PanOriginSV

Group 7

Flowchart

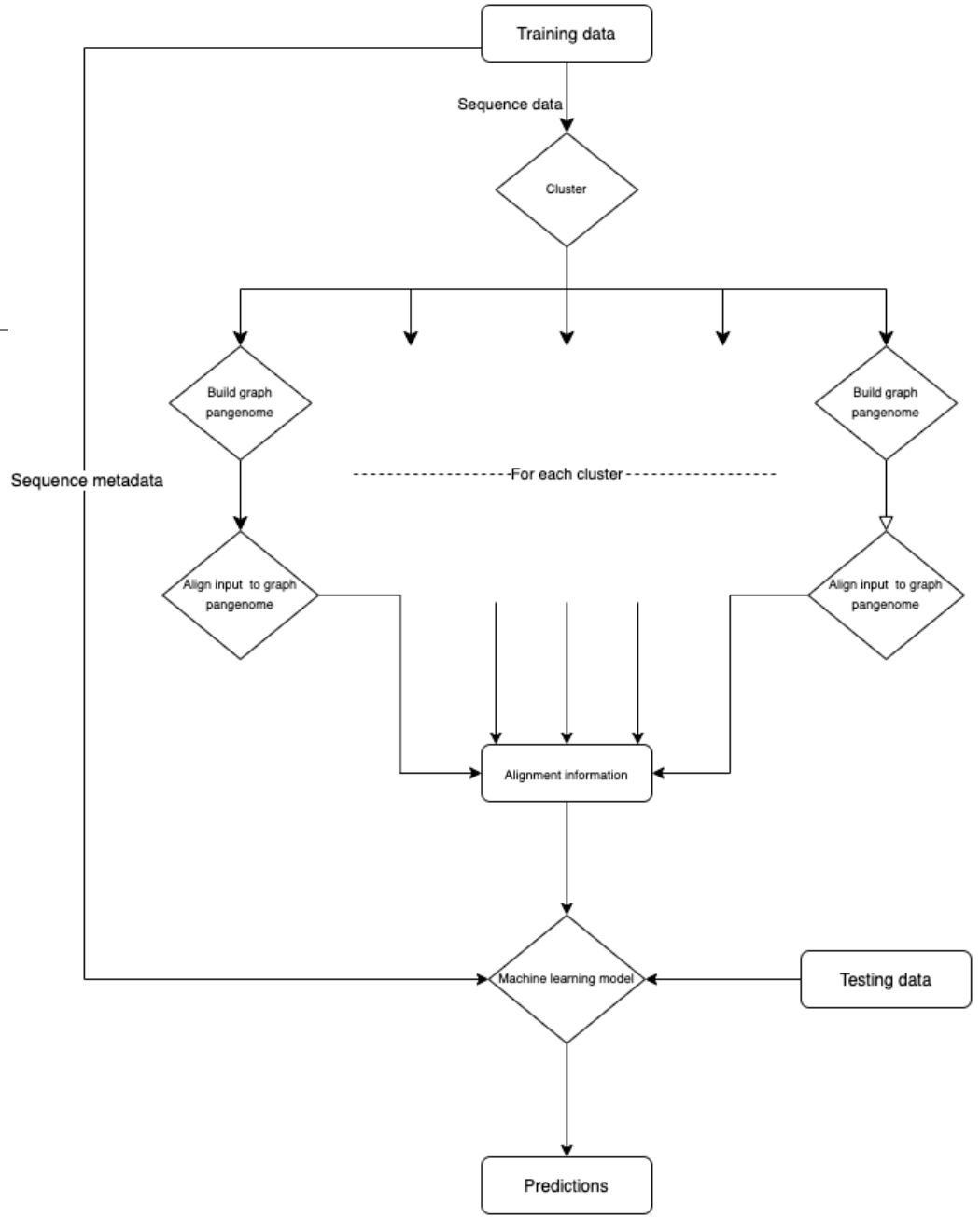
1. Cluster similar training sequences using MMSEQ2

2. For each cluster, create a graph pangenome that incorporates SV information
   - Using bcalm to generate DBG

3. Align the sequences in each cluster to its corresponding pangenome graph
   - Uses minigraph for alignment

4. Use Alignment information and metadata as features to train ML models

5. Predict and benchmark on GEAC dataset



Clustering: MMSEQ2

- mmseqs with `--min-seq-id 0.5 -c 0.8 --cov-mode 1 -s 7.0`
- 13000 clusters (originally 60,000 sequences)
  - 8000 of the clusters are singletons
- Top 10 clusters are above 200 sequences each
- Within a cluster, average nucleotide identity (using FastANI) ranges from 80% to 100%

Benchmarking Against Linear Pangenomes

Cluster	Train Sequences	Test Sequences	Test accuracy (Linear)	Top 5 test accuracy (Linear)	Test accuracy (Graph)	Top 5 test accuracy (Graph)
J7OEM	2870	714	0.96	0.99	0.97	0.99
3PTDM	2397	571	0.82	0.93	0.81	0.93
O3GQU	1046	275	0.65	0.88	0.71	0.83
48073	973	205	0.71	0.89	0.71	0.87
WA905	639	149	0.87	0.94	0.87	0.97
GIGX0	604	119	0.71	0.87	0.79	0.93

Timing data not shown: Graph method uses 10-50x less cpu time

Discussion

- We've shown that by using a graphical pangenome and discarding the order of the alignment to the graph, we achieve comparable results to the linear pangenome method
  - While Top 1 accuracy is slightly higher for the graph method, the Top 5 accuracy is lower. From looking at the training accuracy of Linear vs Graph, it is evident that the graph ML model often overfits the data
- A major issue is ensuring that testing sequences are assigned to clusters that contain training members of the same class. This is a significant issue in the smaller clusters (which there are many of).
- We have yet to try tuning any of the parameters of any of the steps in the pipeline.
- A significant future step would be to train the graph ML model on the sequence of nodes through the graph as opposed to just the bag of nodes

Questions?