A practical fpt algorithm for Flow Decomposition and transcript assembly

Kyle Kloster, Philipp Kuinke, Michael P. O'Brien, Felix Reidl, Fernando Sánchez Villaamil, Blair D. Sullivan, Andrew van der Poel

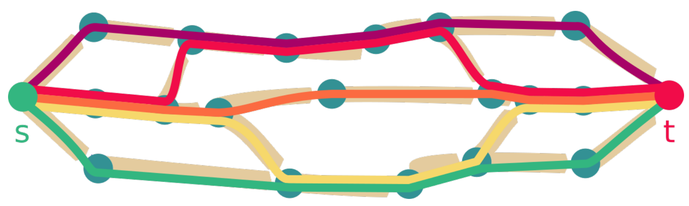
NC STATE UNIVERSITY

RWTH AACHEN UNIVERSITY

The Motivation

Shared segments between DNA/RNA strands create ambiguity in the assembly problem.

Connecting overlapping segments and counting their frequencies yields a DAG and a flow. The problem is to split the flow into the least amount of s-t-paths, to recover the original DNA/RNA strands.



The Problem

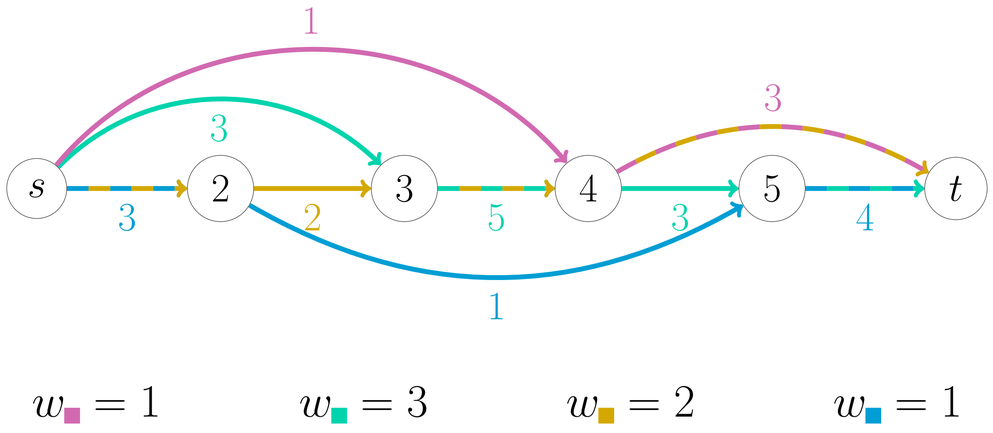
k-Flow Decomposition (*k*-FD)

Input: (G, f, k) with G an s-t–DAG, f a flow on G, and k a positive integer.

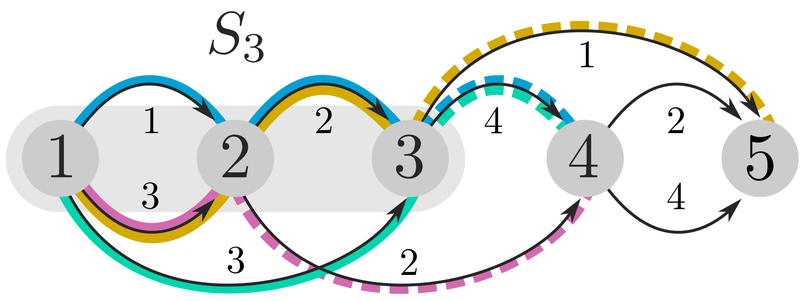
Problem: Is there an integral *flow decomposition* of (G, f) using at most k paths?

About ten years ago, some computer scientists came by and said they heard we have some really cool problems. They showed that the problems are NP-complete and went away!

-Joseph Felsenstein (Biologist)

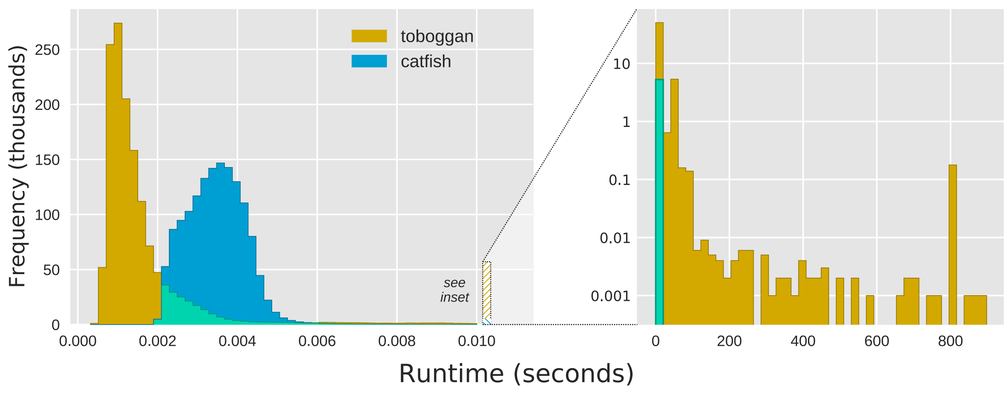


The Algorithm



The routing g out of S_3 (dashed lines) is an extension of the previous routings (solid paths). Each row in the constraint system L on the right corresponds to an arc; those shaded in gray are from arcs inside S_3, and those in white come from g.

The Results



Runtimes of Toboggan and Catfish on all non-trivial instances. The y-axes indicate the number of instances on which the algorithms terminate in the given time window.

dataset	instances	non-trivial	optimal	non-optimal
`zebrafish`	1,549,373	445,880	99.907%	0.053%
`mouse`	1,316,058	473,185	99.401%	0.074%
`human`	1,169,083	529,523	99.490%	0.043%
all	4,034,514	1,448,588	99.589%	0.056%

k	instances	Catfish	Toboggan
2	63.2791%	0.992	**0.995**
3	22.0775%	0.967	**0.969**
4	8.5237%	**0.931**	0.930
5	3.4920%	0.886	0.886
6	1.5375%	**0.830**	0.828
7	0.6698%	**0.788**	0.780
8	0.2889%	**0.767**	0.766
9	0.1241%	0.740	**0.743**
10	0.0070%	0.752	**0.802**
11	0.0004%	0.500	0.500
all	100%	0.973	**0.975**

Since Toboggan finds optimal decompositions we can investigate the Groundtruth for optimality.

All data.

Acknowledgments

This work is supported in part by the Gordon & Betty Moore Foundations Data- Driven Discovery Initiative through Grant GBMF4560 to Blair D. Sullivan.

Resources

The implemented solver is available on Github: `/theoryinpractice/toboggan`