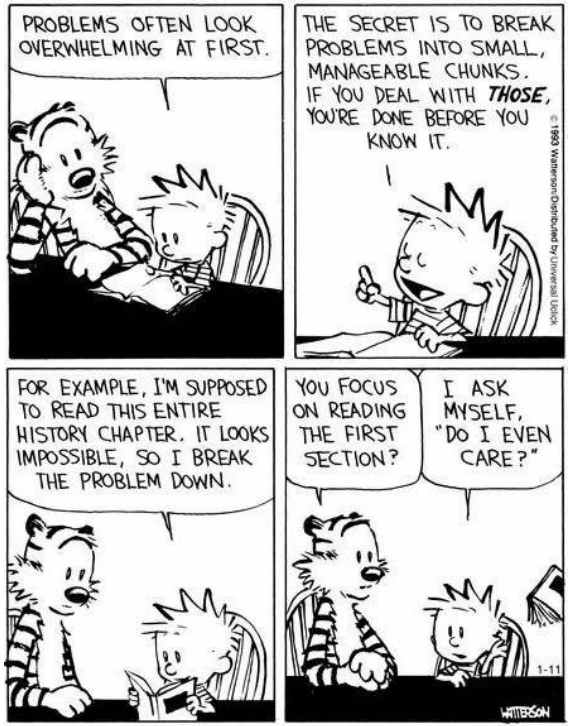
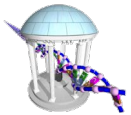
Comp 555 - BioAlgorithms - Spring 2019



- Revised Problem set #4

Problems with Problem Sets

Problem Set #4



Still due next Tuesday…

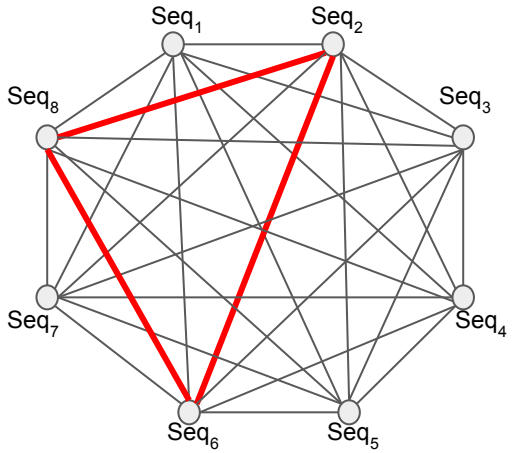
- I have completely changed problems #2, #3, and #4. (Download Version 2)
- Problem #1 is largely intact
- The remainder use that same concepts as the previous version
- But are simplified...

Problem Set #4

Problem 1 is really a graph problem!

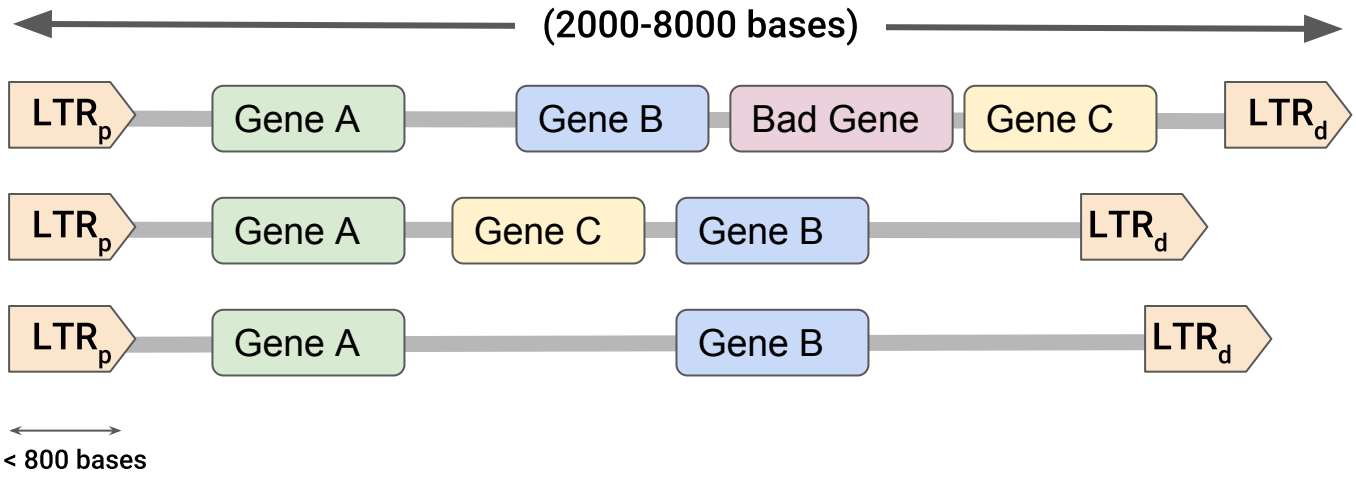
Three parts

1. Get a global aligner to work from class. Strip out all of the backtracking stuff.
2. Find all pairwise alignments (Graph edges) and save them in a normalized dictionary;
   E[i,j] = AligmentScore(seq_i, seq_j), where j > i.
   (Hint: start by using a short prefix to get your code running… ex. first 500 characters of each seq)
3. Find the 3-clique (i,j,k) with the highest edge scores, cliqueScore = E[i,j] + E[i,k] + E[j,k]



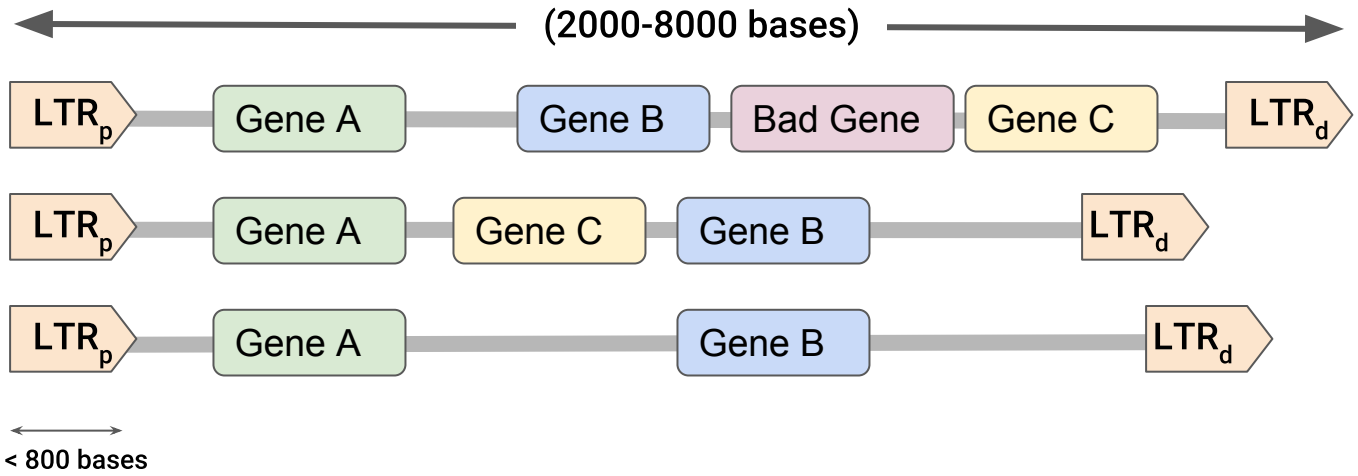
Background

- The 10 sequences I gave you are real endogenous retrovirus (ERV) sequences
- A form of transposable element (jumping gene)
- Viral origin
- TEs make up around 40% of mammalian genomes
- ERVs have a specific Genomic Structure
- Enclosed within a Long-Terminal Repeat sequence (LTR)

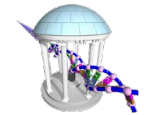


Orignial Problem #2

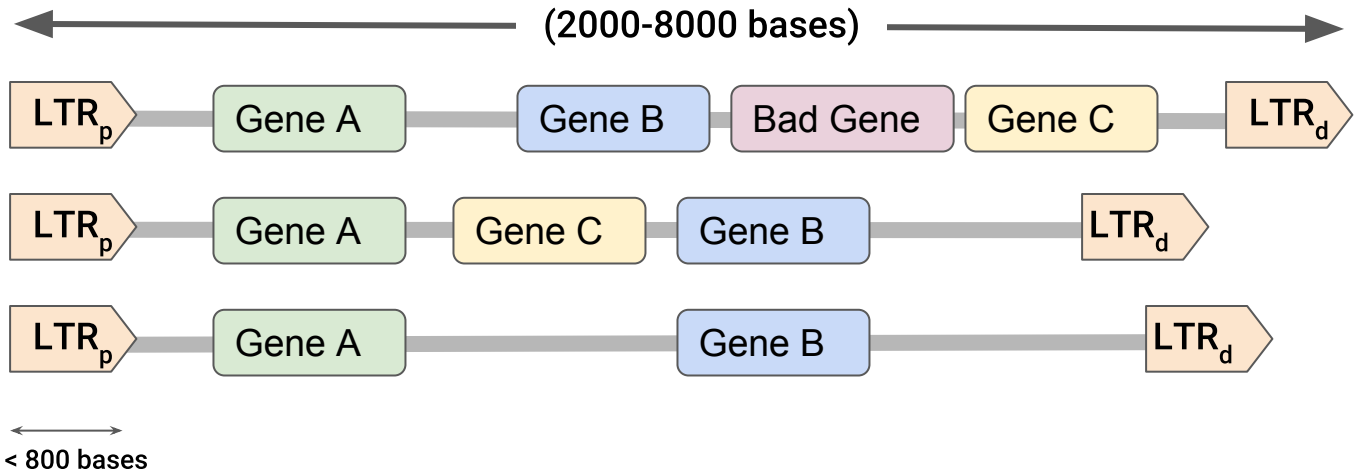
The orignal problem focused on finding the “Essential” genes in the TE’s interior viral sequence. Idea was to find three similar TEs and then find the larger of Gene A, or Gene B.



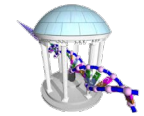
New Problem #2



The new problem focuses on finding the conserved LTR sequence within each of the three closely related genes.

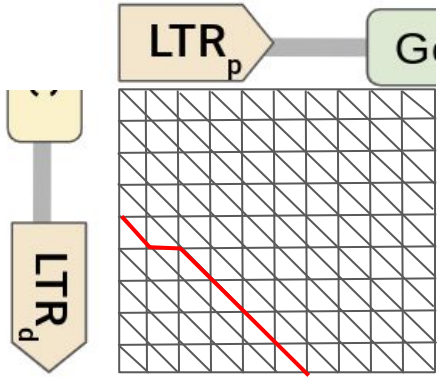


New Problem #2

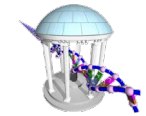


How to do it…

- **Finding the LTR is analogous to finding a Local Alignment of a ERV prefix with a suffix from the same sequence.**
- **Need to find the beginning and end indices in both the prefix and suffix**
- **Repeat for all 3 closely related ERVs**

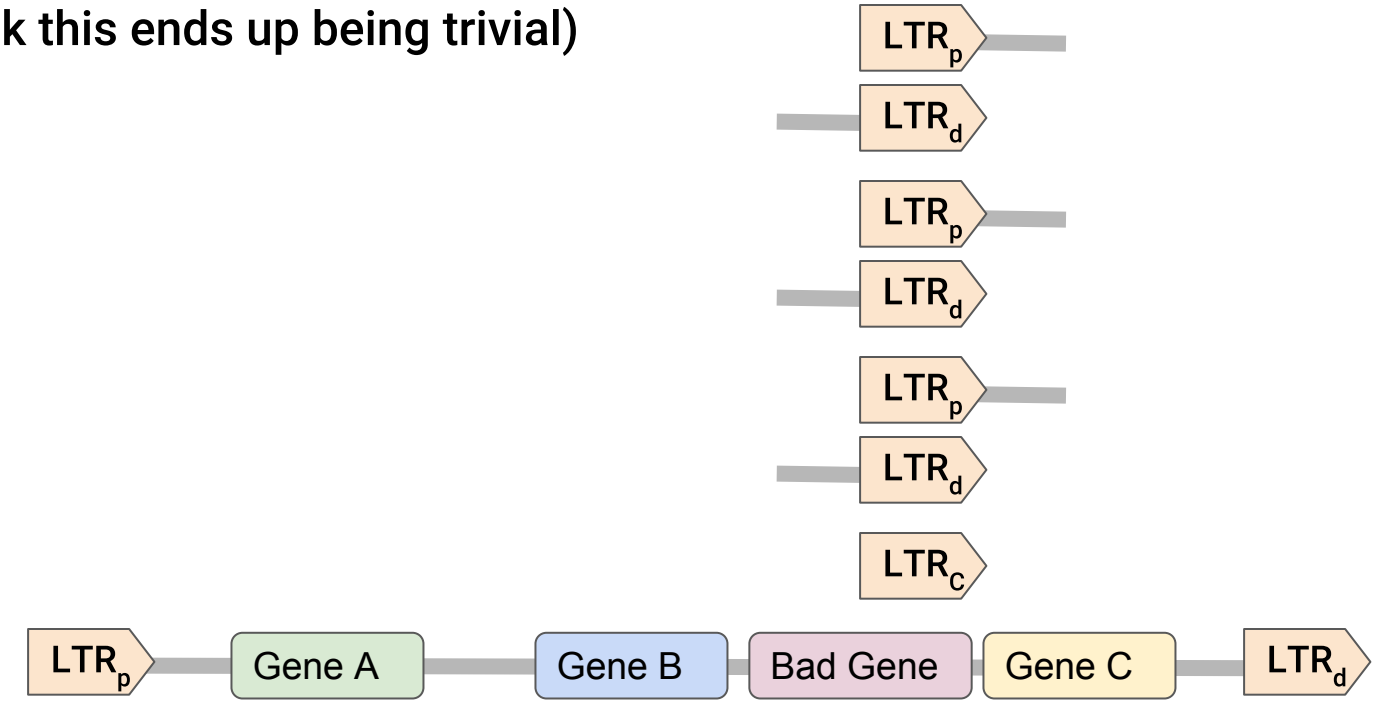


New Problem #3

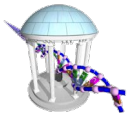


Find a consensus of the 3 LTR alignments

(I think this ends up being trivial)



New Problem #4



Then use a Local Alignment of the consensus LTR to the beginning and ends of each of the original ten sequences, and just keep track of the scores