CSE 491 Introduction to Bioinformatics
Homework 6
To Catch a Serial Killer

Assigned on April 16, 2019.

Source code must be submitted using D2L by 5 pm EST on May 3, 2019.

Reminder: CSE 491 adheres to Michigan State University's policies on academic integrity. No collaboration is allowed on homeworks, and all work must be your own.

1 Introduction

DNA sequence data analysis is playing an increasingly important role in criminal forensics. Recently, computational analyses of DNA sequence data have been critical to solving several notorious cold cases, including investigations involving the Golden State Killer and other serial killers [3, 2].

Breakthroughs in these cold cases relied on the use of public DNA sequence databases to "triangulate" possible suspects based on their genetic relationships to (possibly very distant) relatives who had previously submitted their DNA sequence information and other identifying information to the databases. The topic is part of a nascent field that is known as genetic genealogy, and the computational approaches used in this field make use of fundamental techniques from phylogenetics and population genetics.

In this homework, we will follow the trail of a cold case in much the same way as in these celebrated cases. Other than the fictitious crime discussed below, all other aspects of this homework are real and true. In particular, all of the sequence data and metadata used in this homework are authentic data from a major public database [1].

Note that the computational approach used in this homework differs somewhat from those used by forensic investigators. Both have advantages and disadvantages relative to each other.

2 A (Slightly) Hypothetical Scenario

A crime scene sample has been retrieved from cold storage. The suspect's DNA has been extracted from the sample and sequenced.

Forensic specialists are now interested in using public DNA sequence information to try to help identify the suspect. They have downloaded all of the data from a major public DNA database.

3 Download Datasets

Access the class D2L website, navigate to Homework 6, and download the public database sequences file database-sequences.fasta, the suspect's sequence file query-sequence.fasta, and the metadata file metadata.tsv.

Browse the two sequence data files using a text editor. Both DNA sequence files are encoded in FASTA format. The FASTA format is very simple. Suppose that the file contains n sequences $s_1, s_2, \ldots, s_n$ that correspond to taxa with unique IDs $x_1, x_2, \ldots, x_n$. Each taxon ID appears on a separate line beginning with a ">" symbol, followed by the corresponding sequence on the next line. Thus, the first line of the corresponding file will be "$> x_1$", followed by the sequence s_1 on the next line. Then, the next line will be "$> x_2$", followed by the sequence s_2. The pattern repeats until the last two lines, which will be "$> x_n$", followed by the sequence s_n.

Note that, in general, sequences may be split across multiple lines to improve readability (although this is not the case for the two input files in this homework). For more information about the FASTA file format, see https://www.ncbi.nlm.nih.gov/BLAST/fasta.shtml.

The metadata file metadata.tsv is a tab-delimited spreadsheet. Use any spreadsheet software (e.g., Microsoft Excel or LibreOffice Calc) to view the metadata information. The spreadsheet includes a first header row with descriptive information. Each subsequent row contains metadata for a person (or "taxon"), including a unique ID which can be used to cross-reference the person's DNA sequence in the database file database-sequences.fasta.

4 Parse Sequence Data (10 points)

Write code to parse a FASTA input file and then store the data in a data structure which enables efficient lookup based on taxon IDs. Hint: a good option would be a hash table which efficiently maps a taxon ID to its corresponding sequence string.

Use your code to parse both FASTA input files.

5 Compute Distance Matrix (10 points)

Compute the pairwise Hamming distance matrix for all sequences (i.e., the sequences contained in the file database-sequences.fasta as well as the file query-sequence.fasta).

6 Infer Phylogeny (20 points)

Implement the UPGMA algorithm. Using the pairwise Hamming distance matrix as input to your UPGMA implementation, infer a phylogenetic tree that describes the relationships among the query and public database individuals. (Hint: make use of the tree data structure from past homeworks.)

7 Output Phylogeny (10 points)

Write code to output the Newick string corresponding to your inferred phylogenetic tree. See lecture 6 for class discussion about the Newick string format for rooted phylogenetic trees. (Hint: make use of the tree data structure from past homeworks.)

8 Solving the Cold Case (10 points)

Study the estimated tree. Who are the individuals listed in the database that are most closely related to the suspect? Based on the metadata spreadsheet, what do these individuals have in common, and where should we start looking for the suspect?

9 Submitting Source Code and Other Solutions

Submit the following for full credit:

- Source code for all program(s)
- A text document with your UPGMA-estimated tree and written solutions for "Solving the Cold Case" section questions.

10 Optional Bonus (10 points)

The application of phylogenetics and population genetics to criminal forensics has much promise, but also brings serious risks. Apart from bioethical and criminal justice concerns, there has already been at least one case of misidentification of an individual using these types of computational approaches. Nevertheless, police departments and criminal forensic departments in the U.S. and other countries have dramatically expanded their use of genetic genealogy analysis for criminal forensic purposes.

Can you give a *computational* reason why inferring genetic relationships among a suspect and individuals listed in a public DNA sequence database might not work as intended? Note: bonus points will be assessed based on the algorithmic insights provided by your answer. No bonus points will be given for responses that deal with topics other than algorithmic details. (For example, bioethics/privacy concerns are legitimately worrisome but will not be eligible for bonus points.)

References

[1] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56, 2012.

[2] C. Phillips. The golden state killer investigation and the nascent field of forensic genealogy. *Forensic Science International: Genetics*, 36:186–188, 2018.

[3] N. Ram, C. J. Guerrini, and A. L. McGuire. Genealogy databases and the future of criminal investigation. *Science*, 360(6393):1078–1079, 2018.