NUMB3RS Activity: DNA Sequence Alignment
Episode: “Guns and Roses”

Topic: Biomathematics – DNA sequence alignment **Grade Level:** 10 - 12
Objective: Use mathematics to compare two strings of DNA
Prerequisite: very basic knowledge about DNA
Time: 20 - 30 minutes

Introduction

Charlie Eppes learns that the FBI lifted a number of DNA samples from a victim’s apartment. He wants to try to determine something about the suspect’s ancestry based on the samples. Although examining the DNA would not enable the FBI to identify a specific person, it is possible to identify certain traits like birth defects, red hair, or freckles. A company called DNA Print has created a statistical index that can take information from a DNA sequence and use it to more closely identify its owner.

Discuss with Students

Sequence alignment is particularly useful in tracing the evolution of sequences from a common ancestor, using either protein sequences or actual DNA sequences. It also is used for comparing sequences of successive generations of the same species, or two species believed to be very closely related through evolution. Changes to the code naturally occur from one generation to the other.

A fundamental tool for comparing two closely related sequences of DNA is called **global alignment**. This is one way of determining if two sequences share a common ancestor. DNA is made up of pairs of nucleotides (A, C, G, T, where A pairs with T, and C pairs with G). The nucleotides of the two sequences are compared one at a time to see how well they align.

Example: Compare CAGCT and ATCT

We write these strings so that the letters line up in some way, and so that as many letters as possible match. Because these strings are different lengths we’ll need to use a “gap” symbol like “-”. Here are some sample results:

```
CAGCT      CAGCT      CAGCT      CAGCT      -CAGCT     CA-GCT
-ATCT      A-TCT      AT-CT      ATC-T      ATCT--     -AT-CT
```

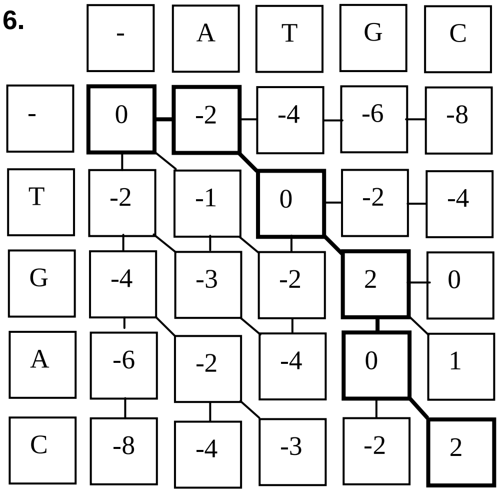
Notice that gaps can be put in one or both strings to improve the alignment.
In each step of the alignment, there are three choices: match a letter from the first string with a gap (1st and 6th example), match a letter from each string (2nd, 3rd, and 4th example), or match a gap from the first string with a letter from the second string (5th example). Because the same three choices exist until all of the letters of the shorter string have been matched with something, there are at least 3^k possible matchings, where *k* is the length of the shorter string. So, which alignment is considered the best?

The purpose of this activity is for students to understand the use of a mathematical algorithm that determines the best alignment for a pair of DNA sequences. It makes use of a scoring system for matches and mismatches, and is a variation of Dijkstra’s algorithm for finding the shortest path through a weighted graph (see Extensions). It may be necessary for you to work through and discuss with students the first part of the student page to help students understand the meaning of the entries in the table and how the scoring system works.

Student Page Answers:

1. $(3 \times 10^9)(0.001) = 3{,}000{,}000$ **2.** $4^{3{,}000{,}000} \approx 10^{602060}$. (This could be a fun one to try to write in scientific notation.) **3.** There is no need to. Both gaps could be taken out and the alignment is still the same. **4.** 3^3, or 27; 3^k. **5.** (–2) + (–1) + (+2) + (+2) = +1; (–1) + (–2) + (+2) + (+2) = +1; (–1) + (–1) + (–2) + (+2) = –2; (–1) + (–1) + (–1) + (–2) = –5; (–1) + (–2) + (+2) + (–2) + (–2) = –5.

6.



	-	A	T	G	C
-	0	-2	-4	-6	-8
T	-2	-1	0	-2	-4
G	-4	-3	-2	2	0
A	-6	-2	-4	0	1
C	-8	-4	-3	-2	2

The best alignment is:

```
ATG-C
-TGAC
```

L. Charles (Chuck) Biehl, Bro. Pat Carney, and Patrick Flynn; DIMACS (Center for Discrete Mathematics and Theoretical Computer Science) Biomathematics Institute, Rutgers University

Name: ______________________________ Date: _______________

NUMB3RS Activity: DNA Sequence Alignment

Charlie Eppes learns that the FBI lifted a number of DNA samples from a victim’s apartment. He wants to try to determine something about the suspect’s ancestry based on the samples. Although examining the DNA would not enable the FBI to identify a specific person, it is possible to identify certain traits like birth defects, red hair, or freckles. A company called DNA Print has created a statistical index that can take information from a DNA sequence and use it to more closely identify its owner.

In this activity you will learn how math is used to perform **global alignment** on a pair of DNA sequences. This is one way of determining if two sequences share a common ancestor. Global alignment is a fundamental tool for comparing two closely related sequences of DNA. Each strand of DNA is made up of pairs of nucleotides (A, C, G, T, where A pairs with T, and C pairs with G), aligned in a double helix. The nucleotides of two sequences are compared one at a time to see how well they align. We write these strings so that the letters line up in some way, with the underlying goal to match up as many letters as possible (without rearranging them, of course). Human DNA contains approximately three billion (3×10^9) base pairs of nucleotides.

1. If all human DNA is 99.9% identical, how many base pairs are left to make different DNA sequences (meaning different humans)?
2. Because bases always pair up the same way (A with T, G with C) it is only necessary to consider one side of the sequence (or one side of the double helix). If each base can be any of the four letters, and they are arranged in a sequence of 3×10^6 letters long, how many different sequences are mathematically possible?

Consider the following two short sample sequences that we want to align: ACTG and TTG. Since they are different lengths, we need to assign a “gap” symbol “-” to assist in the alignment. (The gap symbol can be used in either or both strings.) Some possible alignments are:

String 1:	`ACTG`	`ACTG`	`ACTG`	`ACTG`	`-ACTG`
String 2:	`-TTG`	`T-TG`	`TT-G`	`TTG–`	`T-T-G` (etc.)

When aligning the bases, we always have three choices for each step: align a letter from String 1 with a gap in String 2, align a letter from each string, or align a gap in String 1 with a letter in String 2. (Picture a tree diagram to go with this idea.)

3. Why would we not align a gap in String 1 with a gap in String 2?
4. Because the shorter string is three letters long, what is the least possible number of different matchings? If the shorter string has a length of *k*, how many would this be?
5. Which alignment is the best one? Suppose we use the following scoring system:
    - Add 2 for each alignment of matching letters
    - Subtract 1 for each alignment of mismatched letters
    - Subtract 2 for each alignment of a letter with a gap.

    What is the total score for each of the alignments above?

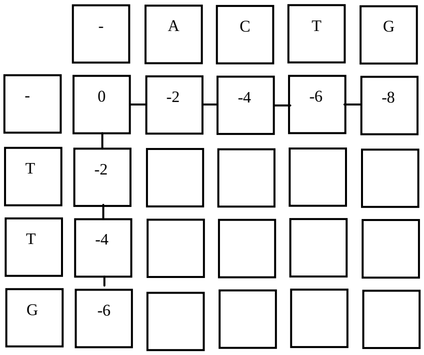
> *The scoring system used in this activity is actually arbitrary, since we don’t know the actual “score” that nature uses. This particular scoring system is useful for the sake of these examples, and is typical.*

The example below uses a matrix to do the alignment. String 1 is on top and String 2 is down the left. *Note that both strings start with a gap in the matrix, just to get things going.* Use the same scoring as before. The goal is to generate a path from the upper left to the lower right. The connections are made and scored in the following way:

- Horizontal connection – a gap is inserted in string #2 (subtract 2)
- Vertical connection – a gap is inserted in string #1 (subtract 2)
- Diagonal connection – a base is paired with a base (add 2 if they match, subtract 1 if they are different)

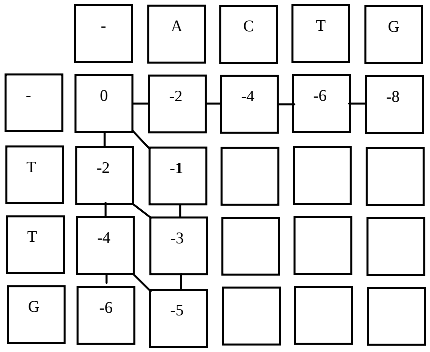
For most of the boxes, there are three ways to get to that box – vertically, diagonally, and horizontally. We only mark the connection that produces the highest total score.

For example:



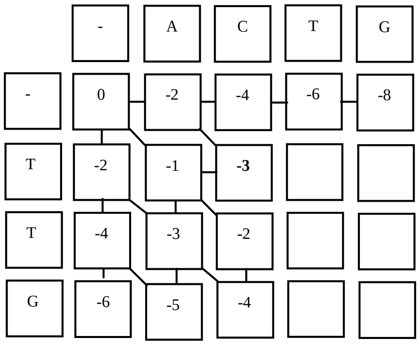
	-	A	C	T	G
-	0	-2	-4	-6	-8
T	-2				
T	-4				
G	-6				

The connections made and scores shown from inserting gaps. (Subtract 2 each time a gap is used.)



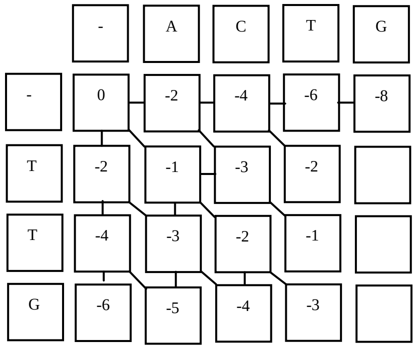
	-	A	C	T	G
-	0	-2	-4	-6	-8
T	-2	**-1**			
T	-4	-3			
G	-6	-5			

Remember to mark only the best connection. The box with a score of –1 means that aligning A with T produces a higher score than the other possible choices for the first letter. (That is, because 0 – 1 = –1, and –2 – 2 = –4, mark the connection only diagonally.)



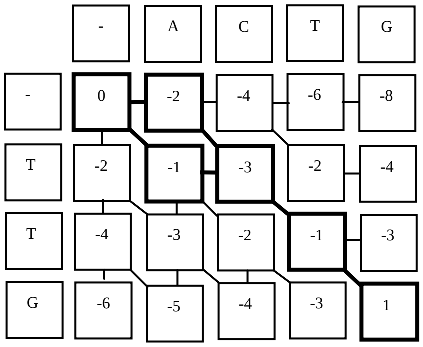
	-	A	C	T	G
-	0	-2	-4	-6	-8
T	-2	-1	**-3**		
T	-4	-3	-2		
G	-6	-5	-4		

For the –3 entry in the C column, note that if there is a tie for best choice, all are marked (here it is marked both diagonally and horizontally, because –2 – 1 = –3 (align C with T) **and** –1 – 2 = –3 (align C with a gap).



	-	A	C	T	G
-	0	-2	-4	-6	-8
T	-2	-1	-3	-2	
T	-4	-3	-2	-1	
G	-6	-5	-4	-3	

Matches are +2, so matching T with T diagonally increases the total score by 2 (–4 + 2 = –2) as shown in the T column (rather than align T with a gap horizontally or a gap with T vertically). In the 3rd row, the highest possible total is –3 + 2 = –1 (align T with T).



	-	A	C	T	G
-	0	-2	-4	-6	-8
T	-2	-1	-3	-2	-4
T	-4	-3	-2	-1	-3
G	-6	-5	-4	-3	1

The total score for this alignment is 1 (in the bottom right box). To determine the alignment(s), start at the end and trace all paths that lead back to the beginning. Since this example has two different paths, this means that the optimal alignments are:

```
ACTG    -and-    ACTG
-TTG             T-TG
```

6. Use the method described above to globally align strings ATGC and TGAC in the space provided below. Trace backwards to show the alignment (this one only has one).

	-	A	T	G	C
-	0	-2	-4	-6	-8
T	-2				
G	-4				
A	-6				
C	-8				

L. Charles (Chuck) Biehl, Bro. Pat Carney, and Patrick Flynn; DIMACS (Center for Discrete Mathematics and Theoretical Computer Science) Biomathematics Institute, Rutgers University

The goal of this activity is to give your students a short and simple snapshot into a very extensive mathematical topic. TI and NCTM encourage you and your students to learn more about this topic using the extensions provided below and through your own independent research.

Extensions

Introduction

Global alignment is mainly used for aligning sequences that are closely related. Due to many factors, evolution can cause “chunks” of a sequence to relocate elsewhere in the string, and even be reversed. These relationships cannot be found using global alignment. Other algorithms are used to perform “local” alignment, in which shorter sequences are identified and aligned within longer strings. This is one way of determining if two sequences share a common ancestor. Scientists associate probabilities with various differences between two strings. This in turn gives them much information about how the two strings are related. Much research is being done to develop and understand mathematical models related to understanding DNA.

Additional Resources

The algorithm used in this activity is called Needleman-Wunsch global alignment. For information about local alignment algorithms, such as Smith-Waterman or Framesearch, as well as a lot more information about the biology related to the mathematics, visit: **http://en.wikipedia.org/wiki/Sequence_alignment**

The algorithm used for aligning strings is based on Dijkstra’s algorithm, used to determine the shortest path between vertices in a graph (like two locations on a map). Dijkstra’s algorithm is at the heart of every Internet service that “computes” driving directions. For more about this algorithm and a demonstration of how it works, visit: **http://www.cs.sunysb.edu/~skiena/combinatorica/animations/dijkstra.html**

The Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) is teaching mathematics and biology teachers about the interface between biology and mathematics (called Biomathematics) and exploring ways of how to teach it. One of the first teaching tools developed is a Java applet developed in 2005 by high school teacher Jim Kupetz that demonstrates global (and local) string alignment. It is available for download at: **http://dimacs.rutgers.edu/dci/2005/education.html**
Just click the “Global Alignment Program” link.

L. Charles (Chuck) Biehl, Bro. Pat Carney, and Patrick Flynn; DIMACS (Center for Discrete Mathematics and Theoretical Computer Science) Biomathematics Institute, Rutgers University