Decomposing inexact flows with application to RNA transcript assembly

Lucia Williams and Brendan Mumey
Gianforte School of Computing, Montana State University

Background

RNA-Seq technology allows for high-throughput, low cost measurement of gene expression. An important step in this process is the assembly of mRNA transcript short reads into full transcripts. We give a **new theoretical formulation** of this problem that better models uncertainty in the short read measurements and an **efficient heuristic algorithm** to solve it**.**

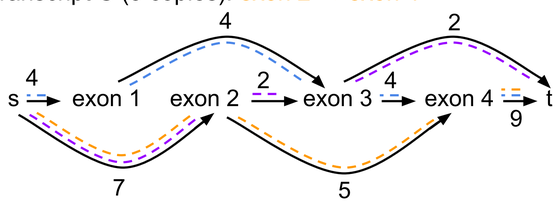
The RNA-Seq assembly problem is analogous to taking many copies of many different but similar books, chopping each one up into pieces of different lengths, and then trying to recover all of the original texts using only the pieces.

RNA-Seq transcripts and their abundances can be represented using a fundamental computer science tool called a **flow network**.

Example

Gene: CGAAGCGCAG GCTGTATTCTTATTCCCGA AGTCTCCGTGC GATCCGAC
exon 1 exon 2 exon 3 exon 4

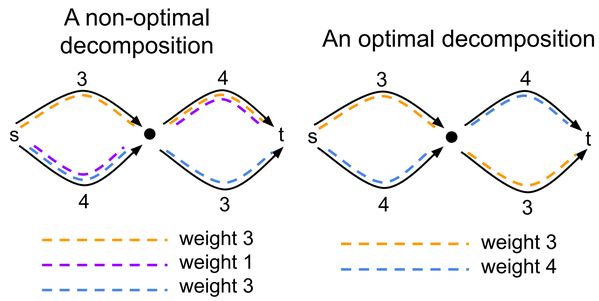
Transcript A (4 copies): exon 1 → exon 3 → exon 4
Transcript B (2 copies): exon 2 → exon 3
Transcript C (5 copies): exon 2 → exon 4



Each transcript is a path from start node s to end node t. Figure adapted from [1].

References
[1] M. Shao and C. Kingsford, "Theory and a heuristic for the minimum path flow decomposition problem," IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017.
[2] A. Tomescu, A. Kuosmanen, R. Rizzi, and V. Makinen: A novel min-cost flow method for estimating transcript expression with rna-seq. BMC Bioinformatics. vol. 14, p. S15. BioMed Central (2013)
[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," European Journal of Operational Research, vol. 185, no. 3, pp. 1390–1401, 2008.

Previous work finds the transcripts and their abundances by **decomposing a flow network into a minimal set of weighted paths** according the principle of parsimony: the simplest explanation for the data is the best.

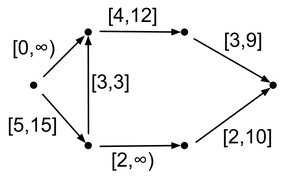


But when we generate these networks from short read data, they will not perfectly represent the true transcripts [2].

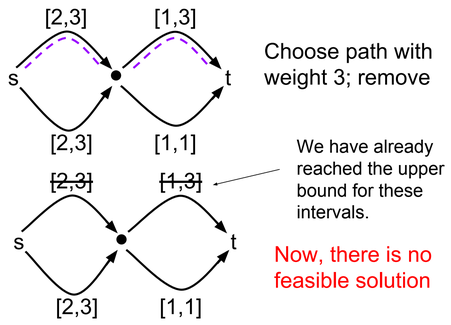
New Theoretical Formulation

We define an **Inexact Flow Network**. Let $G=(V,E)$ be a network with start node s and end node t. Each edge e has an associated flow interval I_e, which can take two forms:

- Bounded intervals: $[l_e, u_b]$, where $0 \le l_e \le u_e$.
- Unbounded intervals: $[l_e, \infty)$, where $0 \le l_e$.



The **Inexact Flow Decomposition** (IFD) problem asks for the minimal set of paths that satisfies all edge constraints. Successively removing paths will always yield a valid solution for exact flow decomposition [3], but this approach can cause problems for IFD. For example:



IFD Algorithm

Our approach has 3 steps:

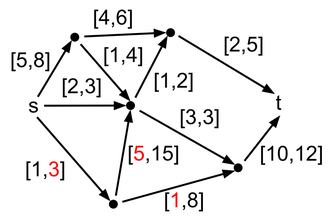
1. Transform the IFD instance into an exact network flow instance.
2. Decompose that exact network flow into a set of paths using an existing method.
3. Refine the set of paths using a new method called *path splicing*.

We can replace all infinite upper bounds with finite values, using knowledge from other edges.

Step 1: IFD Transformation

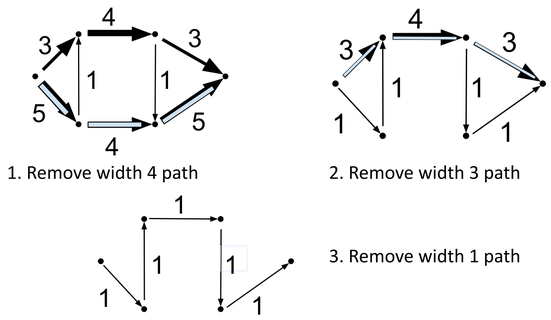
We create and solve a maximum flow problem to find a feasible network flow (if there is one) from the IFD instance.

An infeasible IFD instance. For example, the red upper and lower bounds are incompatible.



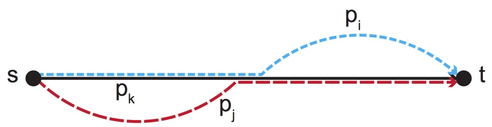
Step 2: Path Decomposition

A common approach to decompose a flow network is called *greedy width*.



Step 3: Splice paths

Once we have a path decomposition we can improve it by *splicing paths.*

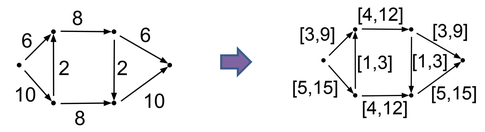


Path splicing turns three paths into two. Here, p_i follows p_k, then p_j follows p_k, and p_i and p_j overlap in the middle.

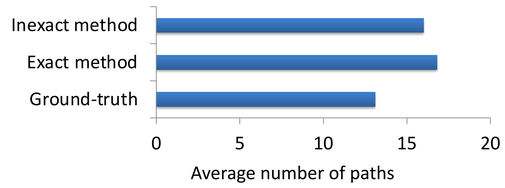
Experiments

We implemented our algorithm in Python.

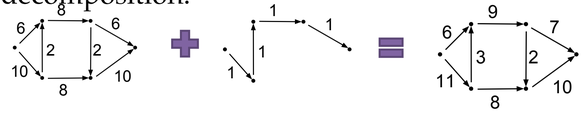
We tested it on 120k graph instances simulated from real transcript data superimposed to form a flow network. We add a 50% error bound to each edge for the inexact method.



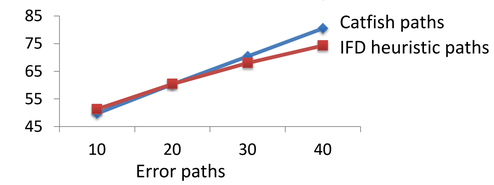
IFD heuristic finds smaller decompositions than exact methods for biological dataset



We also tested on synthetic data sets with many nodes and edges. Here, we added random "error paths" to the graph and compared our algorithm to the current best algorithm (Catfish, [1]) for flow decomposition.



IFD heuristic finds smaller decompositions than Catfish for >30 error paths



Conclusion & Future Work

We give a new formulation for the RNA assembly problem and propose a heuristic to solve it. We would like to extend this work by:

1. Designing and running additional experiments to explore the biological relevance of IFD.
2. Proposing an algorithm to take raw short read data and convert it to an IFD instance.
3. Finding additional algorithms to solve IFD.

This research is supported by the the NSF under grant no. DBI-1759522 and National Institute of General Medical Sciences of the NIH under Award Number P20GM103474.

MONTANA STATE UNIVERSITY