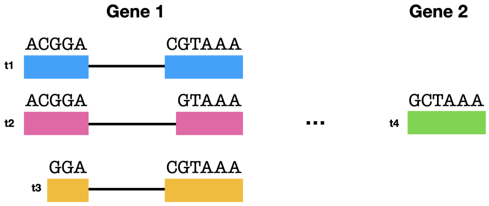
Pseudoalignment

Consider the following transcriptome:

```
>t1_g1
ACGGACGTAAA
>t2_g1
ACGGAGTAA
>t3_g1
GGACGTAAA
>t4_g2
GCTAAA
```

Here is a visual representation of the transcriptome:



(a) Draw the colored de Bruijn graph that one would use for pseudoalignment with $k = 3$. In this particular case, the nodes have 3-mers, not the edges (this is the setup we used in class). Refer to the slides on construction.

(b) Pseudoalign *GGACGT* and show the relevant steps you might take (including skips). Refer to the slides for the pseudoalignment.

(c) Pseudoalign *GGATGT* and show the relevant steps you might take (including skips). Remember, every k-mer has an equivalence class which is a set. If a k-mer is in your data but not in your transcriptome, its equivalence class is the null set.

(d) The previous read might arise if there is an error at position 4 (using 1-based indexing). Describe an algorithm to deal with the error. Describe the benefits and drawbacks of your algorithm. These properties might come in speed, loss of data, or false alignments (or other things). Note: there isn't one correct answer.

(e) The following sequence is a valid RNA-seq read *TTTACG*. Clearly it won't give you a non-empty equivalence class as-is. How did this data arise and how might you pseudoalign it? Hint: think about how RNA-seq is generated. That is, the orientation of the reads matters and can generate some annoying things we have to keep track of.