Math 2300 **DISCRETE MATH** Spring 2018
Dr. Duval **Project 1**
due Tuesday, February 20

Indel

In this project, we'll look at a simplified model of genetic mutation and species evolution, and see how the language of relations can help us formalize some of the ideas.

Phylogenetics is the study of how species evolved (building the "tree of life"), largely by looking at the genetic code (DNA) of different species. The basic underlying assumption is that mutations happen with small changes in the genetic code. The difficulty is that common ancestors of current species may be extinct, and so it is hard to figure out the sequence of steps that led to current species. For our purposes, we'll consider DNA to be strings of letters chosen from the alphabet $\{A, C, G, T\}$.

One relatively simple thing we can do is to measure the "distance" between two species by counting the number of basic changes needed get from the DNA string of one species to another. For the purposes of this assignment, we'll focus on just two kinds of basic changes: insertion and deletion. An insertion is when a single letter is added to a string, and a deletion is when a single letter is removed from a string. For instance, starting with (an extremely short) string

$$p = ACTGGCTA$$

we could insert a G between the first C and the first T to get a new string

$$q = AC\mathbf{G}TGGCTA$$

or we could remove the last T to get a new string

$$r = ACTGGCA$$

There are sophisticated algorithms to find the shortest sequence of insertions and deletions between a pair of strings. But we'll just do some basics.

Define a relation ins on the set of strings so that $x \,\mathsf{ins}\, y$ when we can obtain y from x by an insertion. Similarly, define a relation del on the set of strings so that $u \,\mathsf{del}\, v$ when we can obtain v from u by a deletion. So, for instance, using the examples above, we can say $p \,\mathsf{ins}\, q$ and $p \,\mathsf{del}\, r$.

1. Find all the ins and del relations among the following strings. (In other words, identify all the pairs x and y among the following strings such that $x \,\mathsf{ins}\, y$, and then identify all the pairs u and v among the following strings such that $u \,\mathsf{del}\, v$.)

$$\begin{array}{ll}
b = CGGCG & d = TCACG \\
e = CCAGCG & f = CGAGCG \\
h = TCCACG & i = TCGGAG \\
j = TCGGCG & k = TCCAGCG \\
l = TCCATCG & m = TCGAGCG \\
n = CTCCAGCG & o = CGTCGAGCG
\end{array}$$

2. Is ins reflexive? symmetric? transitive? anti-symmetric? In each case, explain your reasoning.

3. Is del reflexive? symmetric? transitive? anti-symmetric? In each case, explain your reasoning.

4. Remember, we wanted to measure "distance" between two species (or, in our context, between two strings). A careful discussion of defining distances more generally will be too much for this assignment, but one thing to note is that it is generally symmetric (the distance from point A to point B is the same as the distance from point B to point A).

   We can combine ins and del to make a symmetric relation, called indel as follows:

   $$x \,\mathsf{indel}\, y \text{ if:} \quad x \,\mathsf{ins}\, y \text{ or } x \,\mathsf{del}\, y.$$

   Explain why indel is symmetric. Find all indel relations among the strings in question 1.

5. Is indel transitive? Explain your reasoning.

6. Verify that indel is not reflexive. What new relations would you need to add to indel to make it reflexive? Be sure to explain how your new relation works in general, not just for the strings listed in question 1.