SwedishFish

Heuristics for efficient reconstruction of RNA sequences using minimum path decomposition

Dan Wood, Peter Gifford | REU Mentor - Prof. Brendan Mumey, Grad Student Lucy Williams | August 2018

Abstract

We propose two efficient heuristics for reconstructing RNA sequences. The algorithms model the RNA reads using a flow network, and try to decompose the network into the minimum number of path flows. The new heuristics produce results similar to the state of the art with reduced runtime.

Introduction

- RNA sequencing produces short sequencing reads sampled from transcripts
- The process of recovering the full-length sequences, known as the transcript assembly problem, is NP-Hard, so we look for heuristics rather than exact algorithms
- To solve, reads are organized into a de bruijn graph (Figure 1), then a splice graph (Figure 2)
- The minimum path decomposition through the splice graph/flow reveals the full-length sequences

Figure 1: de bruijn graph representing RNA sequence reads

Figure 2: Splice graph generated from the de bruijn graph

Benchmark Algorithms

Greedy Width [3]: Determines which path can carry the largest flow using a dynamic algorithm. It then removes the largest-flow path from the network and repeats until all edges are removed.

Catfish [4]: Uses targeted simplification methods to eliminate potential greedy width errors in the network. Then runs greedy-width with increased accuracy.

Heuristic 1: Greatest Cardinality Cut

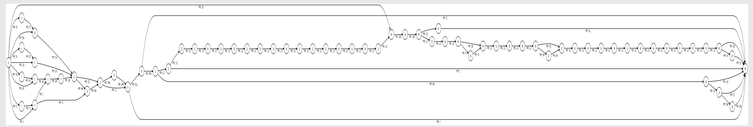

Figure 3: Original network – 62 nodes

- The network is broken down from Figure 1 to Figure 2 through several reversible methods: Serial Edge Removal, Reversal of Subgraphs, and Subset Sums Edge Break Down
- The network is topologically sorted and the greatest cardinality cut is found.
- Let f be the greatest flow value across the cut. We use the longest path algorithm to find a path through the graph such that each edge has flow value greater than or equal to f.
- This process is then repeated until there are no edges left to remove.
- If the greatest cut strategy fails to find a path then greedy width is used in its place for one cycle.

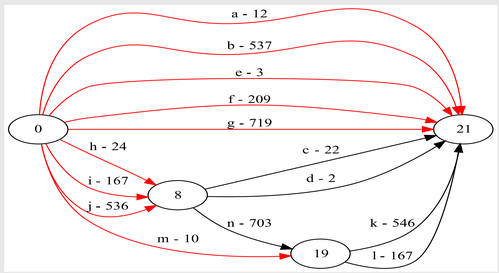

Figure 4: Network after collapse; Greatest cut found by the algorithm is highlighted in red

Heuristic 2: Greedy Edge Remove

- Removes path that will remove the most edges until all edges have been removed
- Tie-breakers that target specific errors make results better than greedy-width, could further improve with more experimentation

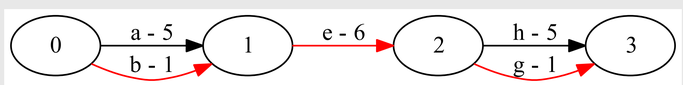

Figure 5: The path highlighted in red will result in 2 edge removals (b and g).

Results

- Tests run on data that models real RNA using Salmon [1] and Flux [2]
- Accuracy is measured by comparing the number of paths generated by each algorithm to the known number of paths (Truth Paths)

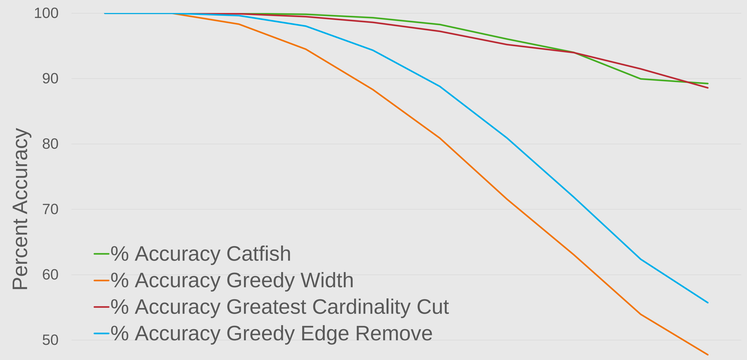

Figure 5: Overall results

Future Work

- Modify Catfish to implement some parts of Greatest Cut
- Modify Greedy Width to implement tie-breaking similar to those used in Greedy Edge Remove
- Implement network reconstruction functionality to show origin paths
- Investigate solving using FPT methods

References

[1] R. Patro, G. Duggal, and C. Kingsford. Salmon: Accurate, versatile and ultrafast quantification from RNA-seq data using lightweight-alignment. bioRxiv, page 021592, 2015.

[2] T. Griebel, B. Zacher, P. Ribeca, E. Raineri, V. Lacroix, R. Guigo, and M. Sammeth. Modelling and simulating generic RNA-seq experiments with the flux simulator. Nucleic Acids Res., 40(20):10073–10083, 2012.

[3] B. Vatinlen, F. Chauvet, P. Chretienne, and P. Mahey. Simple bounds and greedy algorithms for decomposing a flow into a minimal set of paths. Eur. J. Oper. Res., 185(3):1390–1401, 2008.

[4] Shao, M. and Kingsford, C. (2016). Efficient Heuristic for Decomposing a Flow with Minimum Number of Paths.

MONTANA STATE UNIVERSITY