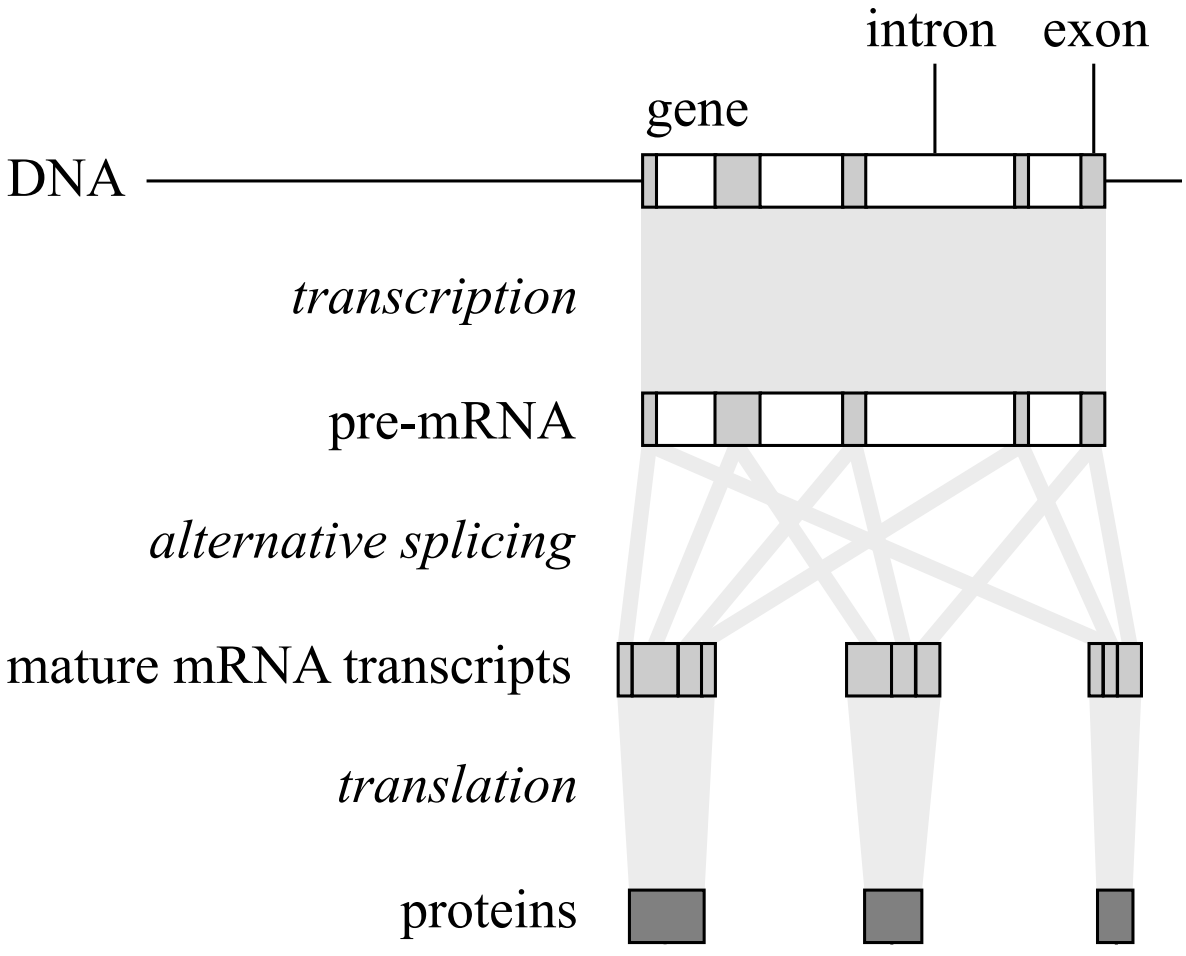
# Multi-Assembly Problems for RNA Transcripts

Alexandru Tomescu
*Department of Computer Science*
*University of Helsinki*

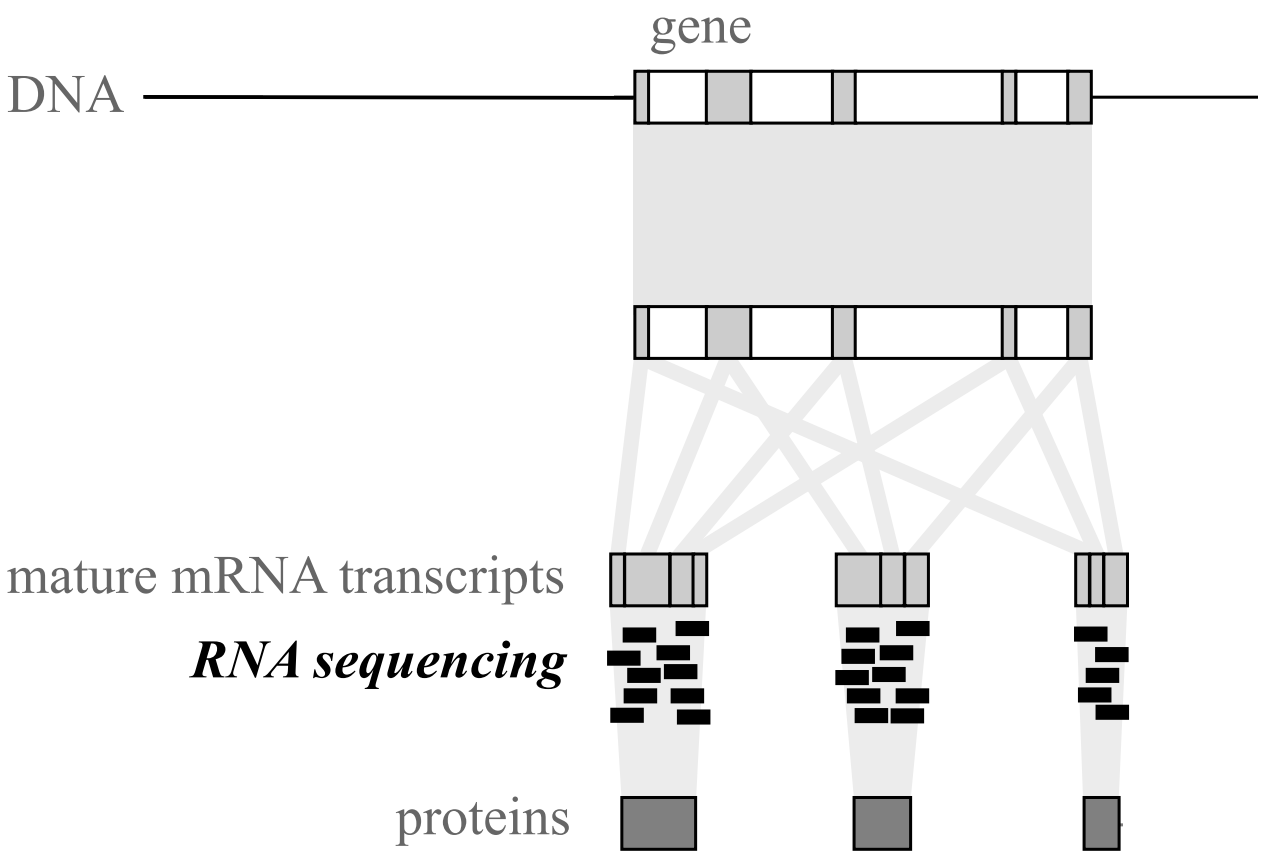
Joint work with
Veli Mäkinen, Anna Kuosmanen, Romeo Rizzi,
Travis Gagie, Alex Popa

CiE
July 3, 2015

# CENTRAL DOGMA OF MOLECULAR BIOLOGY



# RNA-Sequencing




Problem: assemble the RNA transcripts from the RNA-Seq reads and quantify their expression levels

# MULTI-ASSEMBLY

Assembly of fragments from different, but related, sequences

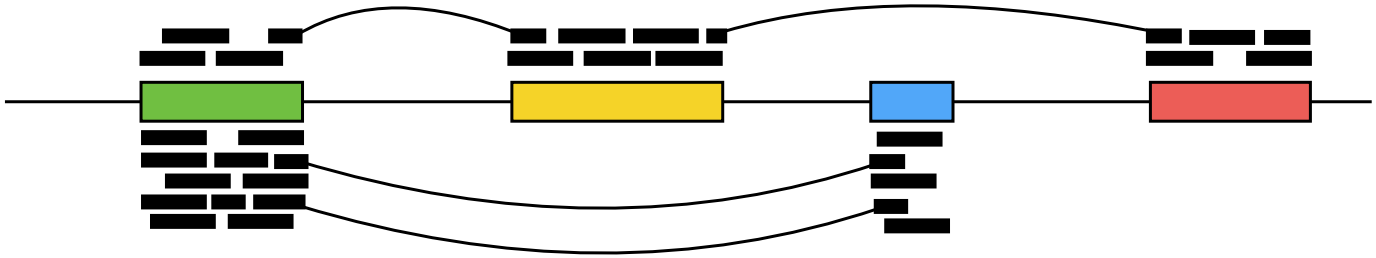
- transcriptomics (RNA-Seq)
- viral quasi-species
- metagenomics

Assumptions:

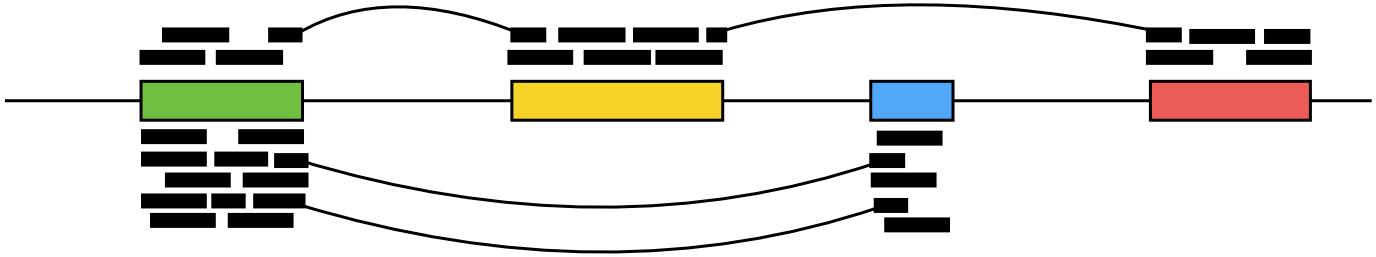
✔ existing reference (genome-guided multi-assembly)

✖ no existing annotation

# Splicing graphs



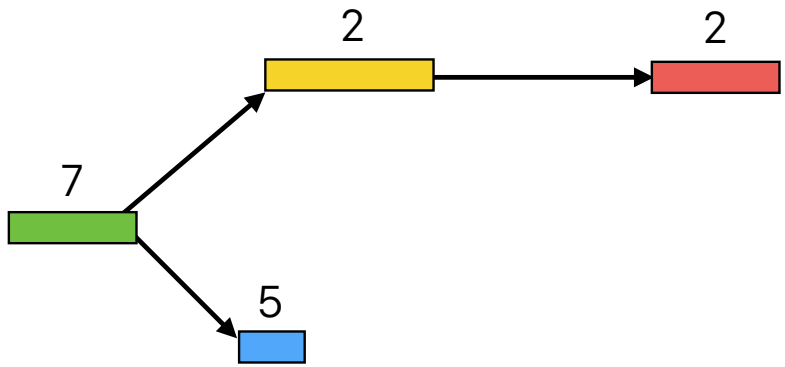
# Splicing graphs



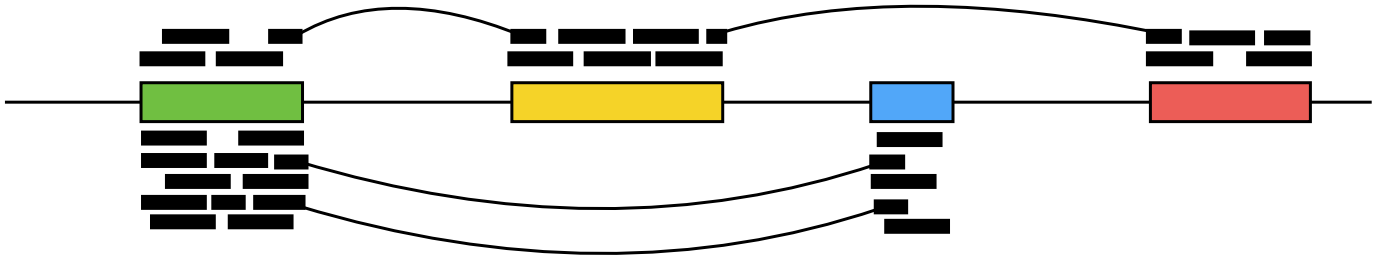
Splicing graphs:

- exons $\equiv$ nodes
- reads overlapping two exons $\equiv$ arcs
- $+$ coverage information

Existing reference $\Longrightarrow$ directed acyclic graphs (DAGs)




# Overlap graphs



- reads $\equiv$ nodes
- overlaps $\equiv$ arcs
- $+$ coverage information

Existing reference $\Longrightarrow$ directed acyclic graphs (DAGs)

# Outline of the talk

Three problem formulations:

1. Assembly only
2. Simultaneous assembly and estimation of expression levels
3. Assembly only, with long reads, or paired-end reads

# Outline of the talk

Three problem formulations:

1. **Assembly only**
2. Simultaneous assembly and estimation of expression levels
3. Assembly only, with long reads, or paired-end reads

# Assembly: Minimum Path Cover (MPC)

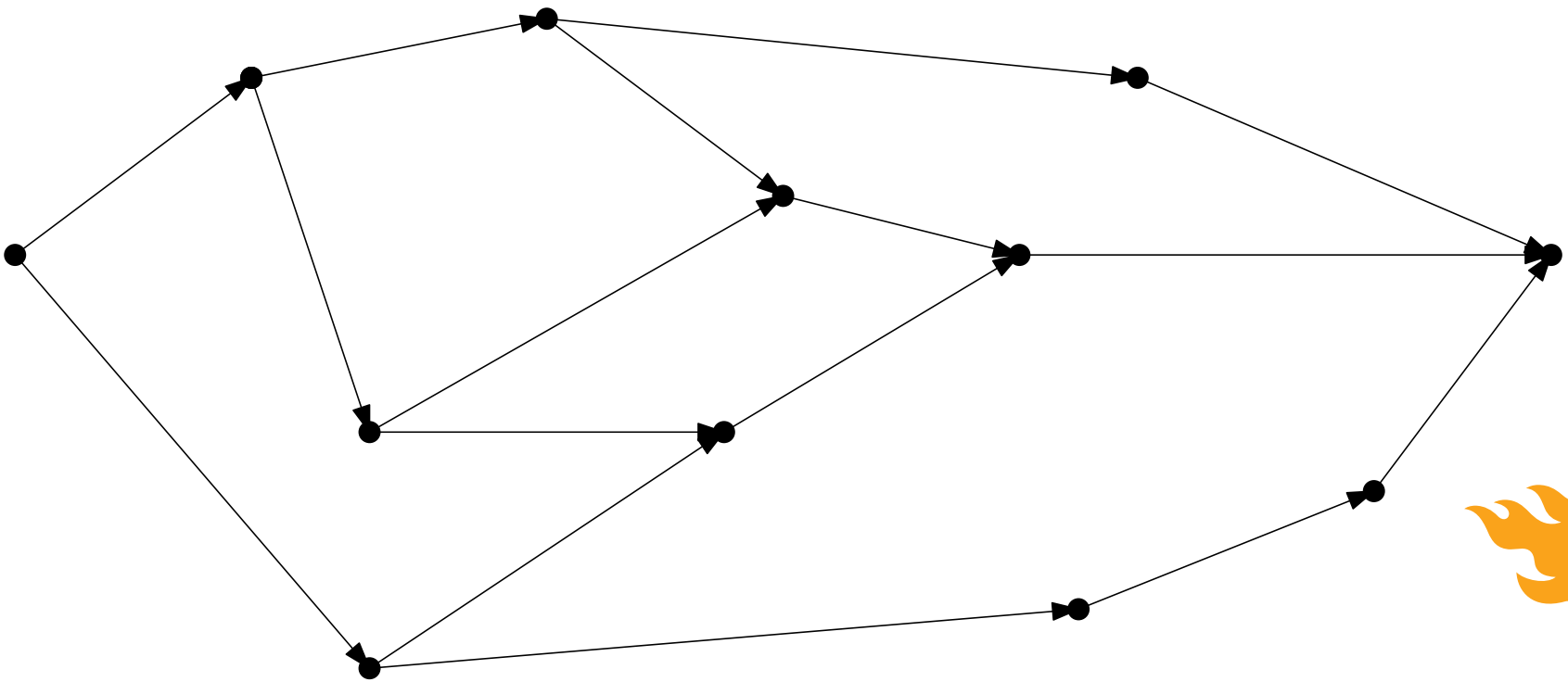
What is the minimum number of paths required to cover all nodes of a DAG?

- RNA-Seq: Cufflinks 2010, CLASS 2012, BRANCH 2013
- Viral quasi-species: ShoRAH 2011

# Assembly: Minimum Path Cover (MPC)

What is the minimum number of paths required to cover all nodes of a DAG?

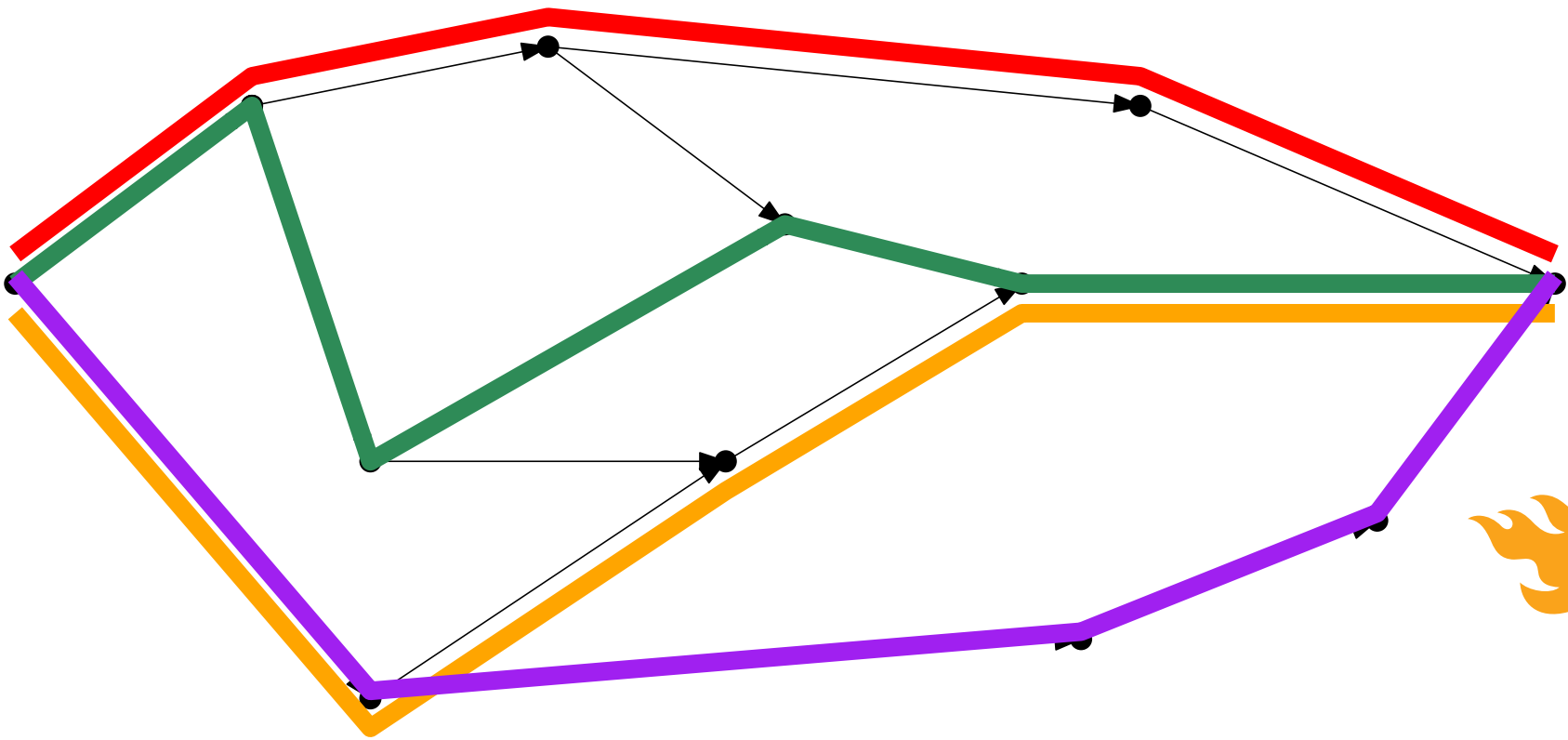
- RNA-Seq: Cufflinks 2010, CLASS 2012, BRANCH 2013
- Viral quasi-species: ShoRAH 2011



# Assembly: Minimum Path Cover (MPC)

What is the minimum number of paths required to cover all nodes of a DAG?

- RNA-Seq: Cufflinks 2010, CLASS 2012, BRANCH 2013
- Viral quasi-species: ShoRAH 2011



# Assembly: Minimum Path Cover (MPC)

In general it is NP-hard (one path iff $G$ has a Hamiltonian path)

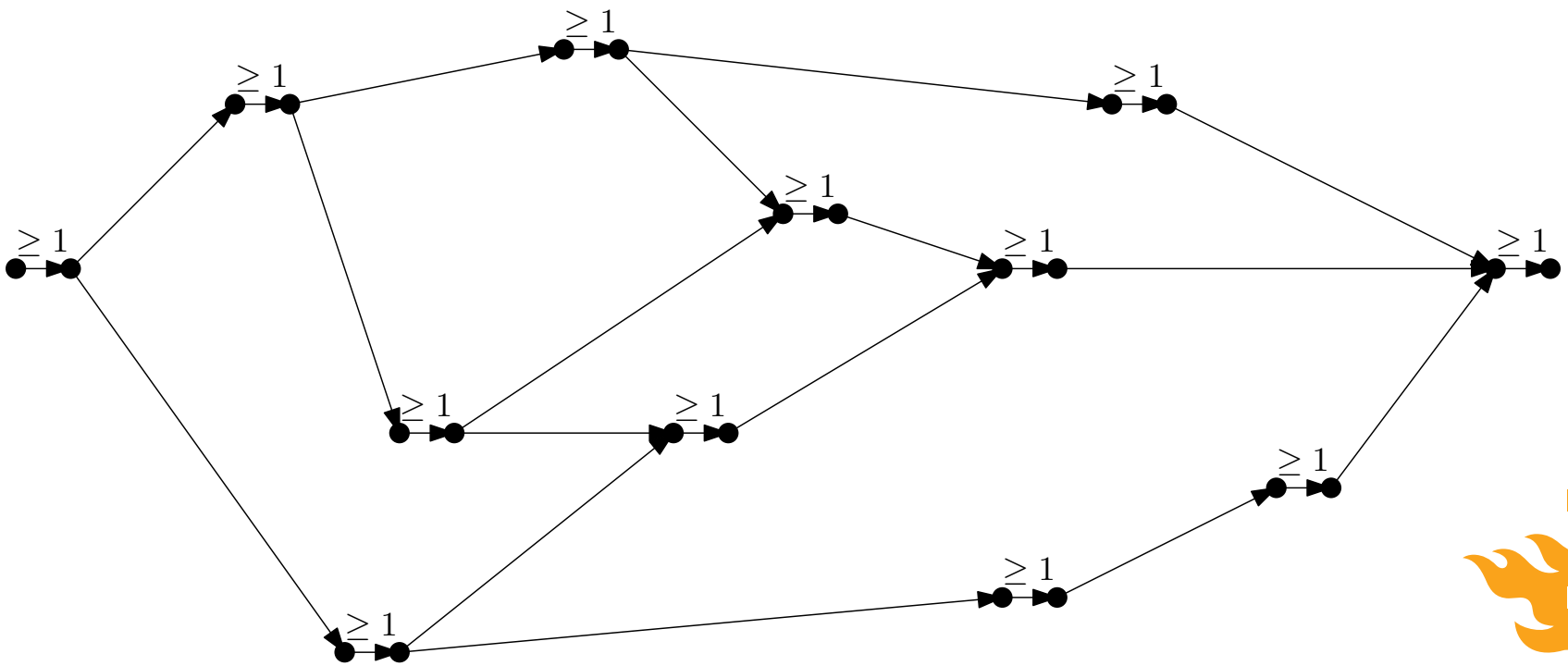
But it is solvable in polynomial-time on DAGs:

- Dilworth's theorem 1950 + Fulkerson's constructive proof 1956
- by a maximum matching algorithm, solvable in time $O(t(G)\sqrt{n})$
- the weighted version can be solved in time $O(n^2 \log n + t(G)n)$

where $m \leq t(G) \leq n^2$ is #arcs in the transitive closure of $G$.

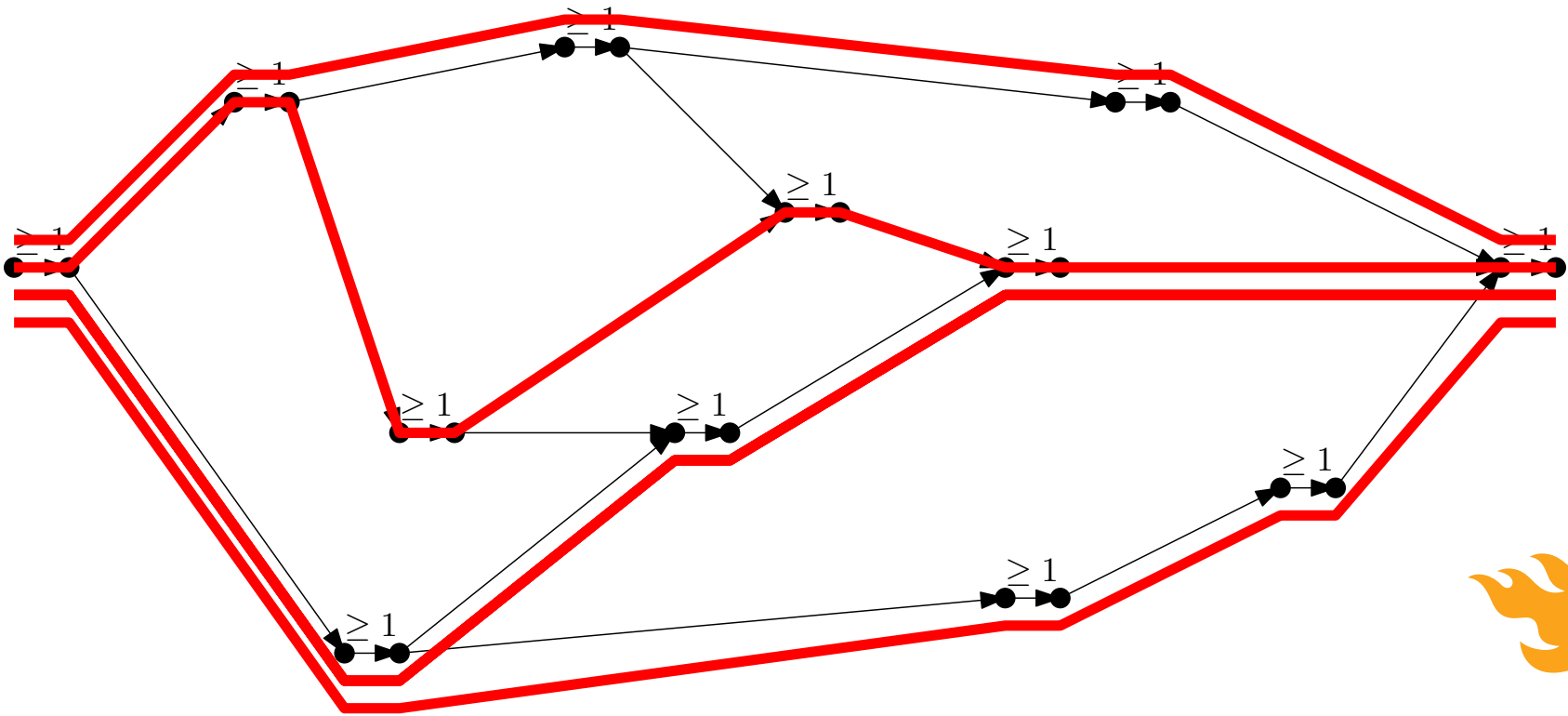
# Min-cost MPC via Min-cost flows

- Unweighted case: MPC via min-flows, e.g. [*Pijls, Potharst, 2013*]
- Weighted case: MPC via min-cost flows



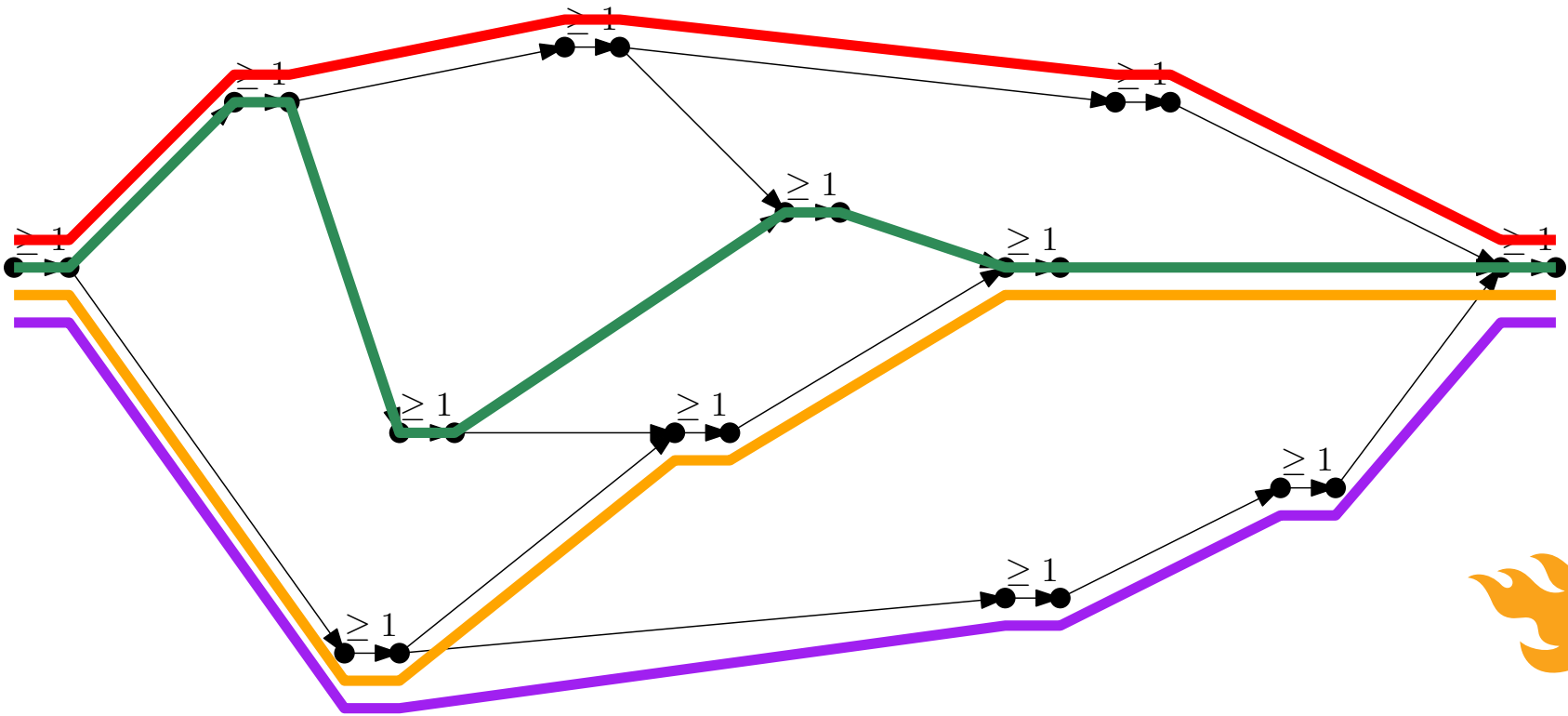
# Min-cost MPC via min-cost flows

- Unweighted case: MPC via min-flows, [*Pijls, Potharst, 2013*]
- Weighted case: MPC via min-cost flows



# Min-cost MPC via min-cost flows

- Unweighted case: MPC via min-flows, [*Pijls, Potharst, 2013*]
- Weighted case: MPC via min-cost flows



# MPC VIA MIN-COST FLOWS

This min-cost flow problem

- can be solved in time $O(n^2 \log n + nm)$ by [*Gabow and Tarjan, 1991*]
- observed in [*Rizzi, T., Mäkinen, 2014*]

This is better than $O(n^2 \log n + nt(G))$, since $m \leq t(G) \leq n^2$

- as soon as there is a path of length $O(n)$, we have $t(G) = O(n^2)$

# Outline of the talk

Three problem formulations:

1. Assembly only
2. **Simultaneous assembly and estimation of expression levels**
3. Assembly only, with long reads, or paired-end reads

# Assembly and estimation of expression levels

**INPUT:** An arc-weighted DAG $G$, and

- A superset $S$ of the sources, and a superset $T$ of the sinks

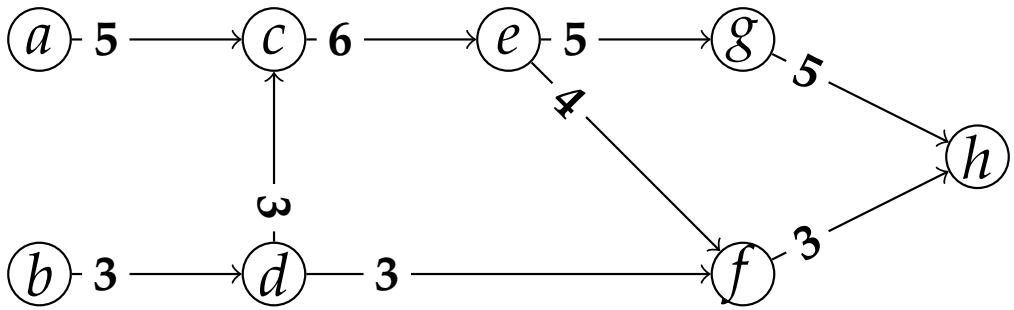
**TASK:** Find a collection of paths $P_1, \ldots, P_k$ in $G$, and their expression levels $e_1, \ldots, e_k$, such that:

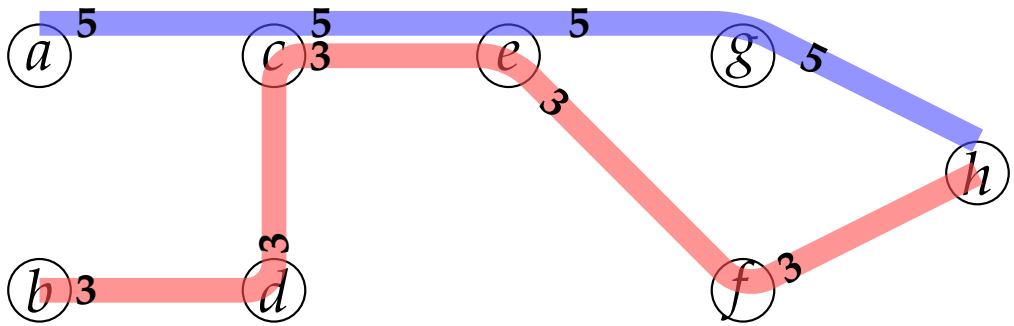
- every $P_i$ starts in $S$, and ends in $T$, and
- the following cost is minimized

$$\sum_{(x,y)\in E} \left| w(x,y) - \sum_{j \,:\, (x,y)\in P_j} e_j \right| .$$

Variants for RNA-Seq in: IsoInfer 2010, IsoLasso 2011, CLIIQ 2012, FlipFlop 2014

# Assembly and estimation of expression levels




Cost is $|6-8|+|3-0|+|4-3|$

# Assembly and estimation of expression levels

Previous solutions based on enumeration of all paths (+ILP)

Solvable in polynomial-time by min-cost flows

- [*T., Kuosmanen, Rizzi, Mäkinen, 2013*]

If number $k$ of paths is given in input, then NP-hard
But solvable in time $O(W^k \mathsf{aw}(G)^k n^2)$

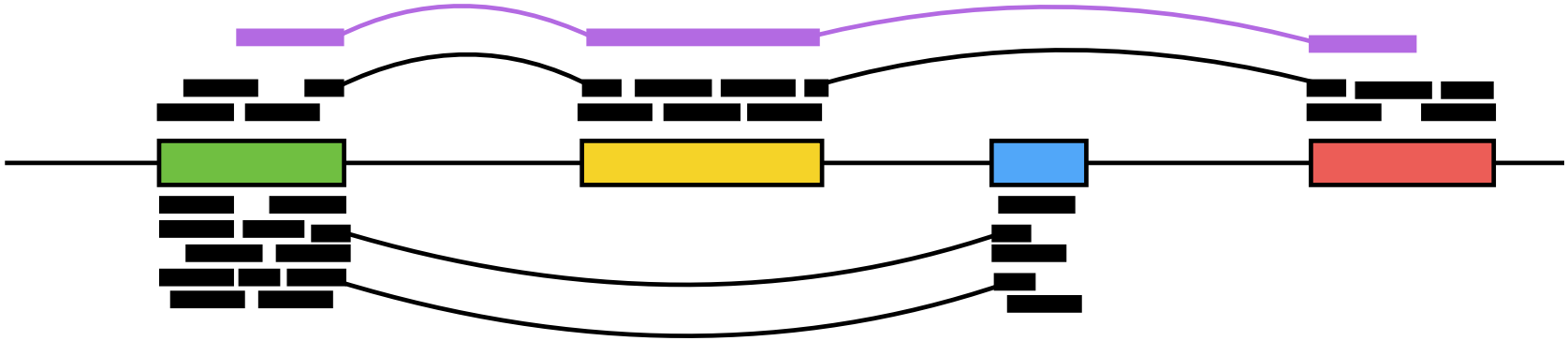
- [*T., Gagie, Popa, Rizzi, Kuosmanen, Mäkinen, 2015*]

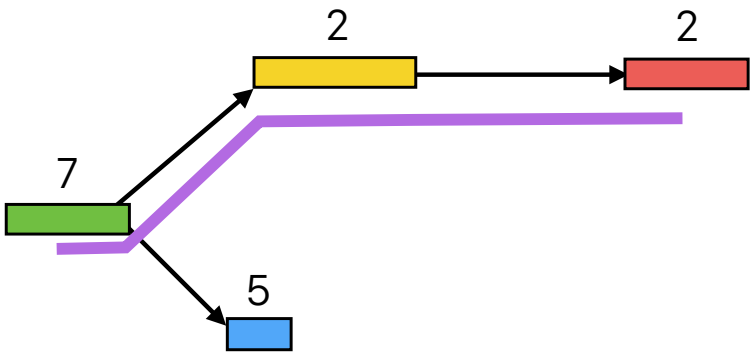
# Outline of the talk

Three problem formulations:

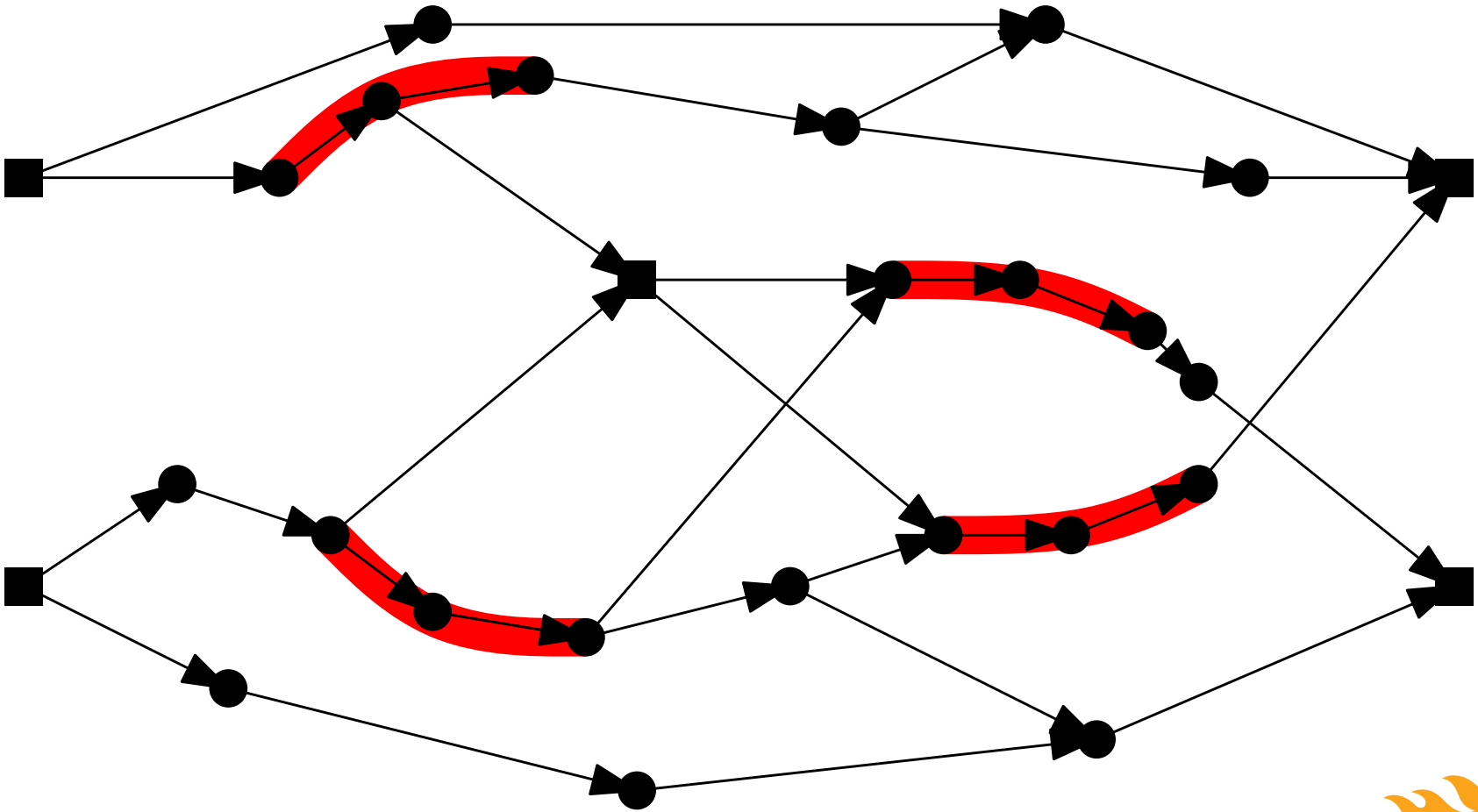
1. Assembly only
2. Simultaneous assembly and estimation of expression levels
3. **Assembly only, with long reads, or paired-end reads**

# ASSEMBLY WITH LONG READS

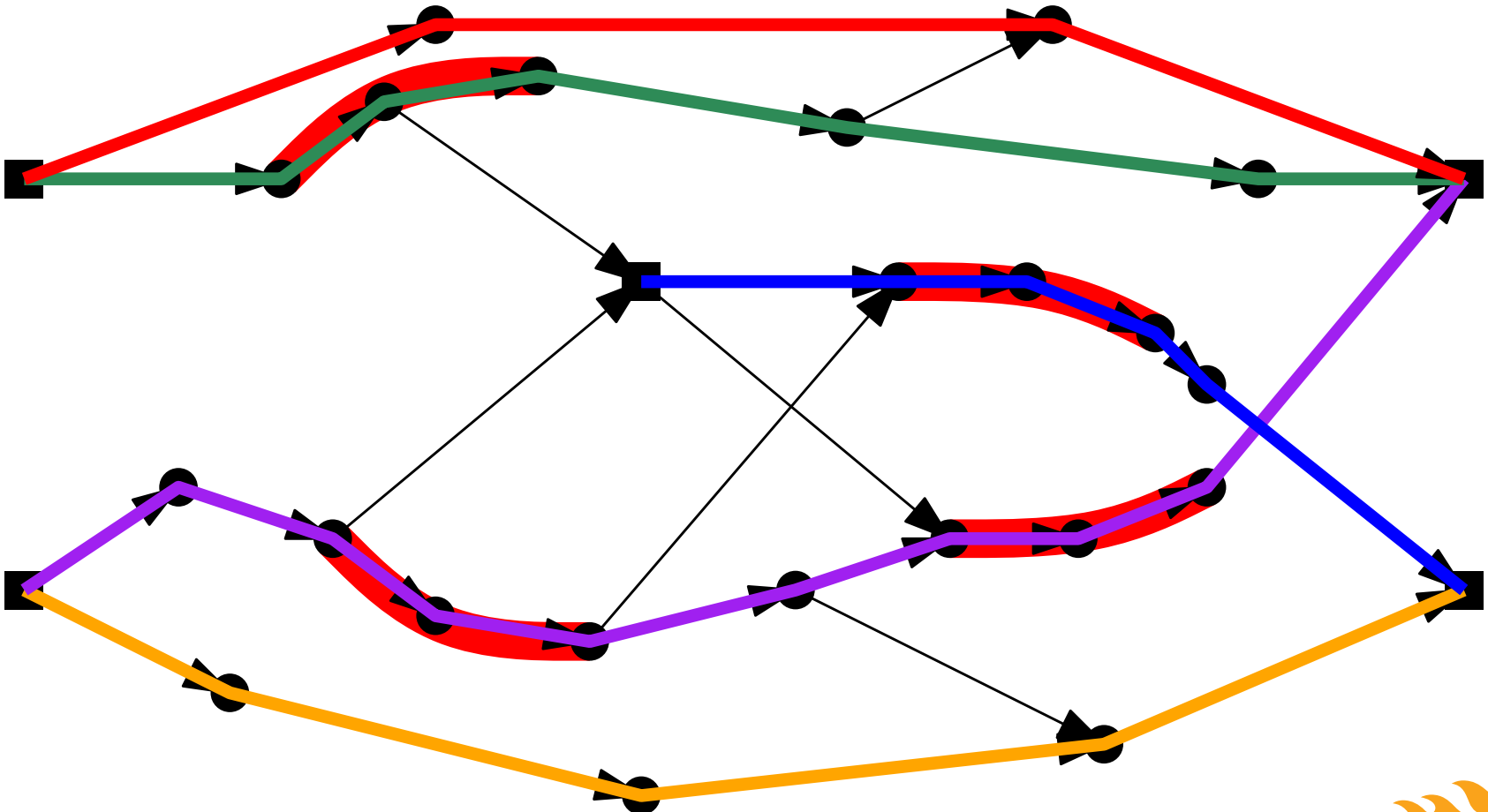




# Assembly with long reads (2)



# Assembly with long reads



# MIN-COST MPC WITH SUBPATH CONSTRAINTS

**INPUT:** An arc-weighted DAG $G$, and

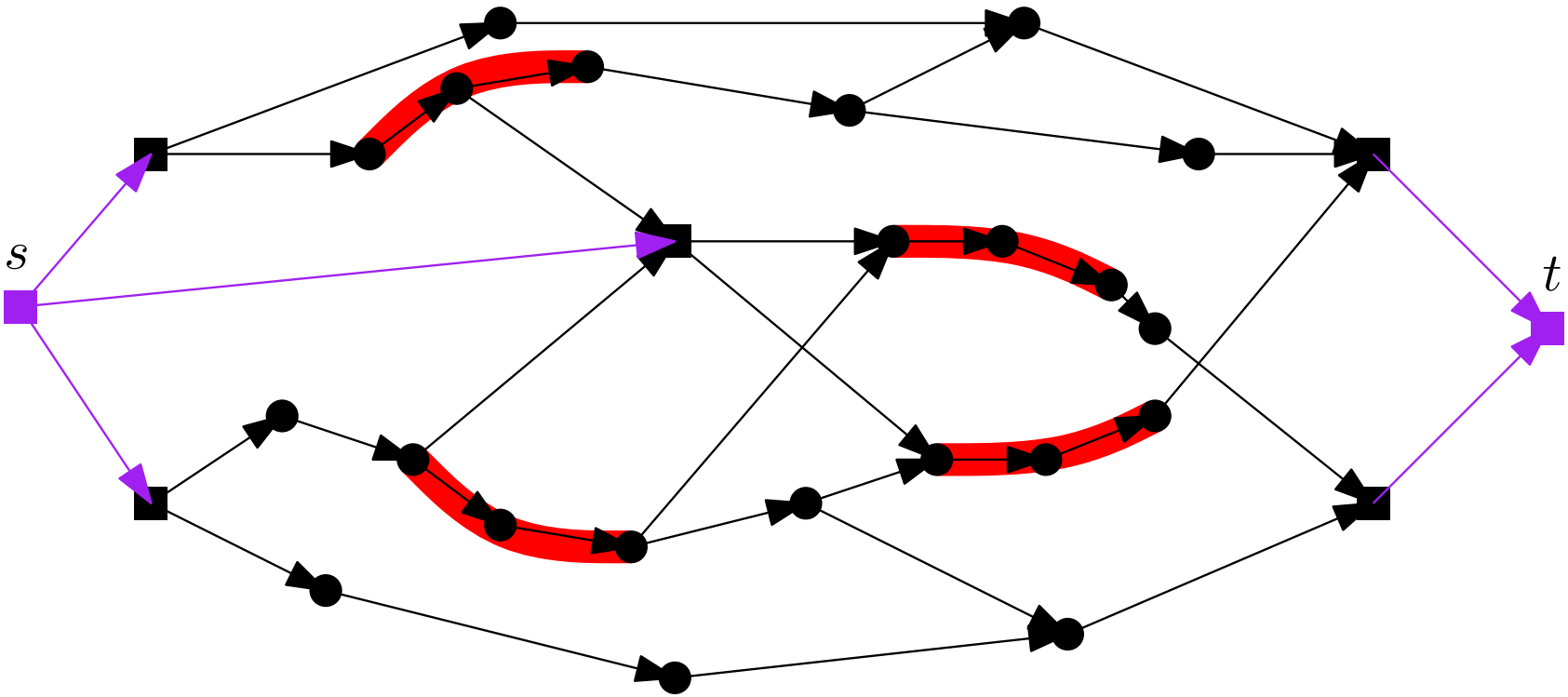
1. A superset $S$ of the sources, and a superset $T$ of the sinks
2. A family $\mathcal{P}^{in} = \{P_1^{in}, \ldots, P_c^{in}\}$ of directed paths in $G$

**TASK:** Find a minimum number $k$ of directed paths $P_1^{sol}, \ldots, P_k^{sol}$ in $G$ such that

1. Every node in $V(G)$ occurs in some $P_i^{sol}$
2. Every path $P^{in} \in \mathcal{P}^{in}$ is a subpath of some $P_i^{sol}$
3. Every path $P_i^{sol}$ starts in $S$ and ends in $T$
4. $\sum_{i=1}^{k} \sum_{\text{edge } e \in P_i^{sol}} w(e)$ is minimum among all such $k$ paths

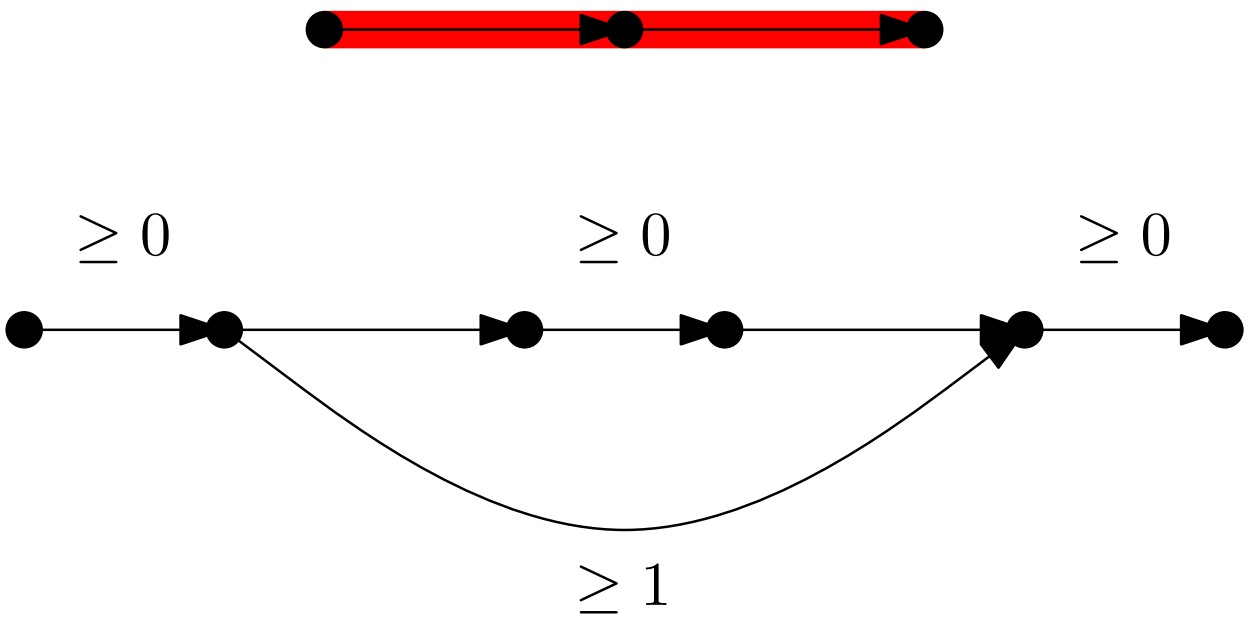
- introduced by [*Bao, Jiang, Girke, 2013, BRANCH*], but the case of overlapping constraints not solved

# Min-cost MPC with Subpath Constraints



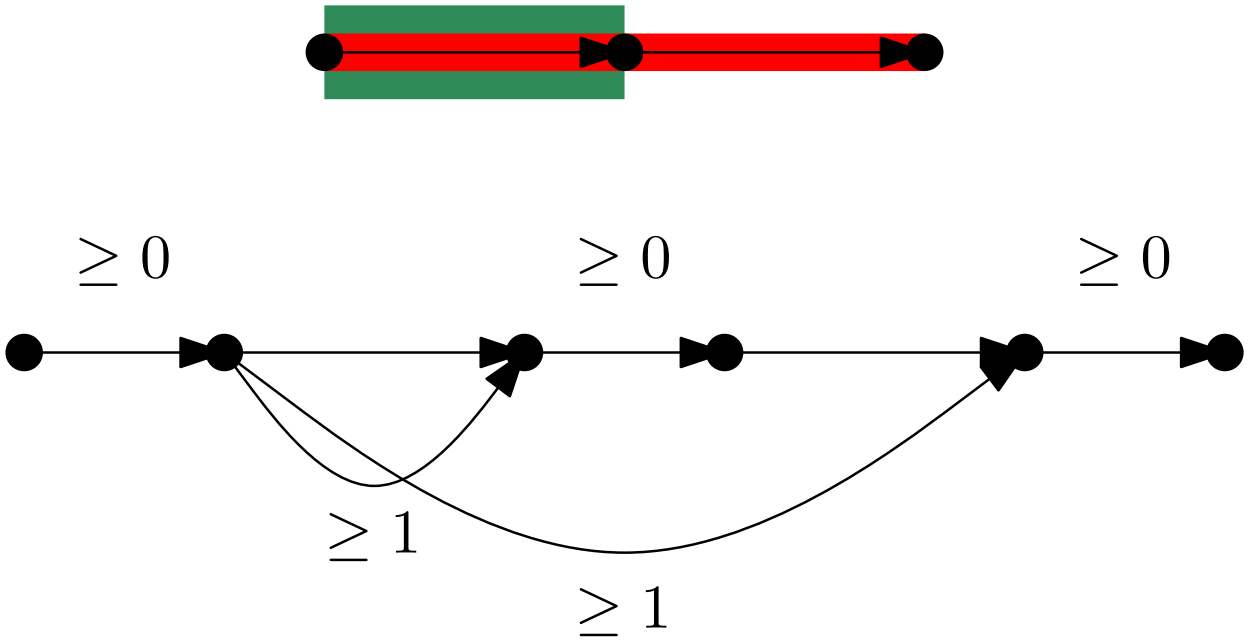

# Min-cost MPC with Subpath Constraints

Subpath constraints as arc demands:



# MIN-COST MPC WITH SUBPATH CONSTRAINTS

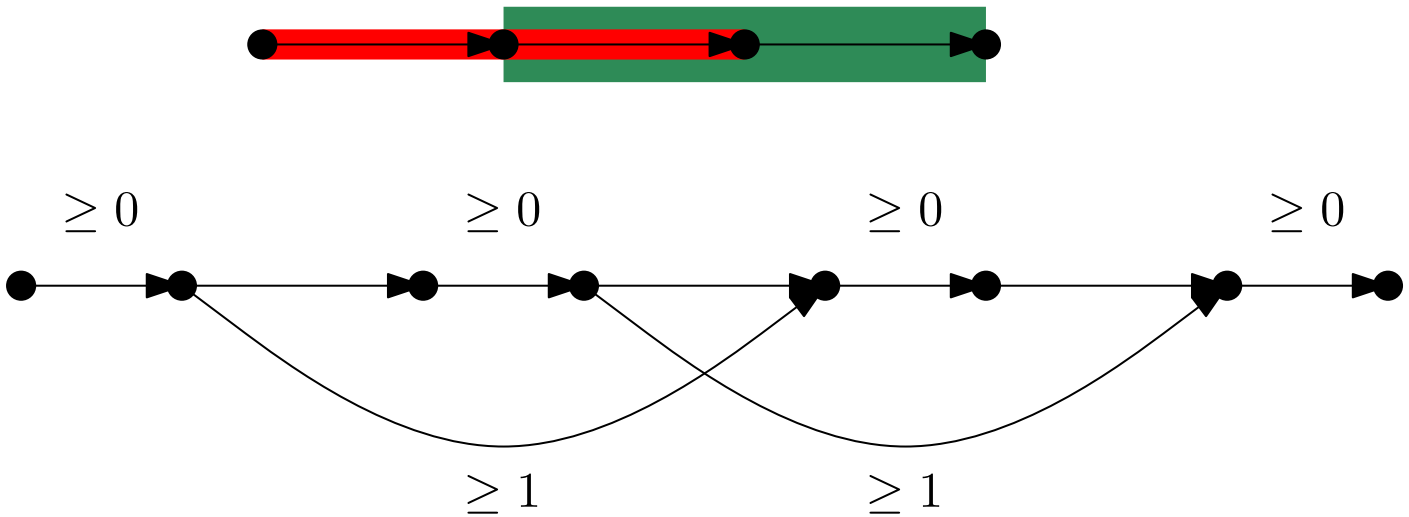
**Problem 1:** a constraint $P$ included in another constraint $Q$




- Remove $P$
- Can be implemented in time $O(N)$ with a suffix tree for large alphabets, [*Farach, 1997*]
  - $N$ = sum of lengths of Subpath Constraints

# MIN-COST MPC WITH SUBPATH CONSTRAINTS

**Problem 2:** Suffix-prefix overlaps




- Iteratively merge constraints with longest suffix-prefix overlap
- All suffix-prefix overlaps can be found in optimal time $O(N + |overlaps|)$ by [*Gusfield, Landau and Schieber, 1992*]
- Our iterative merging also takes $O(N + |overlaps|)$ time

# Min-cost MPC with Subpath Constraints

Pre-processing phase

- $O(N + |overlaps|)$

The flow network has size:

- $O(n)$ nodes and $O(m + c)$ arcs

Min-cost MPC with Subpath Constraints can be solved in time $O(N + |overlaps| + n^2 \log n + n(m + c))$ using [*Gabow and Tarjan, 1991*]

- [*Rizzi, T., Mäkinen, 2014*]

# MPC WITH PAIRED SUBPATH CONSTRAINTS

**INPUT:** A DAG $G$ and

1. A family $\mathcal{P}^{in} = \{(P^{in}_{1,1}, P^{in}_{1,2}), \ldots, (P^{in}_{t,1}, P^{in}_{t,2})\}$ of pairs of directed paths in $G$

**TASK:** Find a minimum number $k$ of directed paths $P^{sol}_1, \ldots, P^{sol}_k$ in $G$ such that

1. Every node in $V(G)$ occurs in some $P^{sol}_i$
2. For every pair $(P^{in}_{j,1}, P^{in}_{j,2}) \in \mathcal{P}^{in}$, there exists $P^{sol}_i$ such that both $P^{in}_{j,1}$ and $P^{in}_{j,2}$ are subpaths of $P^{sol}_i$

- introduced by [*Song and Florea, 2013, CLASS*]
- NP-hard
  - [*Rizzi, T., Mäkinen, 2014*]
  - [*Beerenwinkel, Beretta, Bonizzoni, Dondi and Pirola, 2014*]

# Conclusions

- Min-cost Minimum Path Cover

$$O(n^2 \log n + nm)$$

- Simultaneous assembly and expression estimation

polynomial-time, but NP-hard for given $k$

- Min-cost Minimum Path Cover with Subpath Constraints

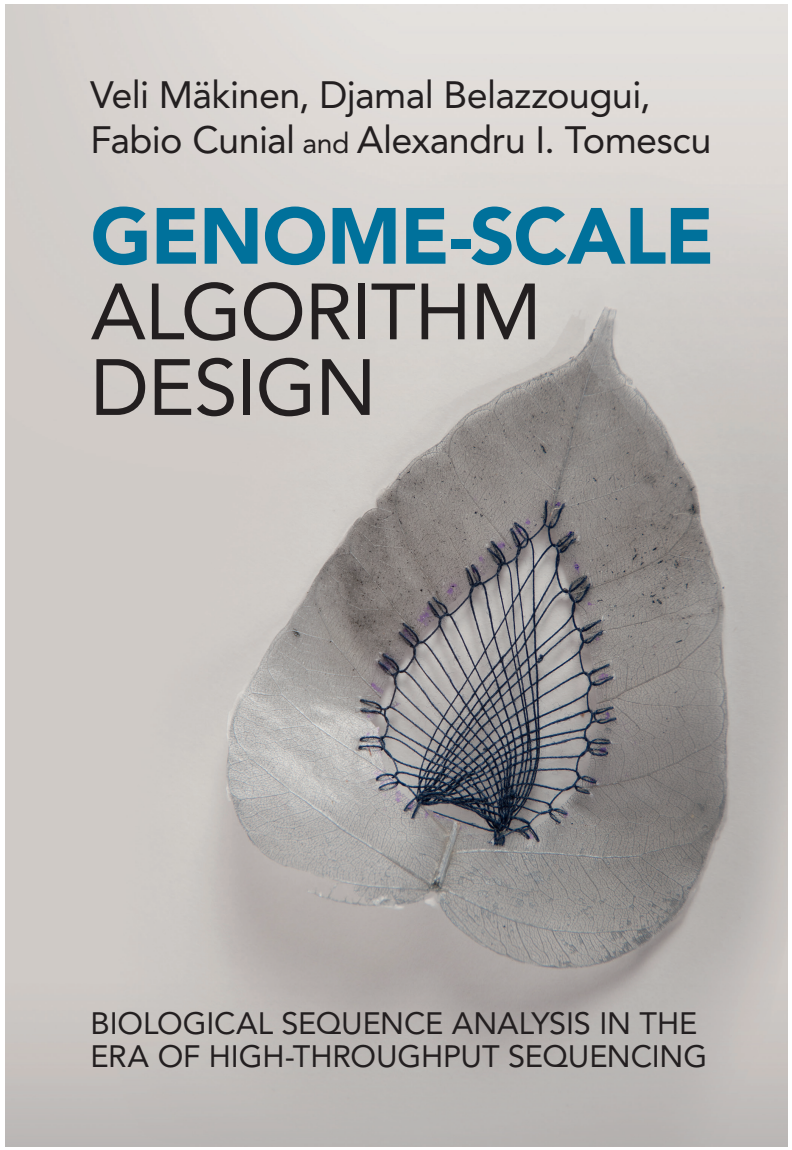
$$O(N + |overlaps| + n^2 \log n + n(m + c))$$

  - $c$ = number of Subpath Constraints
  - $N$ = sum of lengths of Subpath Constraints

- Minimum Path Cover with Pairs of Subpaths Constraints

NP-hard

# Advertisement



Thank you