99.999%</td>
</tr>
</tbody>
</table>
If 99% was good enough

If things only work correctly 99.9% of the time...

- 12 newborns will be given to the wrong parents daily.
- 114,500 mismatched pairs of shoes will be shipped/year.
- 18,322 pieces of mail will be mishandled/hour.
- 2,000,000 documents will be lost by the IRS this year.
- 2.5 million books will be shipped with the wrong covers.
- Two planes landed at Chicago's O'Hare airport will be unsafe every day.
- 315 entries in Webster's Dictionary will be misspelled.
- 20,000 incorrect drug prescriptions will be written this year.
- 880,000 credit cards in circulation will turn out to have incorrect
cardholder information on their magnetic strips.
- 103,260 income tax returns will be processed incorrectly during the year.
- 5.5 million cases of soft drinks produced will be flat.
- 291 pacemaker operations will be performed incorrectly.
- 3056 copies of tomorrow's Wall Street Journal will be missing one of the three sections.

Photo credit
http://www.personal.psu.edu/sxt104/class/99percent.html
A note on controls

At what temperature does ice \((H_2O)\) + Chemical “X” melt?
A note on controls

Positive control: What does the effect look like if present?
A note on controls

**Positive control:** What does the effect look like if present?

**Negative control:** What does the effect look like if absent?
A note on controls

Positive control: What does the effect look like if present?

Negative control: What does the effect look like if absent?

Sensitivity control: Across what range of values can I measure the effect?
A note on controls

Positive control

Negative control

Sensitivity control

Photo credits
https://commons.wikimedia.org/wiki/File:Water_in_a_beaker.JPG
http://www.chem.uiuc.edu/webfunchem/temperature/Temp10.htm
https://www.dreamstime.com/photos-images/alcohol-thermometer.html
A note on controls

Positive control

Negative control

Sensitivity control

Photo credits
- commons.wikimedia.org/wiki/File:Water_in_a_beaker.JPG
- www.chem.uiuc.edu/webfunchem/temperature/Temp10.htm
- en.clipdealer.com/vector/media/A:174945087
Phred are our measure of quality (signal/noise)

Lower score = more noise than signal
Bi-directional sequencing

Photo credit
https://www.omicsonline.org/articles-images/CMBO-2-108-g003.html
Reverse complementation

5’ Plus Strand 3’

ATG  TAA

3’ Reverse Primer 5’

ATT

5’ Forward Primer 3’

ATG

TAC  ATT

3’ Minus Strand 5’

Photo credit
Reverse complementation

- Reverse: change nt. sequence from (5' → 3') to (3' → 5')
- Complement with the reversed sequence

Reverse: 5' CTCCAAGCTCCAAGCTCCAG 3'
Complement: 5' GACCTCGAACCCTCGAACCCT 3'

Photo credit
Clean up and consensus
Introduction to BLAST
Basic Local Alignment Search Tool

- An algorithm for searching a database of sequences
Basic Local Alignment Search Tool

- An algorithm for searching a database of sequences
- “Google for DNA” (although works with any biological sequence, and started before Google ~1985)
Basic Local Alignment Search Tool

- An algorithm for searching a database of sequences
- “Google for DNA” (although works with any biological sequence, and started before Google ~1990 vs 1998)
- NCBI is the most popular interface, but this is software that can be run anywhere (including Subway)
Warning: Analogy

(useful for discussion but not the whole picture)
BLAST algorithm analogy

Query sequence

ACTGACATCGGGGTTGCTACG
BLAST algorithm analogy

Query sequence

ACTGACATCGGGGTGCTACG

Database
BLAST algorithm analogy

National Library of Medicine
Twenty Seven Years of Growth:
NCBI Data and User Services

Photo credit
https://www.nlm.nih.gov/about/2018CJ.html
BLAST algorithm analogy – searching by “word”

Break the *Query sequence* Into “words” (k-mers)

ACT GAC ATC GGG GTG CTA CG

Database
BLAST algorithm analogy – searching by “word”

Break the *Query sequence* into “words” (k-mers)

**ACT** **GAC** **ATC** **GGG** **GTG** **CTA** **CG**

Database

**ACT**... **TCT**... **GCT**...
Let’s BLAST a sequence

>mosquito-1F
CTTTAAGTATATTAATTCGTGCTGAATTAAGTCACCCAGGGATATTTAT
TGGAATGATCAAATTTTATAACGTAATTGTTACAGCTCATGCAATTATT
ATAATTTTTTTATAGTAATACCAATTATAATTGGAGGATTGGGAATT
GATAGTTCTTTTAATATTAGGAGCTCCTGTGATATAGGCAATTCTCCTGGAAT
AAATAATATAAGTTTGGAAATATTACCTCCTTCTTTAACTCTACTACTTT
CTAGTTCAATAGTAAAAATGGAGCAGGGACAGGATGAACAGTTTTA
TCCCTCCTCTTCATCGGAACAGCAGCAGCTGGAGCTTCTGTTGATT
AGCAATTTTCTCTCTCATTTAGCAGGGATTTCATCTATTTTAGGAGC
AGTAAATTTTATTACTACTGTTATTAATATACGATCATCTGGAATTACTT
TAGATCGATTACCTTTATTTGATGTCTGTATATTACTGCTATTTTA
TTACTTTTATCTCTCTCTGTATTAGCTGGAGCTATTATAATTATTACT
GATCGAAATTTAAATACTCTCCTTTTGAACCAATTGGAGGAGGAGA

BLAST and controls

Why smaller databases are better (more sensitive) –
statistics

\[ S' = \lambda S_{raw} - \ln K \cdot m \cdot n \]
\[ S_{bit} = (\lambda S_{raw} - \ln K) / \ln(2) \]
\[ P(S' > x) = 1 - \exp(-e^{-x}) \]
\[ P(S_{bit} > x) = 1 - \exp(-mn2^{-x}) \]
\[ E(S' > x \text{ ID}) = P \cdot D \]

\[ P(\text{B bits}) = mn \cdot 2^{-B} \]
\[ P(40 \text{ bits}) = 1.5 \times 10^{-7} \]
\[ E(40 | D=4000) = 6 \times 10^{-4} \]
\[ E(40 | D=80E6) = 12 \]
BLAST algorithm analogy – searching by “word”

The *Query sequence* is aligned to a *Subject* (a sequence in the database)

Q: **ACTGAC**–**ATCGGGGTGCTACG**

\[\text{||| ||| |||| | || |||| ||} \]

S: **ACTGACCATCGGAGTGCTACG**
BLAST algorithm analogy – alignment

\[
S = \sum (\text{identities, mismatches}) - \sum (\text{gap penalties})
\]

Score = \text{Max}(S)

Photo credit
https://www.ncbi.nlm.nih.gov/books/NBK62051/
Let’s do a BLAST
Some BLAST definitions

- **Max Score**: Highest alignment score (according to a formula)
Some BLAST definitions

- **Max Score**: Highest alignment score (according to a formula)
- **Query Cover**: % of the query length included in aligned segment
Some BLAST definitions

• **Max Score**: Highest alignment score (according to a formula)

• **Query Cover**: % of the query length included in aligned segment

• **E value**: The number of alignments expected by chance with the calculated score or better
Some BLAST definitions

- **Max Score**: Highest alignment score (according to a formula)
- **Query Cover**: % of the query length included in aligned segment
- **E value**: The number of alignments expected by chance with the calculated score or better
- **Per. Identity**: Highest % identity for a set of aligned segments to the same subject sequence.
Does BLAST tell me what species I have identified?
(Some) Limitations to BLAST

- **Homology**: BLAST is trying to indicate which homologous (related by ancestry) sequences are found in the database.
(Some) Limitations to BLAST

- **Homology**: BLAST is trying to indicate which homologous (related by ancestry) sequences are found in the database.

- **Database coverage**: BLAST returns its best result; that is not guaranteed to be the true result.
(Some) Limitations to BLAST

- **Homology**: BLAST is trying to indicate which homologous (related by ancestry) sequences are found in the database.

- **Database coverage**: BLAST returns its best result; that is not guaranteed to be the true result.

- **Locus resolution**: Barcodes are often good for **genus**-level resolution.
A note on resolution (and controls)

Photo credit
Next time:

Multiple sequence alignments and phylogenetics
DNALC Website and Social Media

dnalc.cshl.edu

dnalc.cshl.edu/dnalc-live