Reference-Based Transcriptome Assembly

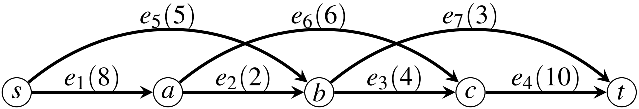
Mingfu Shao

Computational Biology Department, Carnegie Mellon University

April 2, 2016

Formulation

- **Input:** fully connected directed acyclic graph $G = (V, E)$ with source s and sink t, and flow vector f.



$$M = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{array} \begin{array}{c} \begin{array}{ccccccc} e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} \end{array}$$

- **Output:** $P \subset R(M)$ **and** vector s such that $f = s \cdot P$ and that $|R(P)|$ is minimized.

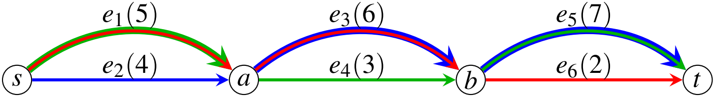
Facts

Let $\varDelta = |E| - |V| + 2$. Let (P^*, s^*) be any optimal solution.

- **Fact 1:** $rank(M) = \varDelta$.
- **Fact 2:** $rank(P^*) = |R(P^*)|$.
- **Fact 3:** $|R(P^*)| \leq \varDelta$.
- **Fact 4:** If $|R(M)| = \varDelta$, then the solution is unique: (M, s), where s is determined by $f = s \cdot M$. **[trivial cases]**
- **Fact 5:** If $|R(P^*)| = \varDelta$, then greedy algorithm is guarenteed to give optimal solution. **[easy cases]**
- **Degenerated cases:** $|R(P^*)| < \varDelta$. **[hard cases]**

Degeneration Theorem

- **Theorem:** there exist $k = \Delta - |R(P^*)|$ linearly independent *non-trivial* vectors $q_1, \cdots, q_k$ satisfying $f \cdot q_i = 0$, $1 \leq i \leq k$.



$$\begin{array}{cc} & \begin{array}{cccccc} e_1 & e_2 & e_3 & e_4 & e_5 & e_6 \end{array} \\ \begin{array}{c} p_1 \\ p_2 \\ p_3 \end{array} & \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \end{pmatrix} \end{array} \begin{array}{c} \begin{array}{ccc} q_1 & q_2 & q_3 \end{array} \\ \begin{pmatrix} +1 & 0 & +1 \\ +1 & 0 & 0 \\ -1 & +1 & 0 \\ -1 & +1 & -1 \\ 0 & -1 & 0 \\ 0 & -1 & -1 \end{pmatrix} \end{array} = 0$$

- Consider the **null space** of P^*, i.e., $N(P^*) = \{q | P^* \cdot q = 0\}$.
- For any $q \in N(P^*)$, we have $f \cdot q = s \cdot P \cdot q = 0$.
- $\dim(N(P^*)) = |E| - rank(P^*) = |E| - |R(P^*)|$.

Identifying Equations

- **Conjecture:** $q_i \in \{+1, -1, 0\}$.
- Only consider two simple forms of equations in the current implementation:

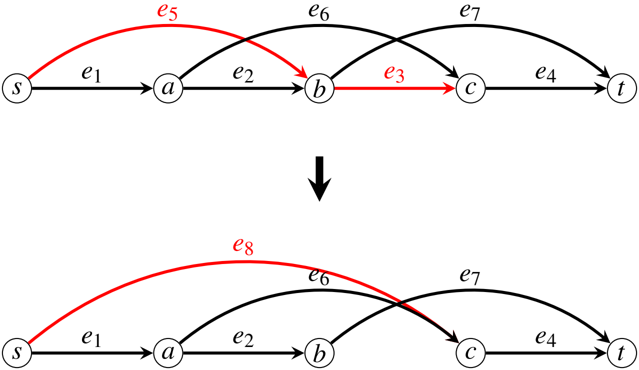
$$f_i = f_{i_1} + f_{i_2} + \cdots + f_{i_k} \tag{a}$$

$$f_i + f_j = f_{i_1} + f_{i_2} + \cdots + f_{i_k} \tag{b}$$

- **Algorithm:** Use the existing pseudo-polynomial-time algorithm for the subset-sum problem.
- For equation of form (a): split e_i into k edges, each with flow value of f_{i_k}; record these k pairs of edges with equal flow.
- For equation of form (b): use heuristics to split both sides into identical set of edges; record these pairs with equal flow.

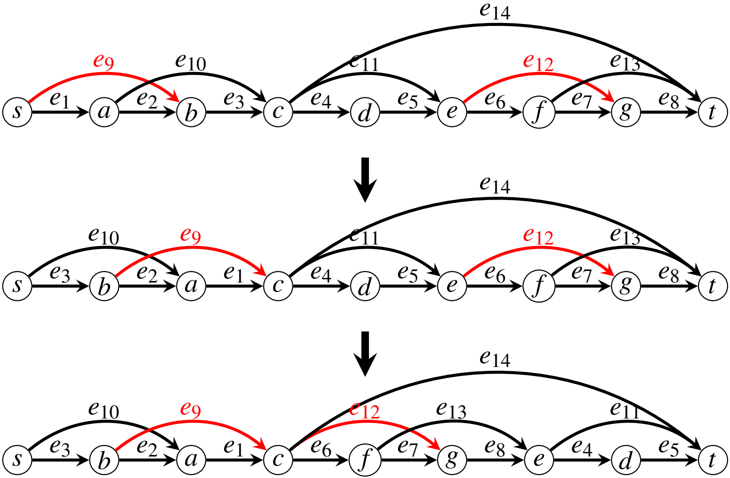
Merge Adjacent Edges with Equal Flow

- **Algorithm:** merge them directly.



Merge Distant Edges with Equal Flow

- **Algorithm: inverse** and **swap** subgraphs based on the its (partial) **nested** structure.


e_{14}
e_9
e_{10}
e_{11}
e_{12}
e_{13}
s e_1 a e_2 b e_3 c e_4 d e_5 e e_6 f e_7 g e_8 t
e_{14}
e_{10}
e_9
e_{11}
e_{12}
e_{13}
s e_3 b e_2 a e_1 c e_4 d e_5 e e_6 f e_7 g e_8 t
e_{14}
e_{10}
e_9
e_{12}
e_{13}
e_{11}
s e_3 b e_2 a e_1 c e_6 f e_7 g e_8 e e_4 d e_5 t

Algorithm

1. Iterate until no change is made to the splice graph. **[core]**

    a. Decompose trivial vertices

    b. Identify type (a) equation and split it

    c. Merge equal edges that can be made adjacent

    d. Merge equal edges that can not be made adjacent (*)

    e. Identify type (b) equation and split it (*)

2. Arbitrarily decompose the remaining graph using greedy algorithm (but optimal solution is not unique).

0. For cases that the estimated flow value are not perfect, we can use LP to correct them when identify equations.

Simulation Results

- Use iGenome annotation (gtf file) of Human genome.
- Use Flux Simulator to only simulate expression.
- Average over 100 independent 100 instances.

	genes	trivial	easy	hard	greedy	scallop
average	11374	8203	129	3036	99.3	**7.8**
ratio(%)	100	72.2	1.1	26.7	3.3	**0.26**

Table: Capability of returning minimized number of paths.

algorithm	gene-level			transcript-level		
	correct	predicted	ratio	correct	predicted	ratio
core	2623	2725	95.6	6763	7029	**96.2**
scallop	2683	3035	88.4	7608	8924	85.2
greedy	2617	3035	86.2	7304	9037	80.8

Table: Accuracy of predicted transcripts.

Next Steps

- Prove/disprove the conjecture.
- Design better algorithms to use type (b) and even more complicated equations.
- Work on generating high quality splice graph (identify exons and estimate flow values) from sequence data.
- Think about applying this algorithm to othe problems, for example, to help the EM-step of the quantification algorithm.