# Homework 8 – Reconstructing a musical genome

This assignment can be done individually or in pairs. Please turn in only one copy of your results. Do not forget to list both participants if working as a pair.

The attached file, hw8_kmers.xlxs, contains 5mers from a musical text. The list has two columns, the first with the word, and the second with the number of times that 5mer appeared in the text. This list of words was constructed as follows:

1. All punctuation and spaces were removed from the text.
2. 30 letter reads were randomly sampled from the text. My goal was coverage=4. I obtained 96 reads with 716 letters.
3. From the 96 reads, I made a list of all 6mers and alphabetized the list for convenience in looking up words. I found a total of 3882 kmers, which fall into 425 unique types. My average coverage is thus 3882/425 = 9.13.
4. Construct the De Bruijn graph from these reads showing the condensed (unbranched) nodes and the connections between nodes. You do not have to draw a graph, a text version is OK (see the example for one approach; anything that is clear is fine). The condensed nodes in this graph should show all the unbranched k-1 overlaps.
5. Calculate the number of contigs, length of the assembly, and the N50 for your assembly. The length of the assembly is the sum of the lengths of your contigs, not counting duplicate overlapping regions at the ends.
6. Although the graph above contains all the unbranched segments, you may be able to extend the contigs using external knowledge, in this case, your knowledge of English grammar.

## Questions

1. What is the song?
2. What are the number of contigs, length of assembly, and N50 for your assembly?
3. How does your assembly length compare to the length you would predict from the kmers?
4. Why are the kmer estimates of the length and the sum of the lengths of your contigs?