Safety in multi-assembly via paths appearing in all path covers of a DAG

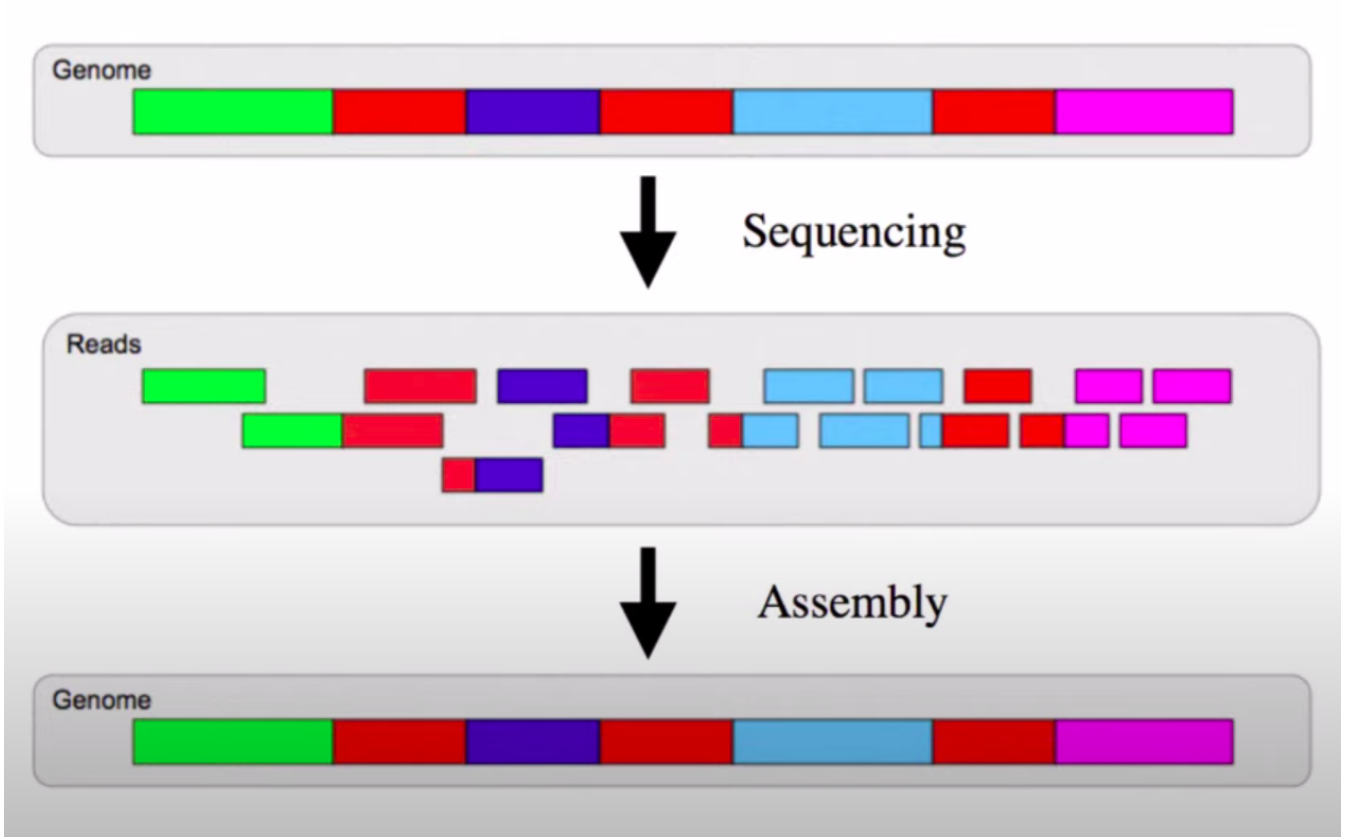
Manuel Cáceres, Brendan Mumey, Edin Husić, Romeo Rizzi, Massimo Cairo, Kristoffer Sahlin, Alexandru I. Tomescu

11.02.2021, DSB

Many problems in bioinformatics can be seen as puzzles

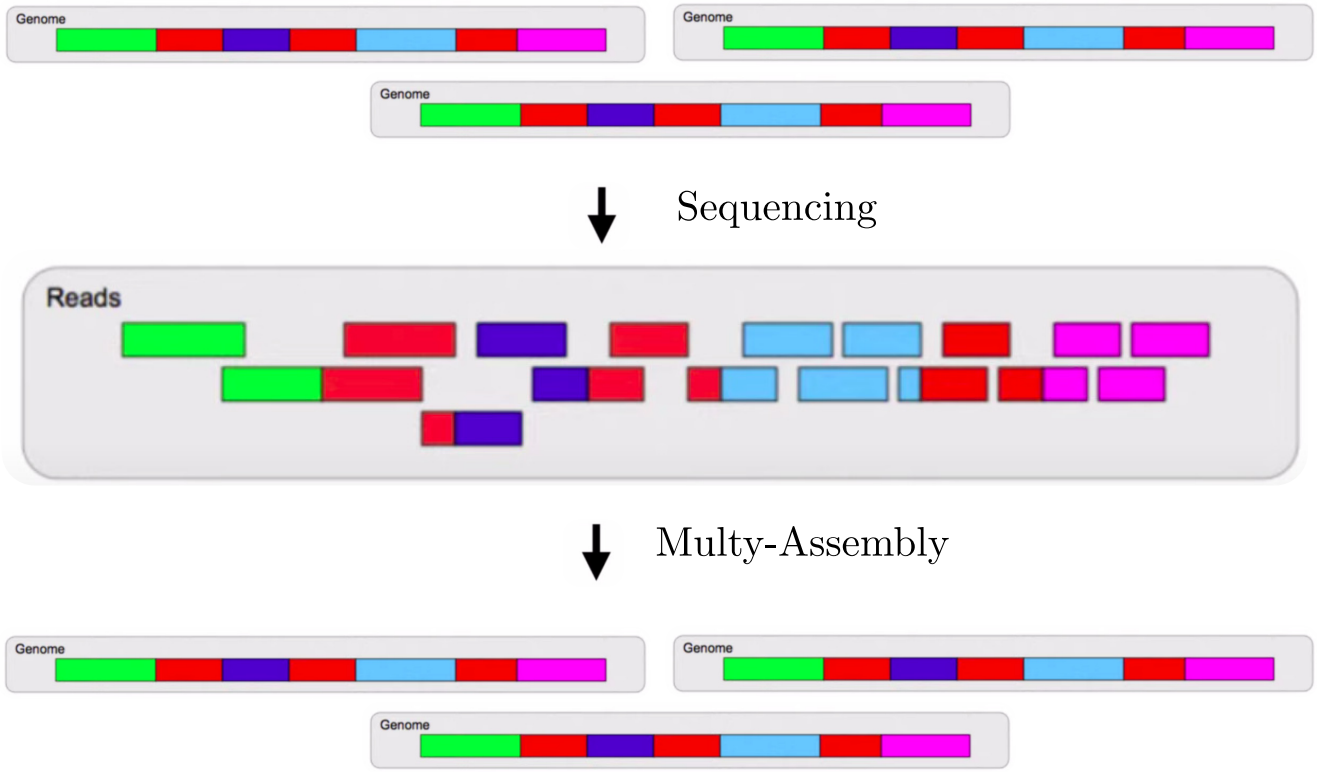
Genome Assembly

Reconstruct a genomic sequence based on *reads* obtained from it.

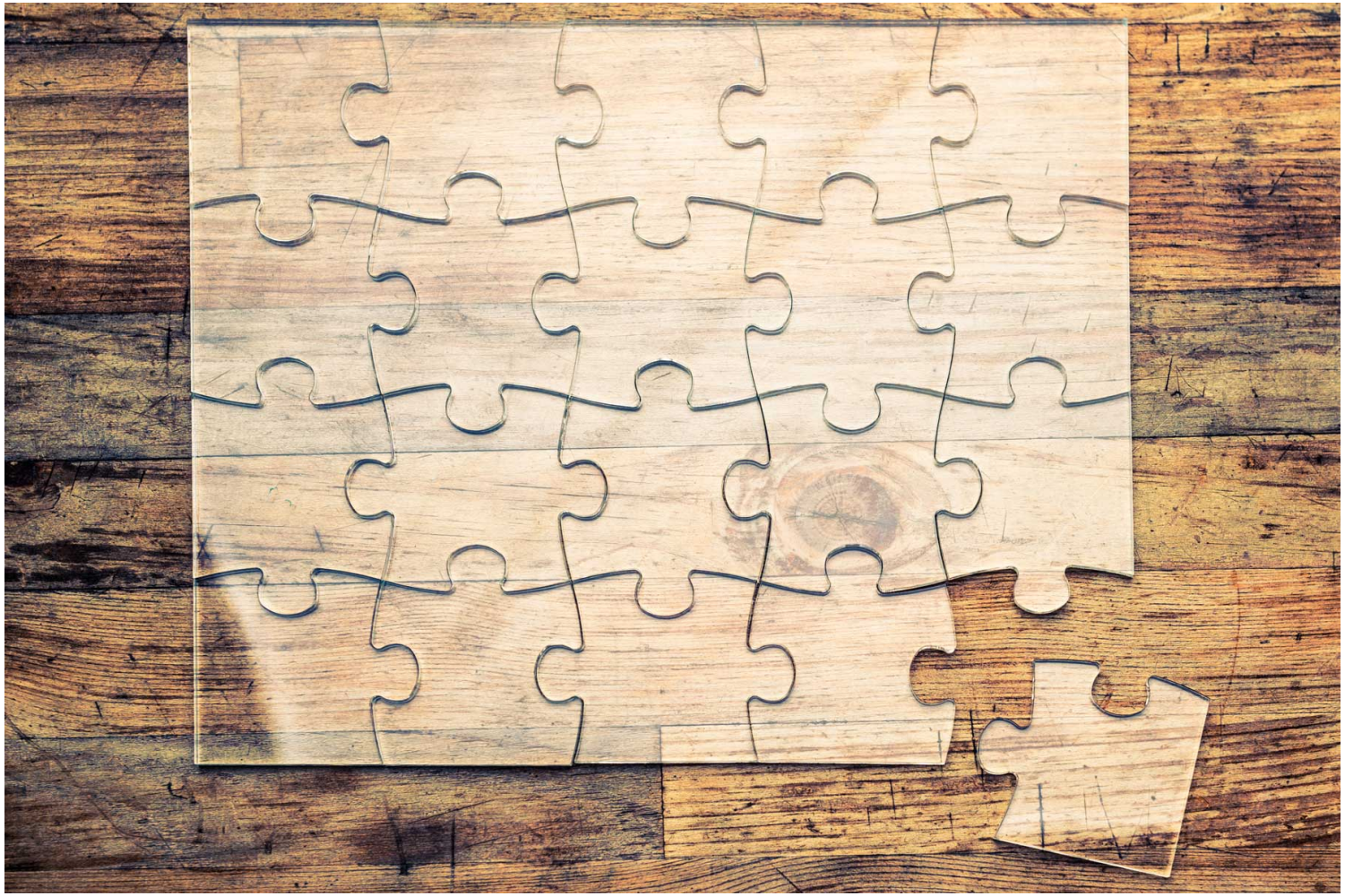


Multi-Assembly

Reconstruct multiple sequences based on *mixed-reads* obtained from all of them.

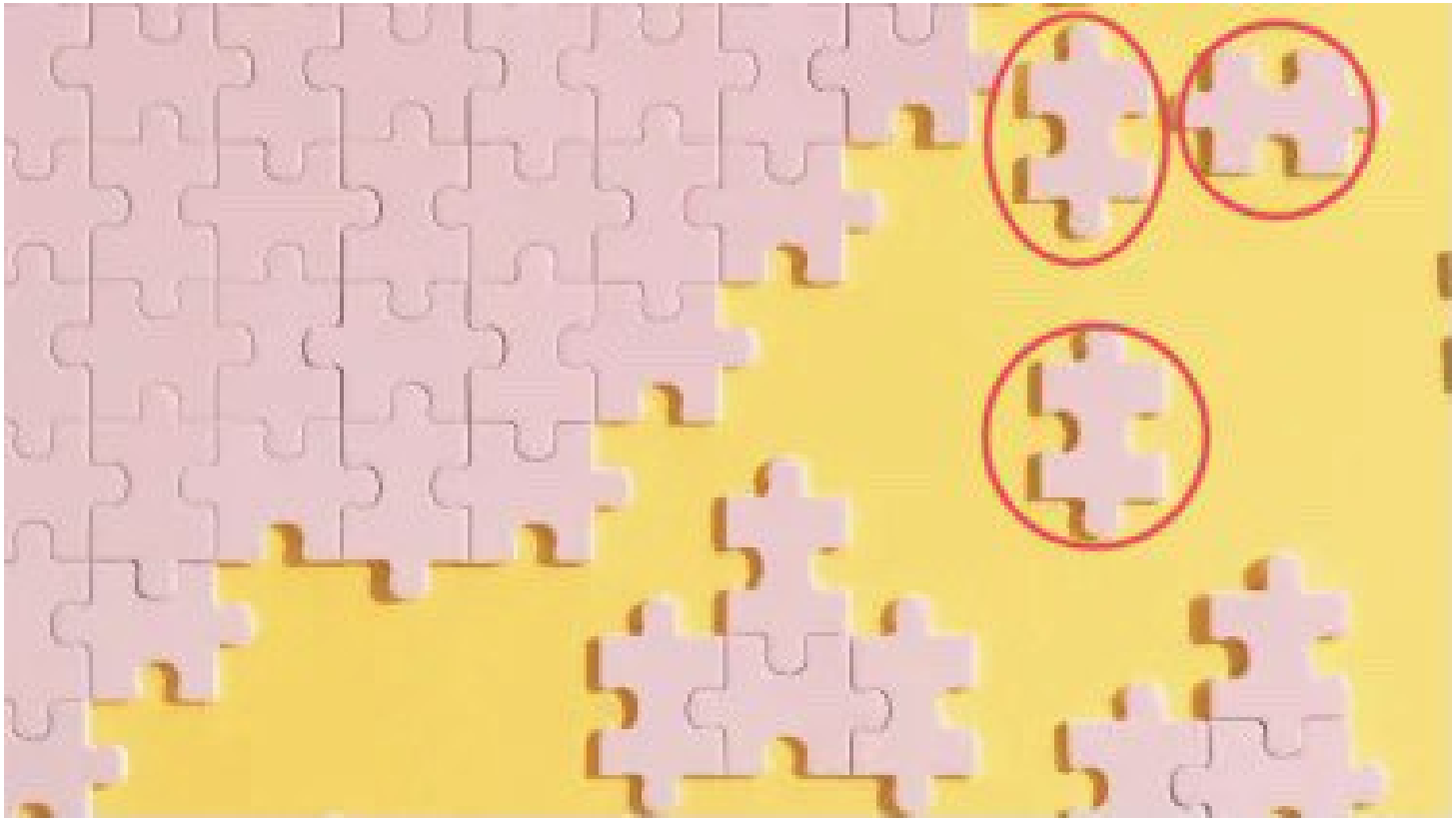


Puzzles solved perfectly

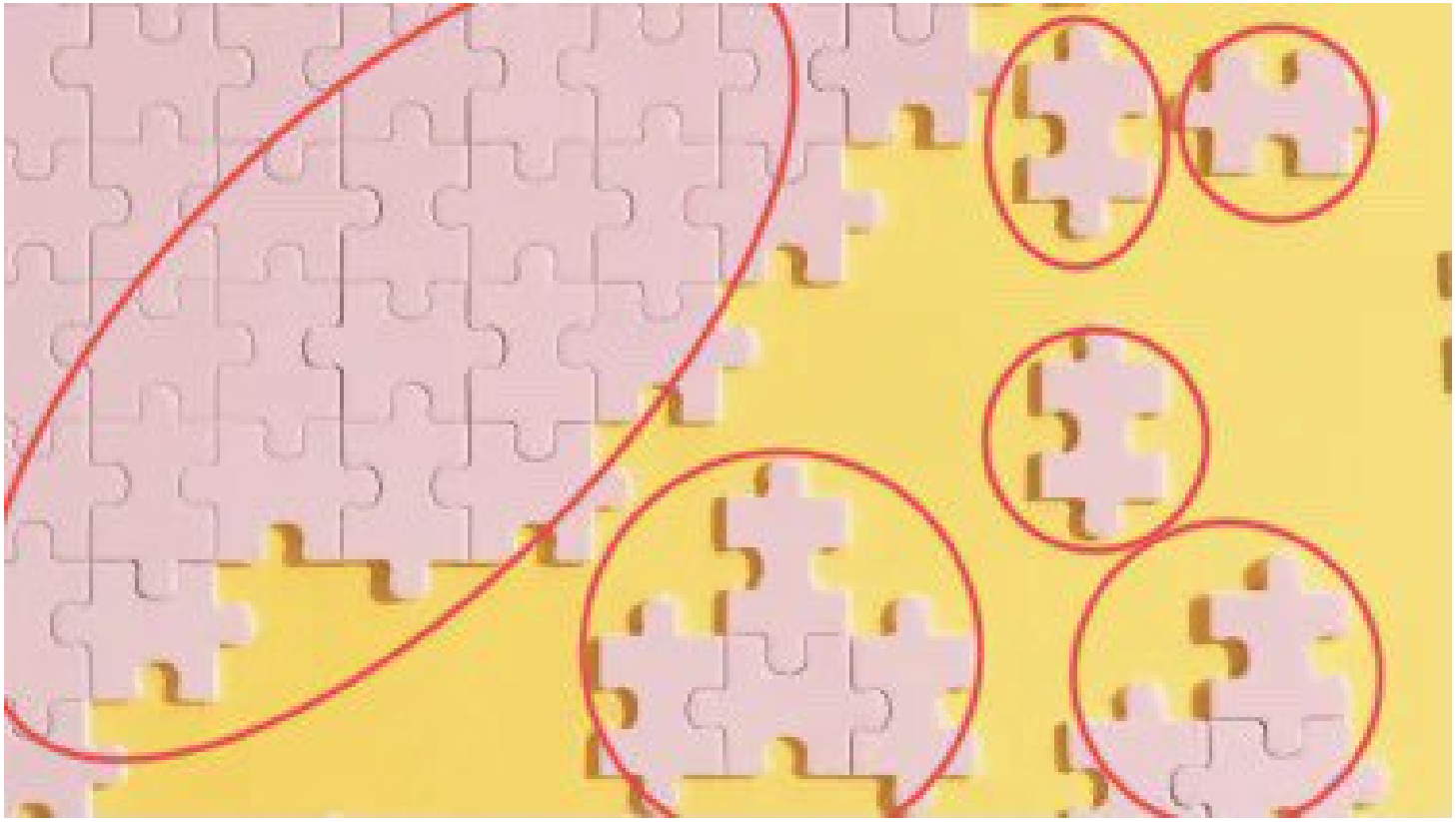


Unsolvable puzzles (multiple solutions)

Puzzles with multiple solutions



Safe parts common to *all* possible solutions



Contigs in Genome Assembly

Genomic fragments *promised* to occur in the original genome.

- Used and reported by practical assemblers [10, 11, 17, 20].
- Developed theoretically [24, 6, 7, 5].
  - *Completeness*

Contigs in Genome Assembly

Genomic fragments *promised* to occur in the original genome.

- Used and reported by practical assemblers [10, 11, 17, 20].
- Developed theoretically [24, 6, 7, 5].
  - *Completeness*

Not used in multi-assembly!

Path cover in a DAG

$$G = (V, E)$$

Model for Multi-assembly

Path cover in a DAG $G = (V, E)$

- Set of paths $\mathcal{P}$ such that every vertex of the graph appears in some path.

Model for Multi-assembly

Path cover in a DAG $G = (V, E)$

- Set of paths $\mathcal{P}$ such that every vertex of the graph appears in some path.
- Directed acyclic graph.

Model for Multi-assembly

Path cover in a DAG $G = (V, E)$

- Set of paths $\mathcal{P}$ such that every vertex of the graph appears in some path.
- Directed acyclic graph.
- E.g.:
  - RNA transcript assembly [25, 21, 3, 13, 9, 16, 19, 4, 12, 23, 22, 14].
  - Viral quasi-species assembly [8, 27, 2, 1]

Two types of path covers

Two types of path covers

- Minimum path cover (MPC).
  - Of size k, known as the width.
  - One can be computed in time $O(|V||E|)$ [18].

Two types of path covers

- Minimum path cover (MPC).
  - Of size k, known as the width.
  - One can be computed in time $O(|V||E|)$ [18].
- Generalized path cover.
  - Of size $\leq \ell$ (parameter).
  - Paths starting at $S \subseteq V$ (parameter).
  - Paths ending at $T \subseteq V$ (parameter).

Our results

Theoretical results

Theoretical results

	Minimum path covers	Generalized path covers
Safe edges	$O(k\|V\|\|E\|)$	$O(k\|V\|\|E\|)$
Maximal safe paths	$O(k^2\|V\|\|E\|)$	$O(max(1, 2k - \ell)k\|V\|\|E\|)$

Recall k is the width of the graph (size of an MPC)

Practical results

Apply to RNA transcript assembly defining *RNA contigs*.

Practical results

Apply to RNA transcript assembly defining *RNA contigs*.

Proof-of-concept study

- Splicing graphs from human annotated transcripts.
- Double length compared to *unitigs*.
- Transcript coverage of 80%.
- Less than 15 seconds (all transcript annotation from Ensembl database [26]).

Theoretical results

(MPC only)

General Approach (Avoid-and-test)

1. Compute a solution to the problem $\rightarrow$ An MPC $\mathcal{P}$.

General Approach (Avoid-and-test)

1. Compute a solution to the problem $\rightarrow$ An MPC $\mathcal{P}$.
   - In $O(|V||E|)$ [18].

General Approach (Avoid-and-test)

1. Compute a solution to the problem $\rightarrow$ An MPC $\mathcal{P}$.
   - In $O(|V||E|)$ [18].
2. For each subpart $\rightarrow$ Every subpath P of a path of $\mathcal{P}$.

General Approach (Avoid-and-test)

1. Compute a solution to the problem $\rightarrow$ An MPC $\mathcal{P}$.
   - In $O(|V||E|)$ [18].
2. For each subpart $\rightarrow$ Every subpath P of a path of $\mathcal{P}$.
   - Test if P is safe $\rightarrow$ Compute the width of G^P.

Safe edges

Theorem (Safe edges MPC)

An edge e is safe if and only if $\mathrm{width}(G) < \mathrm{width}(G \setminus e)$.

Theorem (Safe edges MPC)

An edge e is safe if and only if $\mathrm{width}(G) < \mathrm{width}(G \setminus e)$.

- Naive algorithm: $O(k|V|^2|E|)$.

Theorem (Safe edges MPC)

An edge e is safe if and only if $\mathrm{width}(G) < \mathrm{width}(G \setminus e)$.

- Naive algorithm: $O(k|V|^2|E|)$.
- But, we can do better.

Shrinking Primitive

Lemma (Shrinking [15])

Given a path cover $\mathcal{P} = P_1, \ldots, P_t$ of G, we can obtain a MPC of G in time $O(|E|(t - k + 1))$

Theorem (Safe edges MPC)

An edge e is safe if and only if $\mathrm{width}(G) < \mathrm{width}(G \setminus e)$.

- Naive algorithm: $O(k|V|^2|E|)$.
- But, we can do better.
  - $O(\sum_{e \in \mathcal{P}} \mu_e |E|) = O(k|V||E|)$

Maximal safe paths

Path Reduction

Definition

Given a path $P = x_1, \ldots, x_p$ of G, we define $G^P = (V, E^P)$, where

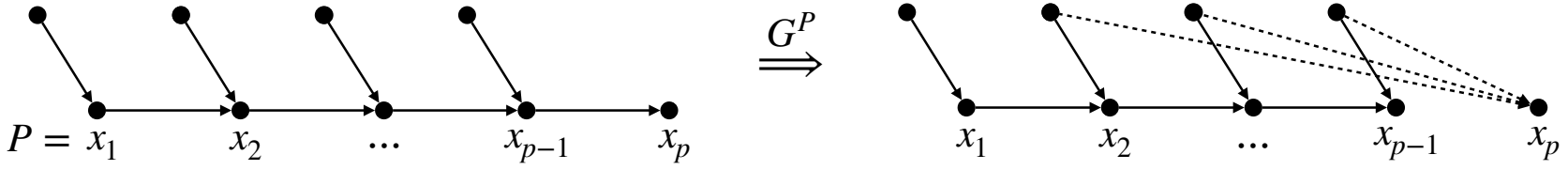
$$E^P = (E \setminus \{(x_{p-1}, x_p)\}) \cup \bigcup_{i=2}^{p} \left\{(u, x_p) \mid u \in N^-(x_i) \setminus \{x_{i-1}\}\right\}.$$

Path Reduction

Definition

Given a path $P = x_1, \ldots, x_p$ of G, we define $G^P = (V, E^P)$, where

$$E^P = (E \setminus \{(x_{p-1}, x_p)\}) \cup \bigcup_{i=2}^{p} \left\{(u, x_p) \mid u \in N^-(x_i) \setminus \{x_{i-1}\}\right\}.$$



Theorem (Maximal safe paths MPC)

Let $P = x_1, \ldots, x_p$ *be a path of* G*, such that* $x_1, \ldots, x_{p-1}$ *is a safe path. It holds that* P *is safe if and only if* $\mathrm{width}(G) < \mathrm{width}(G^P)$*.*

Theorem (Maximal safe paths MPC)

Let $P = x_1, \ldots, x_p$ be a path of G, such that $x_1, \ldots, x_{p-1}$ is a safe path. It holds that P is safe if and only if $\mathrm{width}(G) < \mathrm{width}(G^P)$.

- Paths of increasing size algorithm: $O(k|V|^3|E|)$.

Theorem (Maximal safe paths MPC)

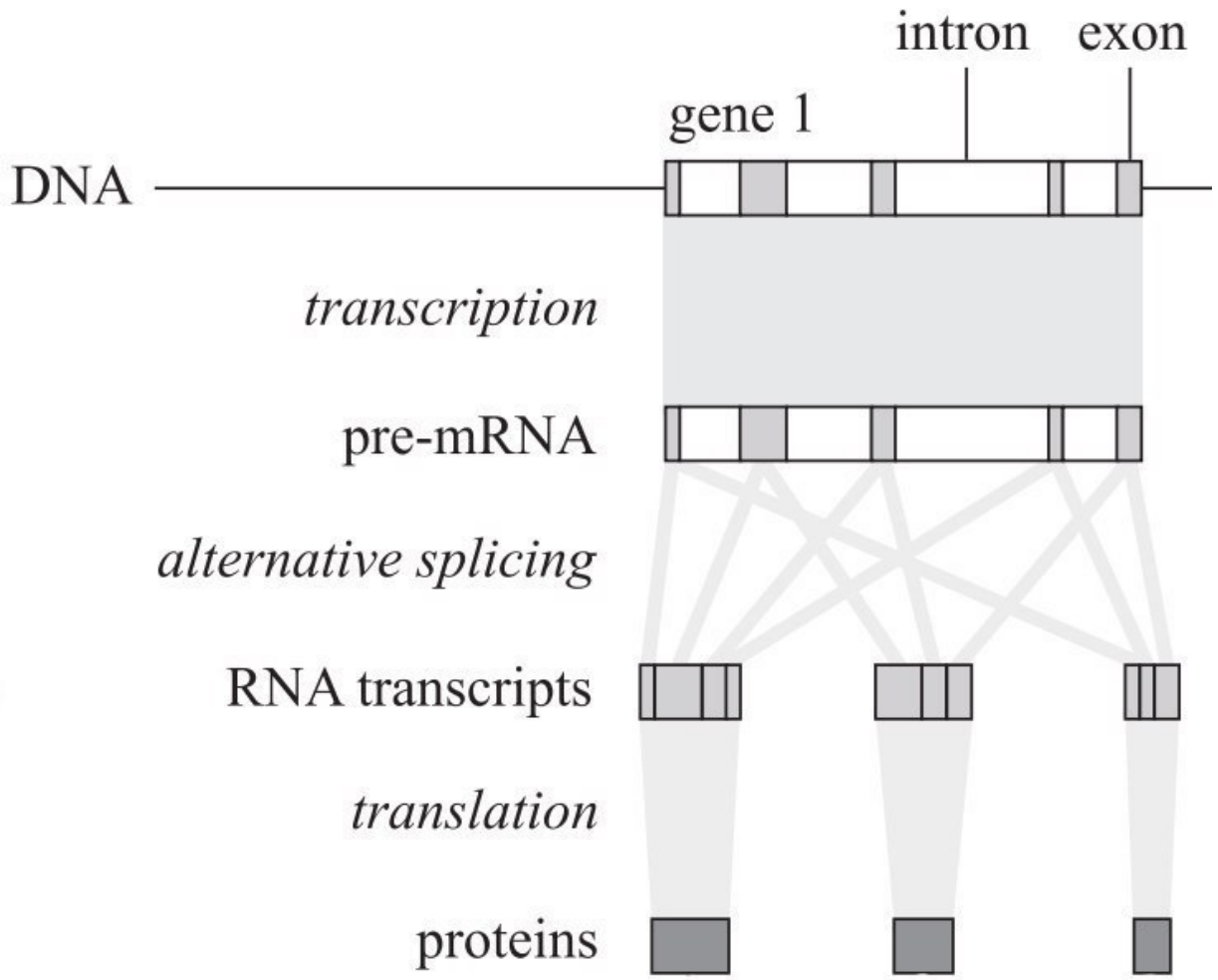
Let $P = x_1, \ldots, x_p$ be a path of G, such that $x_1, \ldots, x_{p-1}$ is a safe path. It holds that P is safe if and only if $\mathrm{width}(G) < \mathrm{width}(G^P)$.

- Paths of increasing size algorithm: $O(k|V|^3|E|)$.
- By using shrinking, and two-finger on each path: $O(k^2|V||E|)$.

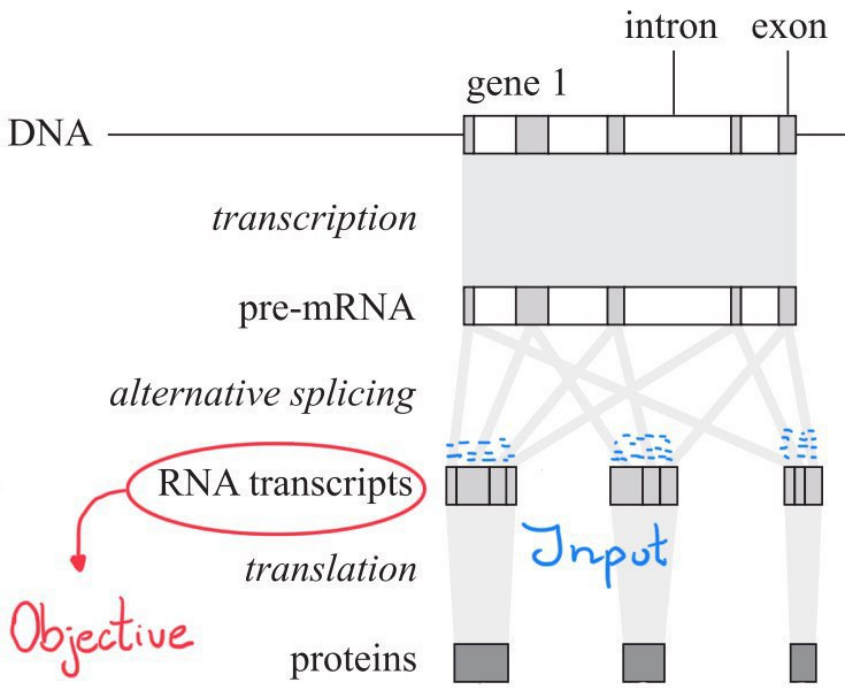
Practical results

(RNA contigs)

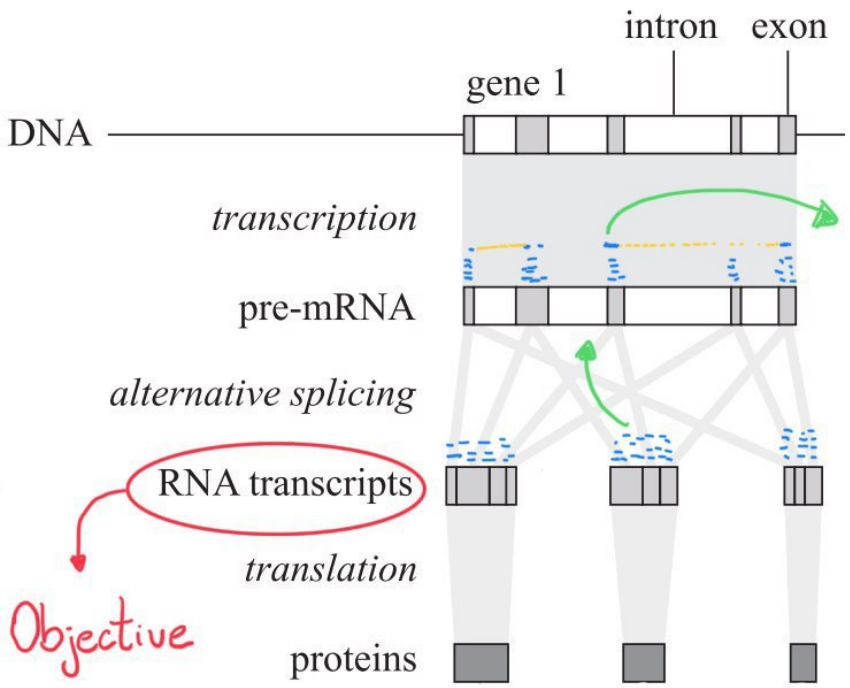
Central dogma of molecular biology



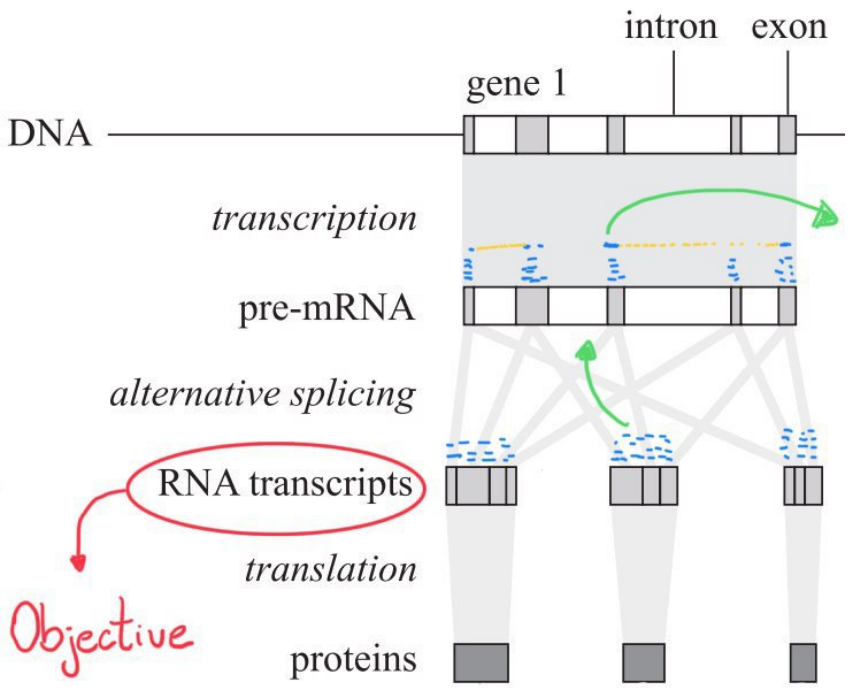
RNA transcript assembly

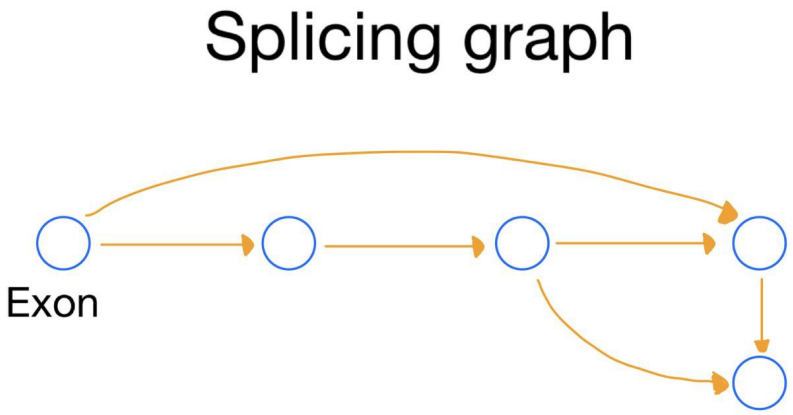


RNA transcript assembly

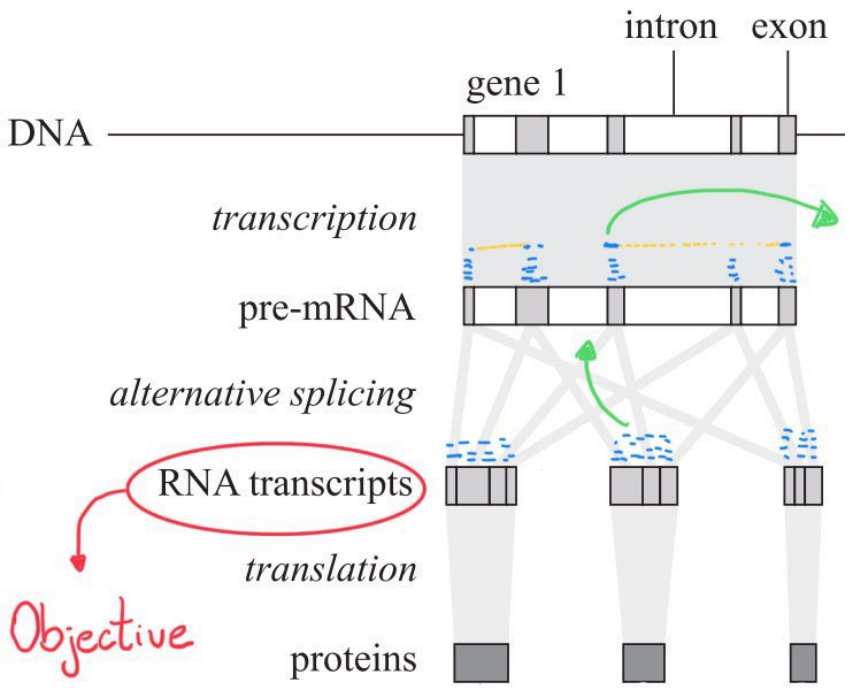


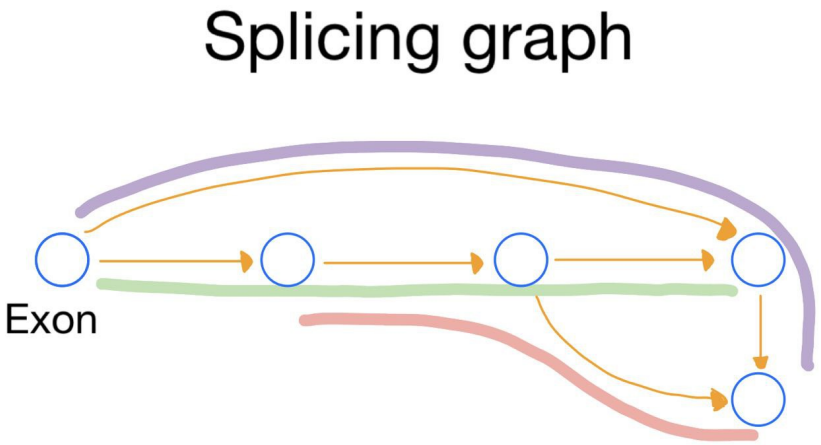
RNA transcript assembly



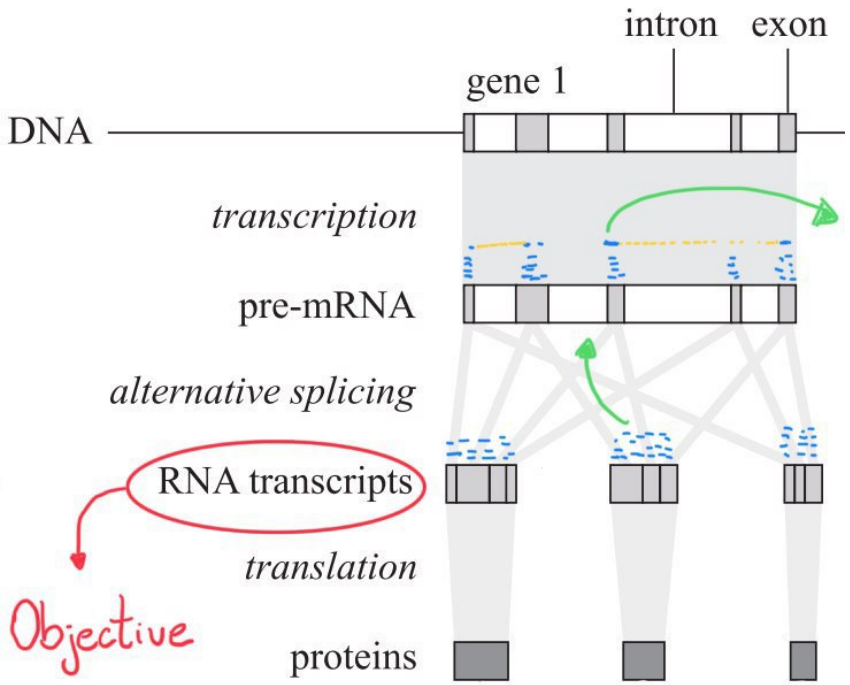


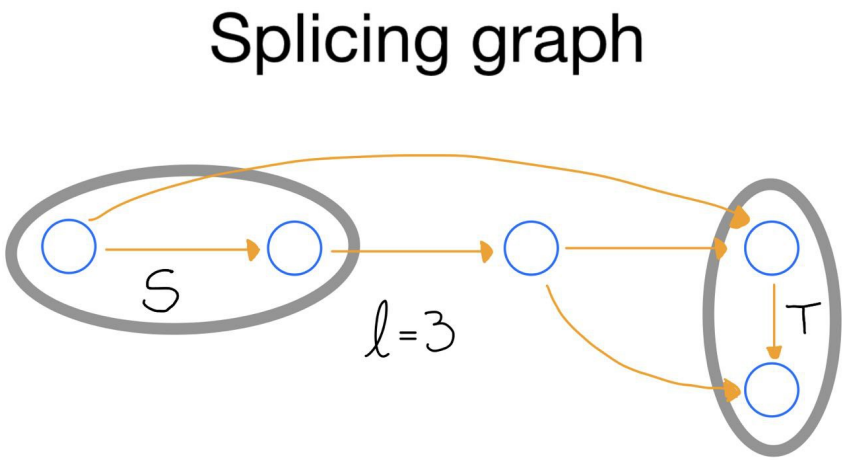
RNA transcript assembly





RNA contigs





RNA contigs = Maximal safe paths for generalized path cover

Experimental setup

- Human gene annotation (Ensembl [26],`GRCh38.p13`).
  - All transcripts in chromosomes 1 to 22, on the forward strand.
- Build splicing graph.
  - Perfect scenario.
  - Known answers.
- $\ell \in \{k, k+1, \ldots, 2k\}$.
- Baseline comparison with ST-unitigs.

Results

(less than 15 seconds to run)

Results

ℓ	small graphs (3-21 vertices)		medium graphs (22-52 vertices)		large graphs (53-725 vertices)	
	prec	mcr	prec	mcr	prec	mcr
k	0.84	0.82	0.81	0.64	0.84	0.56
	1.00	0.86	0.92	0.62	0.89	0.53
$k+1$	1.00	0.79	0.99	0.61	0.99	0.53
	1.00	0.83	1.00	0.59	1.00	0.50
t	1.00	0.79	1.00	0.61	1.00	0.53
	1.00	0.83	1.00	0.59	1.00	0.50
$2k$	1.00	0.79	1.00	0.61	1.00	0.53
	1.00	0.83	1.00	0.59	1.00	0.50
ST-unitigs	1.00	0.64	1.00	0.49	1.00	0.42
	1.00	0.67	1.00	0.47	1.00	0.39

Precision and Relative Maximum Coverage (of transcripts) for RNA contigs and ST-unitigs.

Results

ℓ	small graphs (3-21 vertices)	medium graphs (22-52 vertices)	large graphs (53-725 vertices)
k	2.65×	3.54×	3.50×
$k+1$	1.43×	1.83×	1.97×
t	1.42×	1.82×	1.94×
$2k$	1.42×	1.82×	1.94×

Relative length of longest RNA contig containing a ST-unitig.

Conclusions

Conclusions

- Efficient algorithms obtaining all maximal safe paths for generalized path covers.

Conclusions

- Efficient algorithms obtaining all maximal safe paths for generalized path covers.
- Proof-of-concept RNA contigs on human annotated transcripts.
  - Publicly available code, datasets, and manipulation scripts.
  - `https://github.com/elarielcl/SafePathsRNAPC`

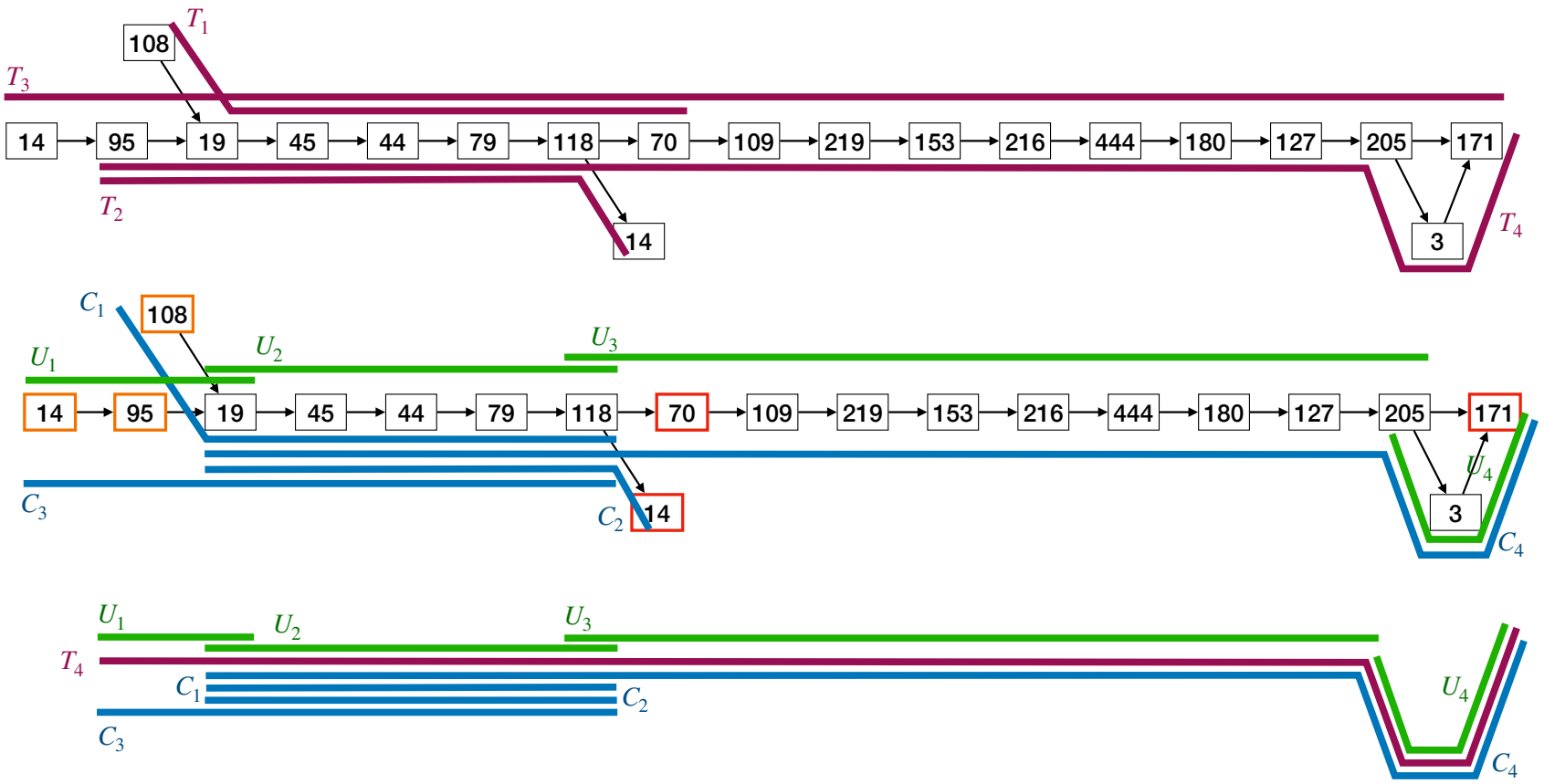
Conclusions

- Efficient algorithms obtaining all maximal safe paths for generalized path covers.
- Proof-of-concept RNA contigs on human annotated transcripts.
  - Publicly available code, datasets, and manipulation scripts.
  - `https://github.com/elarielcl/SafePathsRNAPC`
- Future developments of safe paths
  - Output of transcript assemblers.
  - Use as preprocessing step.
  - Validate transcript assemblies.
  - Apply to other multi-assembly problem.

Safety in multi-assembly via paths appearing in all path covers of a DAG

Manuel Cáceres, Brendan Mumey, Edin Husić, Romeo Rizzi, Massimo Cairo, Kristoffer Sahlin, Alexandru I. Tomescu

11.02.2021, DSB


T_1
T_2
T_3
T_4
108
14
95
19
45
44
79
118
70
109
219
153
216
444
180
127
205
171
14
3
C_1
C_2
C_3
C_4
U_1
U_2
U_3
U_4

References I

Baaijens, J. A., der Roest, B. V., Köster, J., Stougie, L., and Schönhuth, A.
Full-length de novo viral quasispecies assembly through variation graph construction.
Bioinform. 35, 24 (2019), 5086–5094.

Baaijens, J. A., Stougie, L., and Schönhuth, A.
Strain-aware assembly of genomes from mixed samples using flow variation graphs.
In *Research in Computational Molecular Biology - 24th Annual International Conference, RECOMB 2020, Padua, Italy, May 10-13, 2020, Proceedings* (2020), R. Schwartz, Ed., vol. 12074 of *Lecture Notes in Computer Science*, Springer, pp. 221–222.

References II

Bao, E., Jiang, T., and Girke, T.
Branch: boosting rna-seq assemblies with partial or related genomic sequences.
Bioinformatics 29, 10 (2013), 1250–1259.

Bernard, E., Jacob, L., Mairal, J., and Vert, J.
Efficient RNA isoform identification and quantification from rna-seq data with network flows.
Bioinformatics 30, 17 (2014), 2447–2455.

Cairo, M., Khan, S., Rizzi, R., Schmidt, S. S., Tomescu, A. I., and Zirondelli, E. C.
Genome assembly, a universal theoretical framework: unifying and generalizing the safe and complete algorithms.
CoRR abs/2011.12635 (2020).

References III

Cairo, M., Medvedev, P., Acosta, N. O., Rizzi, R., and Tomescu, A. I.
An optimal $O(nm)$ algorithm for enumerating all walks common to all closed edge-covering walks of a graph.
ACM Trans. Algorithms 15, 4 (2019), 48:1–48:17.

Cairo, M., Rizzi, R., Tomescu, A. I., and Zirondelli, E. C.
Genome assembly, from practice to theory: safe, complete and linear-time.
CoRR abs/2002.10498 (2020).

Eriksson, N., Pachter, L., Mitsuya, Y., Rhee, S.-Y., Wang, C., Gharizadeh, B., Ronaghi, M., Shafer, R. W., and Beerenwinkel, N.
Viral population estimation using pyrosequencing.
PLoS Computational Biology 4, 5 (2008).

References IV

Feng, J., Li, W., and Jiang, T.
Inference of isoforms from short sequence reads.
In *Research in Computational Molecular Biology, 14th Annual International Conference, RECOMB 2010, Lisbon, Portugal, April 25-28, 2010. Proceedings* (2010), B. Berger, Ed., vol. 6044 of *Lecture Notes in Computer Science*, Springer, pp. 138–157.

Jackson, B. G.
Parallel methods for short read assembly.
PhD thesis, Iowa State University, 2009.

Kingsford, C., Schatz, M. C., and Pop, M.
Assembly complexity of prokaryotic genomes using short reads.

BMC bioinformatics 11, 1 (2010), 21.

References V

Li, J. J., Jiang, C., Brown, J., Huang, H., and Bickel, P.
Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation.
Proc. of the National Academy of Sciences 108, 50 (2011), 19867–19872.

Li, W., Feng, J., and Jiang, T.
Isolasso: A LASSO regression approach to rna-seq based transcriptome assembly.
Journal of Computational Biology 18, 11 (2011), 1693–1707.

References VI

Lin, Y.-Y., Dao, P., Hach, F., Bakhshi, M., Mo, F., Lapuk, A., Collins, C., and Sahinalp, S. C.
CLIIQ: Accurate Comparative Detection and Quantification of Expressed Isoforms in a Population.
In *Proc. WABI 2012* (2012), vol. 7534 of *LNCS*, Springer, pp. 178–189.

Mäkinen, V., Tomescu, A. I., Kuosmanen, A., Paavilainen, T., Gagie, T., and Chikhi, R.
Sparse Dynamic Programming on DAGs with Small Width.
ACM Transactions on Algorithms (TALG) 15, 2 (2019), 1–21.

Mangul, S., Caciula, A., Al Seesi, S., Brinza, D., Banday, A. R., and Kanadia, R.
An integer programming approach to novel transcript reconstruction from paired-end RNA-Seq reads.

References VII

In *ACM International Conference on Bioinformatics, Computational Biology and Biomedicine, BCB' 12, Orlando, FL, USA - October 08 - 10, 2012* (2012), S. Ranka, T. Kahveci, and M. Singh, Eds., ACM, pp. 369–376.

Medvedev, P., and Brudno, M.
Maximum likelihood genome assembly.
Journal of Computational Biology 16, 8 (2009), 1101–1116.

Orlin, J. B.
Max flows in $O(nm)$ time, or better.
In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing* (2013), pp. 765–774.

References VIII

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L.
Stringtie enables improved reconstruction of a transcriptome from rna-seq reads.
Nature Biotechnology 33, 3 (2015), 290–295.

Pevzner, P. A., Tang, H., and Waterman, M. S.
An Eulerian path approach to DNA fragment assembly.
Proceedings of the National Academy of Sciences 98, 17 (2001), 9748–9753.

Song, L., and Florea, L.
CLASS: constrained transcript assembly of RNA-seq reads.
BMC Bioinformatics 14, S-5 (2013), S14.
Proceedings paper from RECOMB-seq: Third Annual RECOMB Satellite Workshop on Massively Parallel Sequencing, Beijing, China. 11-12 April 2013.

References IX

Tomescu, A. I., Kuosmanen, A., Rizzi, R., and Mäkinen, V.
A novel combinatorial method for estimating transcript expression with rna-seq: Bounding the number of paths.
In *WABI 2013– 13th Workshop on Algorithms for Bioinformatics, Sophia Antipolis, France, September 2-4, 2013. Proceedings* (2013), A. Darling and J. Stoye, Eds., 8126, Springer Berlin Heidelberg, pp. 85–98.

Tomescu, A. I., Kuosmanen, A., Rizzi, R., and Mäkinen, V.
A novel min-cost flow method for estimating transcript expression with rna-seq.
BMC Bioinformatics 14, S-5 (2013), S15.
Proceedings paper from RECOMB-seq: Third Annual RECOMB Satellite Workshop on Massively Parallel Sequencing Beijing, China. 11-12 April 2013.

References X

Tomescu, A. I., and Medvedev, P.
Safe and complete contig assembly via omnitigs.
In *Research in Computational Molecular Biology - 20th Annual Conference, RECOMB 2016, Santa Monica, CA, USA, April 17-21, 2016, Proceedings* (2016), M. Singh, Ed., vol. 9649 of *Lecture Notes in Computer Science*, Springer, pp. 152–163.

Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M., Salzberg, S., Wold, B., and Pachter, L.
Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.
Nature Biotechnology 28 (2010), 511–515.

References XI

Yates, A. D., et al.
Ensembl 2020.
Nucleic Acids Research 48, D1 (11 2019), D682–D688.

Zagordi, O., Bhattacharya, A., Eriksson, N., and Beerenwinkel, N.
ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data.
BMC Bioinformatics 12, 1 (2011), 119+.