Activity NN

Telephone–***Phylogenetics***

Summary

Bioinformatics is a discipline that combines computer science and biology. Bioinformatics uses the algorithms and technology of computer science, mathematics, and statistics to solve problems for biology. For example, it has allowed biologists to reconstruct phylogenetic (evolutionary) trees using computer science techniques such as string matching, edit distance, and matrices.

This activity lets students participate in the process of reconstructing a phylogenetic tree and introduces them to several core bioinformatics concepts.

Curriculum Links

- Biology: evolution
- Bioinformatics: phylogenetics
- Mathematics: matrices

Skills

- Following instructions
- Writing
- Basic math

Ages

- 10 years and up

Materials

You will need:

- 20 to 40 minutes
- 10+ students
- A predetermined "secret message" written on a slip of paper
- A blackboard or whiteboard

Nine students will each need:

- A small slip of paper and pencil to write down the secret message.

The rest of the students will each need:

- Paper and pencil to try to reconstruct the path of the message.

Instructions and Discussion

1. Construct a message consisting of five to seven words chosen from a set of four simiilar-sounding words (example sets are available in the resources section). For example, from the set:

$$\{\textit{chair, prepare, affair, stair}\},$$

the message might be:

prepare prepare affair stair chair stair.

Write the set of four words on the board, and write the message on the slip labeled "Student 1" to begin the telephone chain.

2. Select nine students to participate in constructing the telephone chain. Hand each student one of the slips of paper prepared for you in the resources section. Send the rest of the students out of the room.

3. Ask Student 1 to memorize the message written on their slip of paper and return the slip to you. Instruct Student 1 to whisper the message they have memorized to Student 2 and then separately to Student 3. Allow the students to whisper the message only once.

4. Instruct Students 2 and 3 to write the message that they heard on their slip of paper. From memory, have each student whisper the message to the appropriate people, following the tree structure shown in figure 1. The process continues until the tree is complete, with each of the students writing the message they hear on their slip of paper, whispering it to the next students in the chain, and then returning the piece of paper to you.

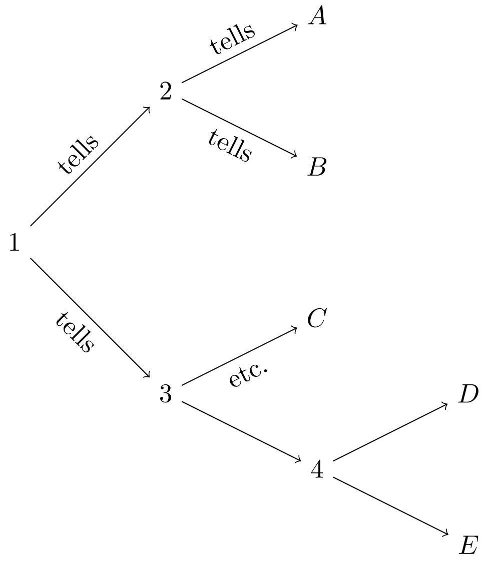

Figure 1: The telephone chain.

5. Using the slips of paper the students have written their messages on and returned to you, translate the words in each message into single letter codes for simplicity. For example, if Student 3 wrote down

prepare prepare affair chair stair affair on their slip, the message would become abbreviated as

p p a c s a.

Then assign the leaf messages (those from students *A*, *B*, *C*, *D*, and *E*) new identifying letters in a scrambled order. For example, message *A* could become message *X*, message *B* could become message *V* and so on.

6. Bring the rest of the students back into the room.

7. Draw the tree structure from figure 1 on the board. Then write the five leaf messages next to the tree structure on the board, in random order, identified by the new IDs you created in step six. For example:

V: p a a s a c
W: p a s c a c
X: p p a s c p
Y: p a s s a c
Z: p p a s c s

8. Now you can ask the rest of the students to try to guess where each of these leaf messages belongs in the tree structure you have drawn on the board. They can write their guesses on paper or on the board.

9. After the students have guessed the phylogenetic tree, attempt to re-create the true phylogenetic tree using a species distance matrix (see Matrix Theory in the *What's It All About?* section). First, construct a lower triangular 4 × 4 matrix of the pairwise distances between each pair of leaf messages, as shown in figure 2. The distances in the matrix are calculated by looking at the number of differences (or "mutations") when comparing the messages.

For example, the distance from *V* to *W* is 2, since for these messages only the third and fourth codes differ:

V: p a ***a s*** *a c*
W: p a ***s c*** *a c*

	V	*W*	*X*	*Y*
W	*2*			
X	*3*	*5*		
Y	$\boxed{1}$	*1*	*4*	
Z	*3*	*4*	*1*	*3*

Figure 2: The 4 × 4 distance matrix. The minimum distance is 1.

10. This step in the tree reconstruction uses the concept of maximum parsimony which is explained in more detail in the next section. Find the smallest entry in the matrix. It is important to note that there may be multiple smallest entries in the matrix. A ties between them is broken by randomly choosing one.

In this example, we randomly choose *V* and *Y* as our minimum distance (although we could just have easily choose *Y* and *W* or *Z* and *X*). Join the column and row headers of the

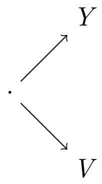

Figure 3: The beginning of the current tree.

chosen entry. They become two leaf nodes in the phylogenetic tree connected to a common internal node (or ancestor), as shown in figure 3.

10.1 Compute a new distance matrix, in this case a 3 × 3 lower triangular one, in which the chosen column and row headers (in this case V and Y) are combined into a single header entry. To calculate the distance between this new header entry and the rest of the entries, arithmetic averages of the previous distances are used. For example, in this matrix, the (YV, W) entry consists of the average of the (Y, W) and (V, W) entries in the matrix in figure 2. The resulting matrix should appear as shown in figure 4.

	YV	*W*	*X*
W	1.5		
X	3.5	5	
Z	3	4	1

Figure 4: The 3 × 3 distance matrix.

10.2 Repeat this process by creating a 2 × 2 matrix from the 3 × 3 matrix, as shown in figure 5. The smallest entry in the 3 × 3 matrix is 1 from (*Z*,*X*) so these column and row headers have been joined, forming a new internal node seen in figure 6.

	VY	*W*
W	1.5	
XZ	3.25	3

Figure 5: The 2 × 2 distance matrix.

10.3 Select the minimum entry in the 2 × 2 matrix. At each of these steps, a new interior node of the phylogenetic tree will be created.

11. When this procedure is complete the final tree is drawn (figure 9). Replace the symbols in the tree you have just constructed with the “real” A – E and compare the final tree with the original tree. Be aware that the original tree and the computed tree may not always match.

If the computed tree is not correct see if the students can figure out this discrepancy. The two main sources of error are:

1. *randomly breaking ties* – where there is more than one smallest entry in the matrix and the wrong one is randomly chosen, and

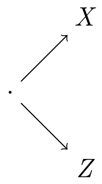

Figure 6: Next step of the tree.

	YVW
XZ	3.125

Figure 7: The 1 × 1 distance matrix.

2. *neutral mutations* – where a sequence of mutations first changes a word and then changes it back to the original. These two mutations cancel each other out. This gives a distance of 0 instead of a distance of 2.

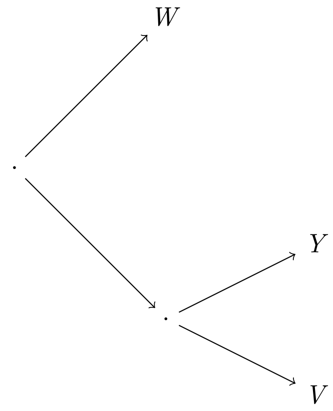

Figure 8: The tree resulting from figure 7.

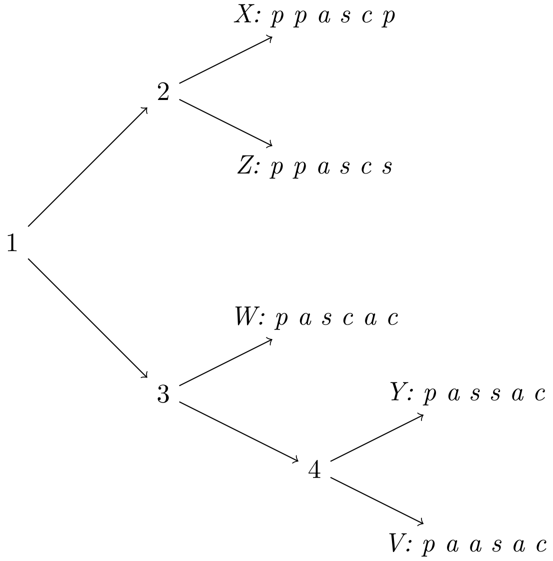

Figure 9: The telephone chain.

What's it all about?

Bioinformatics utilizes the theory and technology of computer science to address biological questions especially in the areas of genetics, evolution, protein function, cellular networks, etc. In our exercise, we addressed an area of bioinformatics related to evolution.

Evolutionary theory is vital to the work of biologists. To investigate how evolution works we look at DNA. DNA has informational content and computer science is the study of information, so computer science provides many tools for working with DNA.

The analogy of this exercise to real life is that the leaf messages represent extant evolutionarily related species whose DNA is available and can be sequenced, while the interior messages represent the DNA of extinct ancestral species whose DNA is not available. Here our task is to recreate the true course of evolution to explain how today's species came into existence.

Phylogenetic Trees

Tracing your family tree for your immediate family is easy. However, the further you go back, the harder this reconstruction becomes. Now think about how hard it would be to go back millions of years! That far back would be impossible to ask your grandma about.

Extend this idea to species; instead of mapping how family members are related, we are looking for the way today's species evolved from species in the past. A phylogenetic tree depicts the evolutionary path by which today's species evolved from species in the past.

Since we are going back millions of years we cannot keep track of all the mutations that happen and, therefore, we cannot always recreate the correct phylogenetic tree. We use the computational idea of string edit distance to record the number of mutations from generation to generation, or in the game, words changed from student to student. String edit distance calculates the minimum number of insertions, substitutions or deletions to transform one string into another. Creating a phylogenetic tree to keep track of these changes is exactly what the game does.

Maximum Parsimony

Maximum parsimony embodies in biology the philosophical principle of Occam's razor, where the most likely explanation is the simplest. Similarly, evolution is believed to happen through the smallest number of mutations from generation to generation. Consider the DNA sequence `AAGTCCAG`. If we compare this to the two sequences `ACGTCCAG` and `ATGCACGG`, we would assume that the first sequence is more closely related to the original one since only one mutation separates them.

Matrix Theory

Matrices are a very important concept in math and computer science. For our purposes, they are a convenient way to organize the number of mutations that have occurred between the leaf nodes. For this reason, we call it a *species distance matrix*. Using this method of organization, maximum parsimony is easy to see and calculate.

The species distance matrix contains the pairwise distances between species calculated by evaluating all pairs of the sequence and using the differences between the sequences as the distance.

Contributed By

Tru Women in Computer Science (TWiCS) `http://twics.truman.edu`
Truman State University, Kirksville, Missouri, USA
Mariya Davidkova
Amy McNabb
Molly Smith
Julia Stefani
Michelle VanKleeck
Allie Wehrman
and Jon Beck

Resources

Sample Message Sets

{*chair, prepare, affair, stair*}
{*school, rule, cool, fool*}
{*dream, stream, theme, beam* }
{*gum, bum, thumb, plumb* }
{*math, path, wrath, bath*}
{*play, day, stay, clay*}
{*pink, think, drink, wink*}

Message Forms

Student 1

Message:

Student 2

Message:

Student 3

Message:

Student 4

Message:

Student A

Message:

Student B

Message:

Student C

Message:

Student D

Message:

Student E

Message: