Sequence Alignment Games

created by Lisa Wang

What is DNA Sequence Alignment?
DNA strands consist of four nucleotides: A, C, T, G
Over time and through diseases, mutations can happen to these strands, resulting in insertions of nucleotides, deletions or point mutations (one nucleotide replaced by another). In order to find causes for genetic diseases like down syndrome, to design drugs, to find out more about evolution or to do research on the human genome, sequence alignment is one of the most important tools.
Through proper alignment, scientists can identify where mutations have happened. Since mutations are often only very small changes to a DNA strand, we want to find the best alignment of two strand e.g. from a healthy animal and its counterpart with a genetic disease.

Appetizer Activity (~10 minutes):
Play Phylo (http://phylo.cs.mcgill.ca/), a DNA puzzle game that helps molecular biologists to find out more about genetic diseases and genetic differences between species. If you beat the computer (the "par" score), you get to the next level. While you are playing, write down what decisions increase or decrease your score. Pay attention to gaps and mismatches. How would a computer program solve this puzzle?
Increases score: _____
Decreases score: _____

Today's challenge: Align them all!
Now it's your turn! We will write a program that aligns two DNA strands.
First, we need to define how we are computing the score of an alignment.
Discuss the scorings for match, mismatch and gap together. This is a crucial step in modeling!

Match: _____ points
Mismatch: _____ points
Single Gap: _____ points

If three gaps are right next to each other, please count them as three separate gaps for the following exercise. However, in a different model, you could also count gaps right next to each other as one gap, so it wouldn't matter how long the gap is. Both models are used in research.
The decisions made in modeling aim at imitating reality as closely as possible and rely on observations made by biologists. Hence, creating a good alignment program requires close collaboration between computer scientists and biologists.

Example:
string strand1 = "ACGGTG";
string strand2 = "CGGCCG";
Here is a possible alignment:

A	C	G	G	T	-	G
-	C	G	G	C	C	G

_____ match(es), _____ mismatch(es), _____ gap(s) Score: _____

Here is another one:

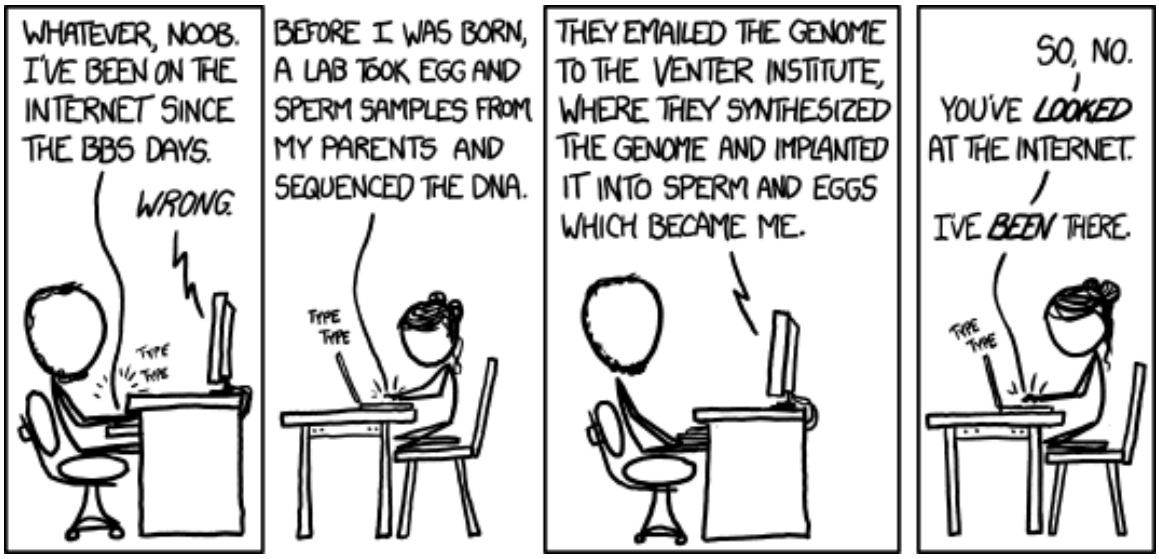
A	C	G	G	T	-	-	G
-	C	-	G	G	C	C	G

_____ match(es), _____ mismatch(es), _____ gap(s) Score: _____

a) Write a program that finds the "score" of the best alignment.
b) Output the alignment with the highest score.
c) What is the Big-Oh of your program?
d) Huge Challenge: Try to optimize your program so it runs faster. Ask your section leader for hints!

Craving for more?
Research paper on Phylo (Yes, you can do research on creating games!): "
A Citizen Science Approach For Improving Multiple Sequence Alignment"
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0031362

Old-Timers
"You were on the internet before I was born? Well, so was I."



(http://xkcd.com/1058/)