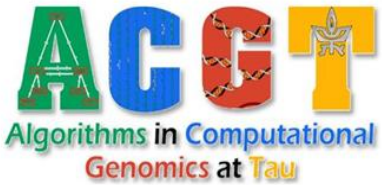


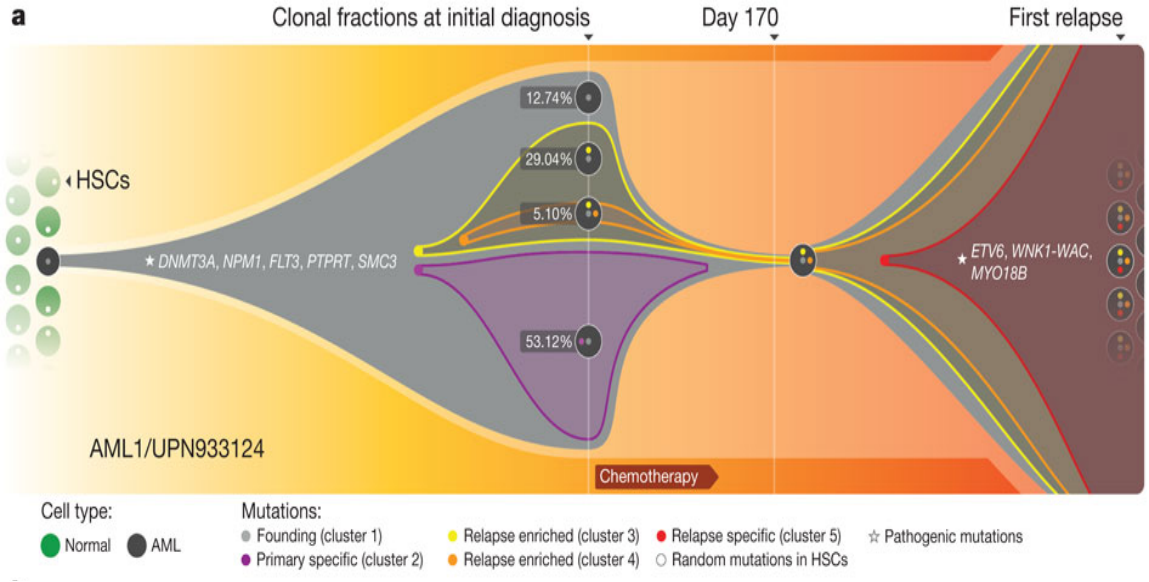
Copy-Number Evolution Problems: Complexity and Algorithms

Mohammed El-Kebir, Ben Raphael, Ron Shamir, Roded Sharan, **Simone Zaccaria**, Meirav Zehavi and **Ron Zeira**

August 23
WABI 2016

Evolution of Cancer

- Cancer is an evolutionary process characterized by the accumulation of somatic mutations
- Different populations of cells form a tumor

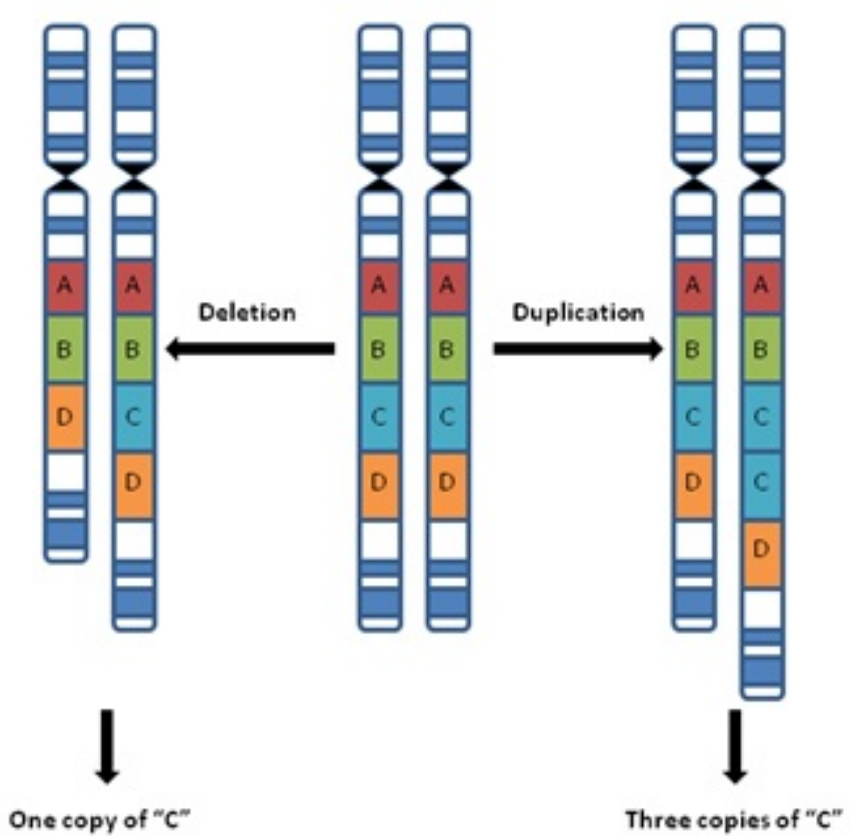


Ding *et al*. Nature 2012

- Inference of evolution for understanding:
  - Order of mutations
  - Dynamics of clones
  - Effects of treatment
  - Driver mutations
  - …

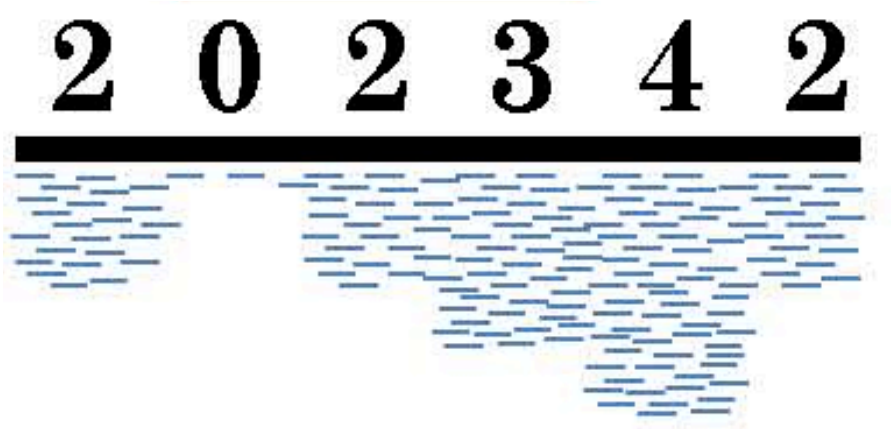
Copy-Number Aberrations

- For most tumor types, *copy-number aberrations* are ubiquitous



Copy-number profiles

- *Copy-number profiles* encode the number of copies of each region along a chromosome.
- Inferred from experimental data (sequencing, aCGH, FISH)



Operations on profiles

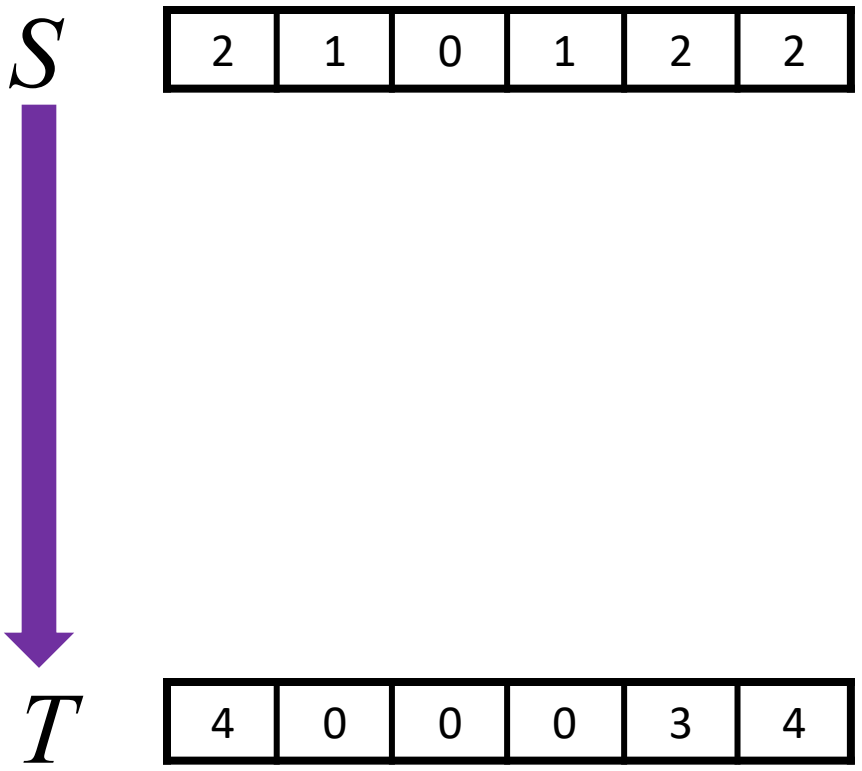
Chromosome =
Copy number profile

S

Operations on profiles

Chromosome = *Copy number profile*

Evolves by copy number change events

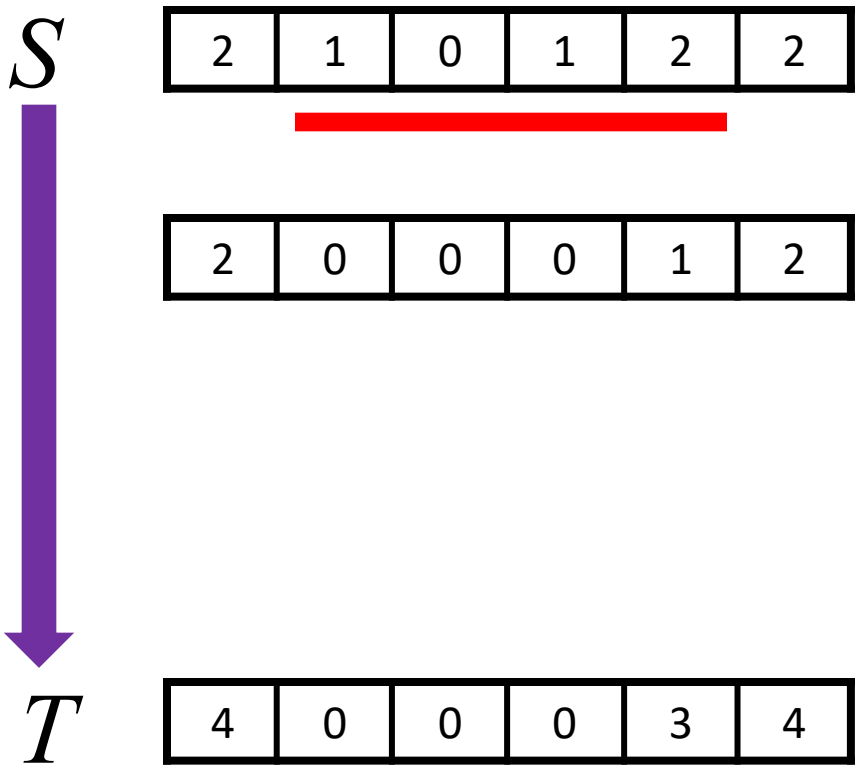


Operations on profiles

Chromosome = *Copy number profile*

Evolves by copy number change events

Event: segmental *duplication* / *deletion*

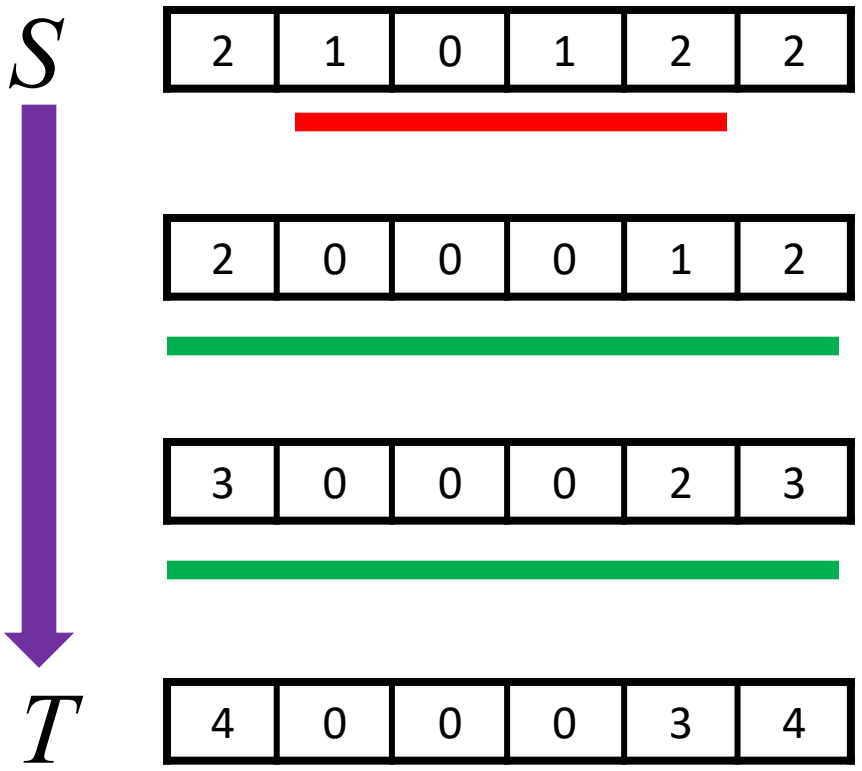


Operations on profiles

Chromosome = *Copy number profile*

Evolves by copy number change events

Event: segmental *duplication* / *deletion*



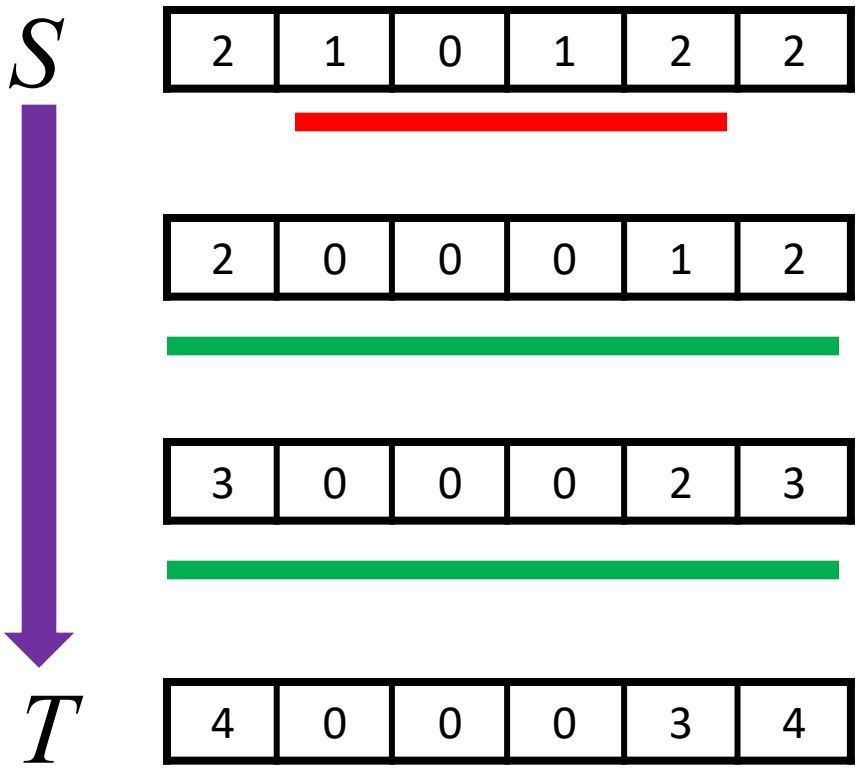
Operations on profiles

Chromosome = *Copy number profile*

Evolves by copy number change events

Event: segmental *duplication* / *deletion*

Distance from S to T: *least no. of operations*



$$dist(S,T) = 3$$

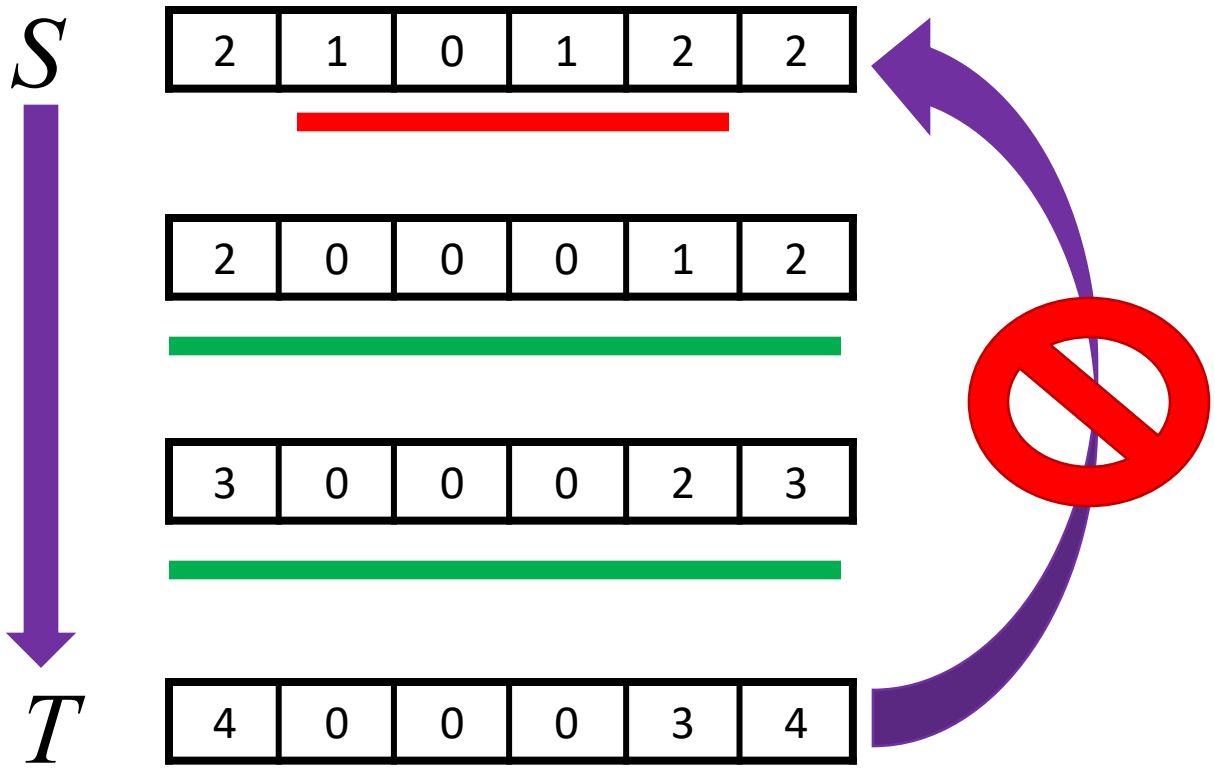
Operations on profiles

Chromosome = *Copy number profile*

Evolves by copy number change events

Event: segmental *duplication* / *deletion*

Distance from S to T: *least no. of operations*



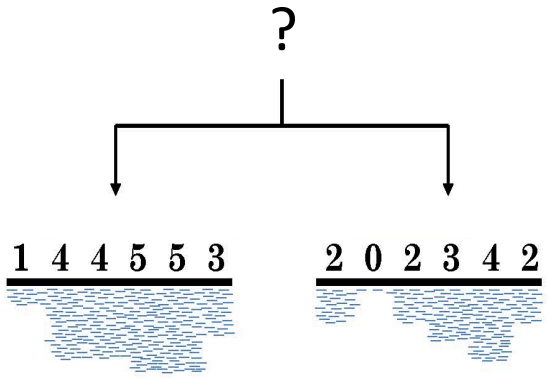
$$dist(S,T) = 3$$

Previous work

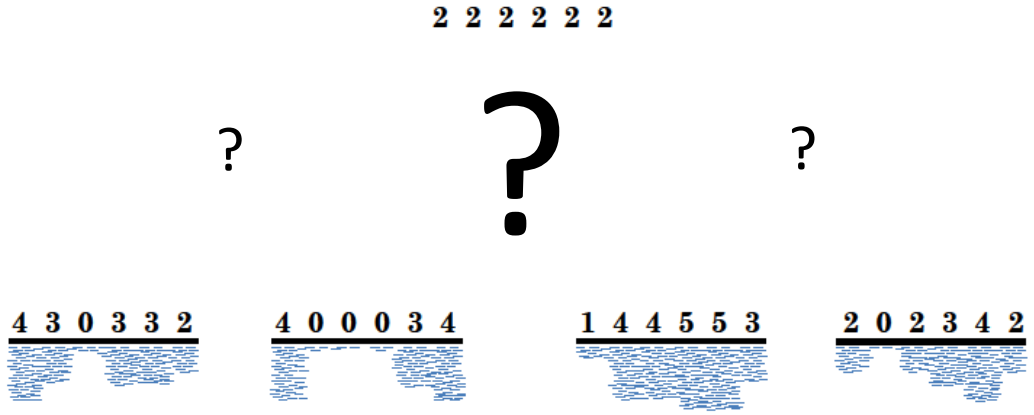
- Schwarz et al. PLoS CB 2014:
  - Presented model, developed a heuristic procedure for tree reconstruction
  - Reconstructed ovarian cancer sample phylogeny.
- Shamir, Zehavi, and Zeira CPM 2016:
  - A linear time algorithm for S → T distance.

This work

- Copy-number triplet (CN3):

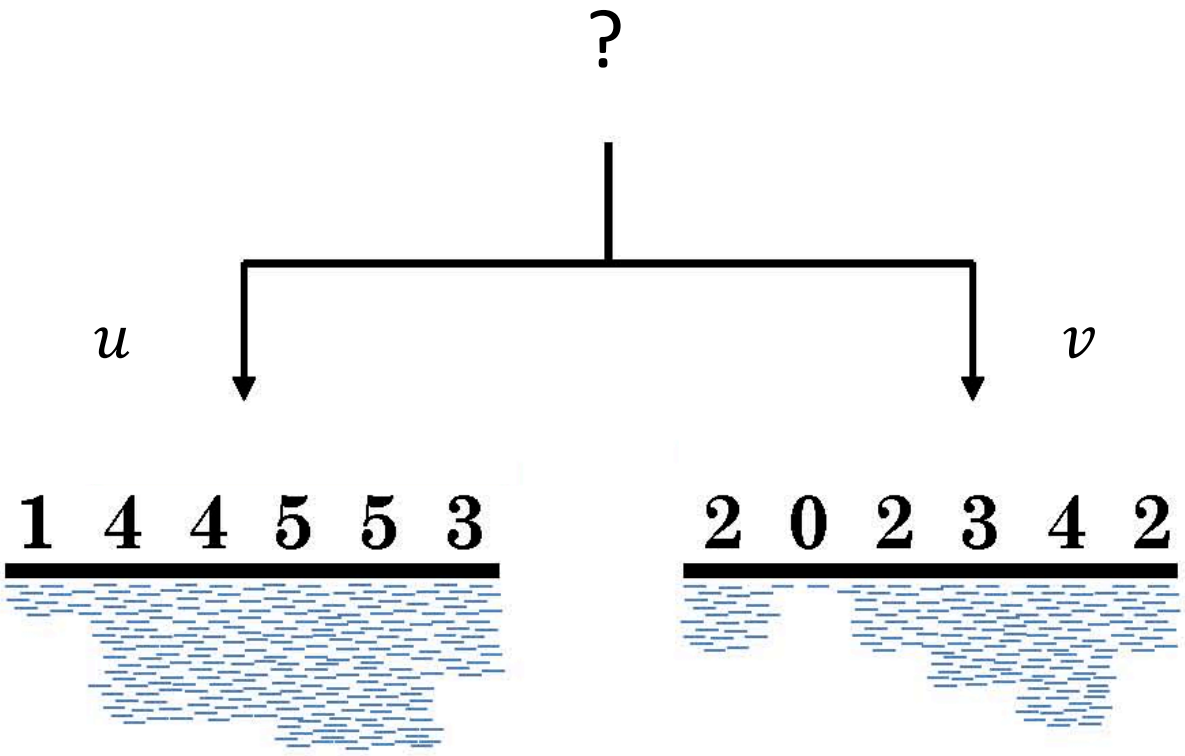


- Copy-number tree (CNT):



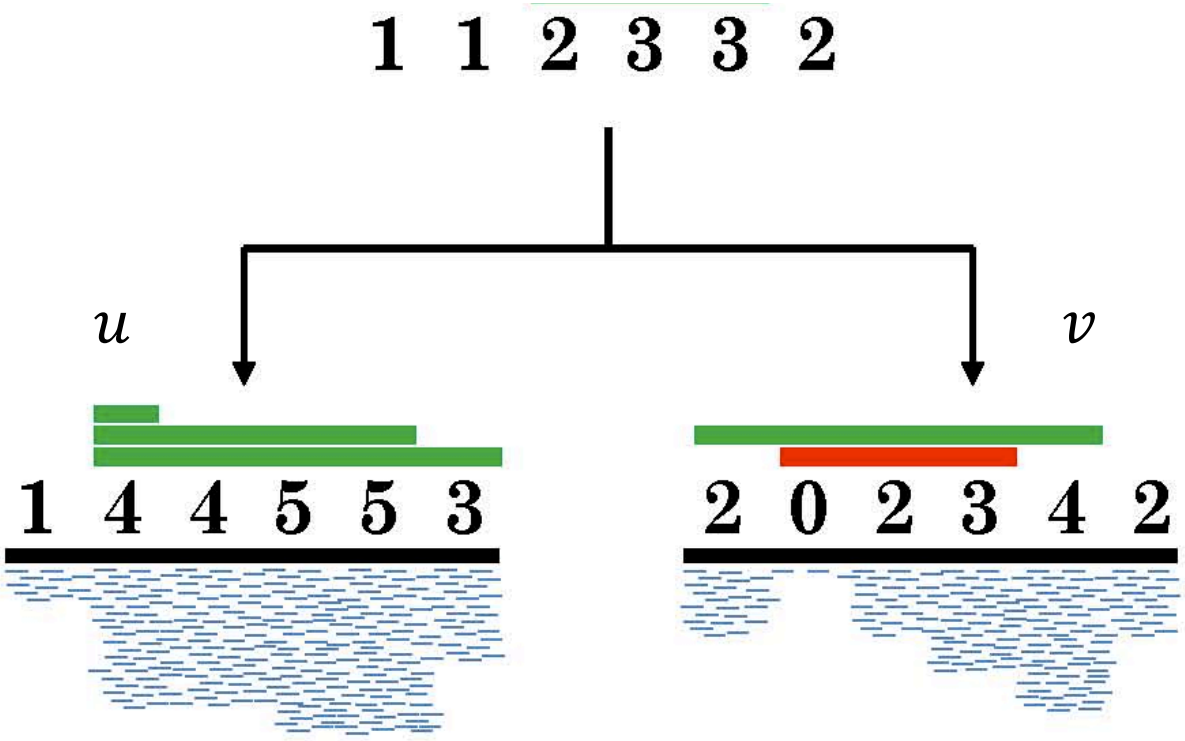
Copy-Number Triplet Problem (CN3)

- Given two profiles, find a parent profile minimizing the sum of distances to them.



Copy-Number Triplet Problem (CN3)

- Given two profiles, find a parent profile minimizing the sum of distances to them.



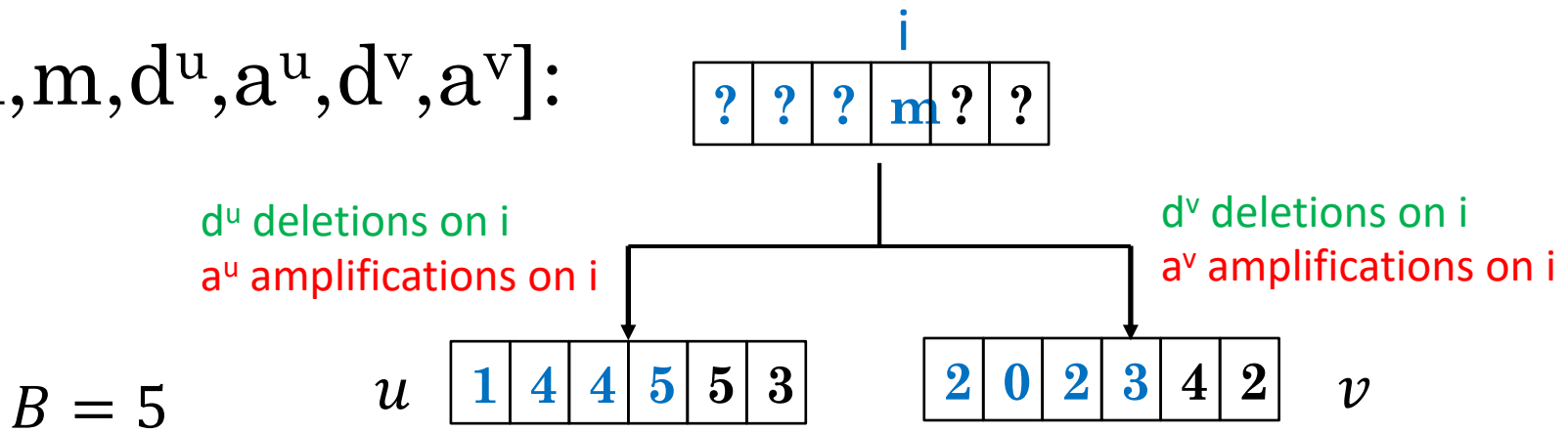
Problem properties

- Shamir et al. 2016:
  - There is an optimal S→T sorting scenario where all deletions precede all amplifications.
  - Let B be the maximal copy-number. There is an optimal S → T sorting scenario that deletes/amplifies each position ≤ B times.

- **Lemma**: There is an optimal solution to CN3 where all positions in the parent are ≤B.

Dynamic programming solution

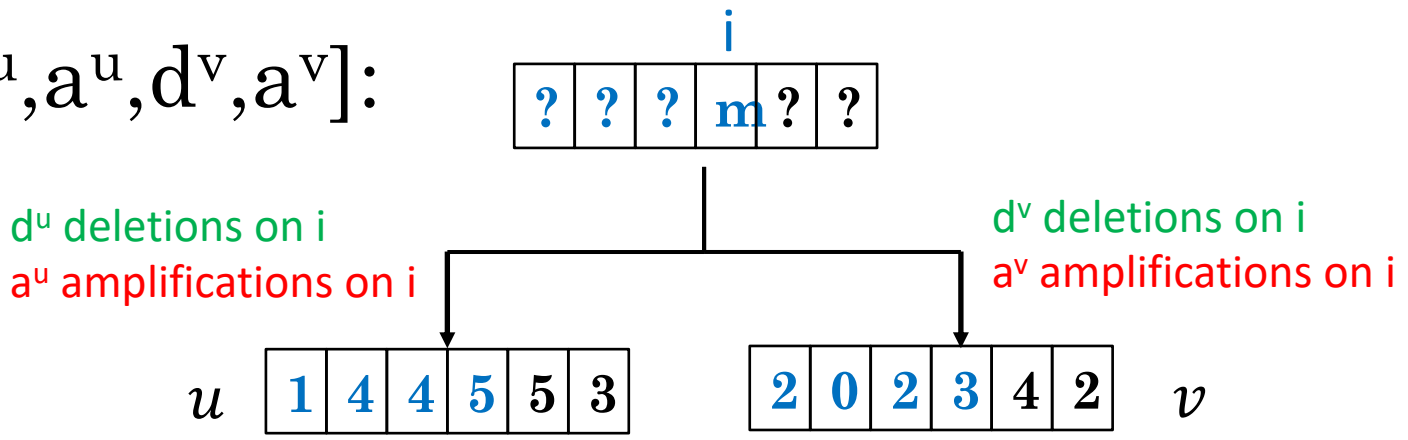
- L[i,m,d^u,a^u,d^v,a^v]:



L: Optimal value of solution for prefixes 1,..i given the values of the parameters.

Dynamic programming solution

- L[i,m,d^u,a^u,d^v,a^v]:



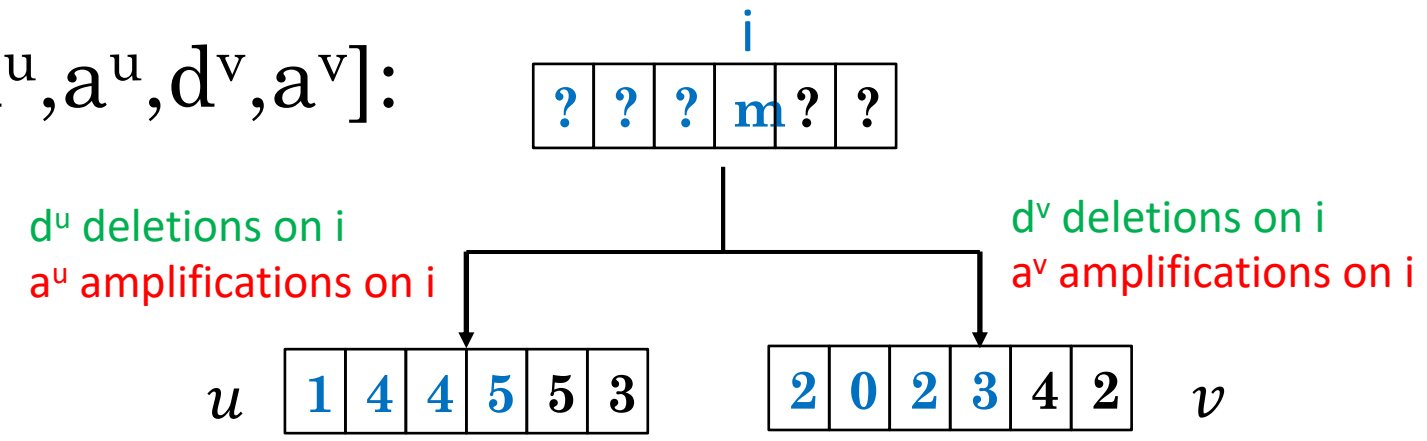
$$\mathrm{L}[i,m,d^{\mathbf{u}},a^{\mathbf{u}},d^{\mathbf{v}},a^{\mathbf{v}}] = \min_{\substack{0\le m'\le N \\ 0\le d^{\mathbf{u}\prime},a^{\mathbf{u}\prime},d^{\mathbf{v}\prime},a^{\mathbf{v}\prime}\le N}} \Big\{ \mathrm{L}[i-1,m',d^{\mathbf{u}\prime},a^{\mathbf{u}\prime},d^{\mathbf{v}\prime},a^{\mathbf{v}\prime}] + \max\{d^{\mathbf{u}}-d^{\mathbf{u}\prime},0\} + \max\{a^{\mathbf{u}}-a^{\mathbf{u}\prime},0\} + \max\{d^{\mathbf{v}}-d^{\mathbf{v}\prime},0\} + \max\{a^{\mathbf{v}}-a^{\mathbf{v}\prime},0\} \Big\}$$

new deletions

new amplifications

Dynamic programming solution

- L[i,m,d^u,a^u,d^v,a^v]:



$$\mathrm{L}[i, m, d^{\mathbf{u}}, a^{\mathbf{u}}, d^{\mathbf{v}}, a^{\mathbf{v}}] = \min_{\substack{0 \le m' \le N \\ 0 \le d^{\mathbf{u}\prime}, a^{\mathbf{u}\prime}, d^{\mathbf{v}\prime}, a^{\mathbf{v}\prime} \le N}} \Big\{ \mathrm{L}[i-1, m', d^{\mathbf{u}\prime}, a^{\mathbf{u}\prime}, d^{\mathbf{v}\prime}, a^{\mathbf{v}\prime}] + \max\{d^{\mathbf{u}} - d^{\mathbf{u}\prime}, 0\} + \max\{a^{\mathbf{u}} - a^{\mathbf{u}\prime}, 0\} + \max\{d^{\mathbf{v}} - d^{\mathbf{v}\prime}, 0\} + \max\{a^{\mathbf{v}} - a^{\mathbf{v}\prime}, 0\} \Big\}$$

new deletions

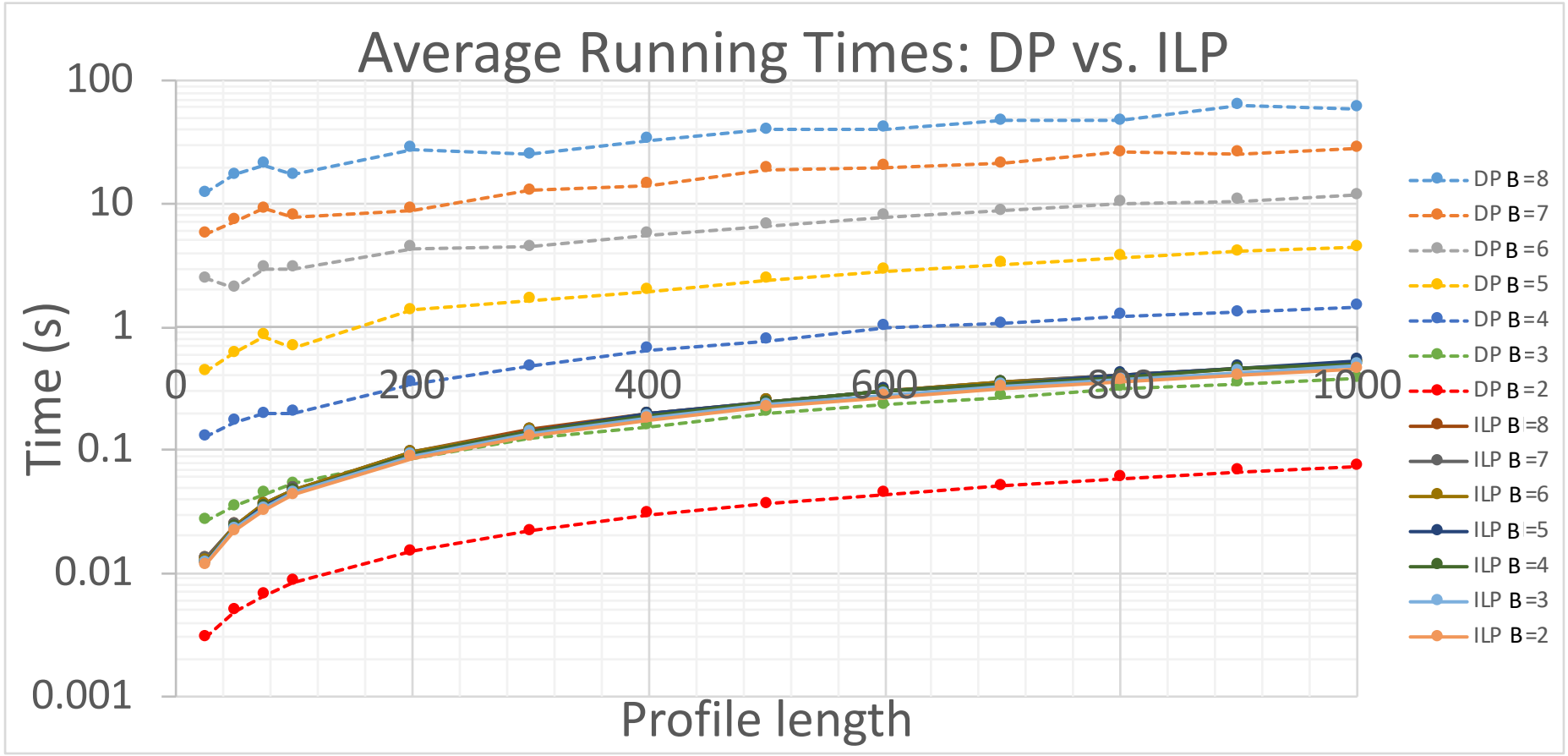
new amplifications

$O(nB^{10})$ time
$O(nB^5)$ space

Can be improved to $O(nB^7)$ time, $O(nB^4)$ space

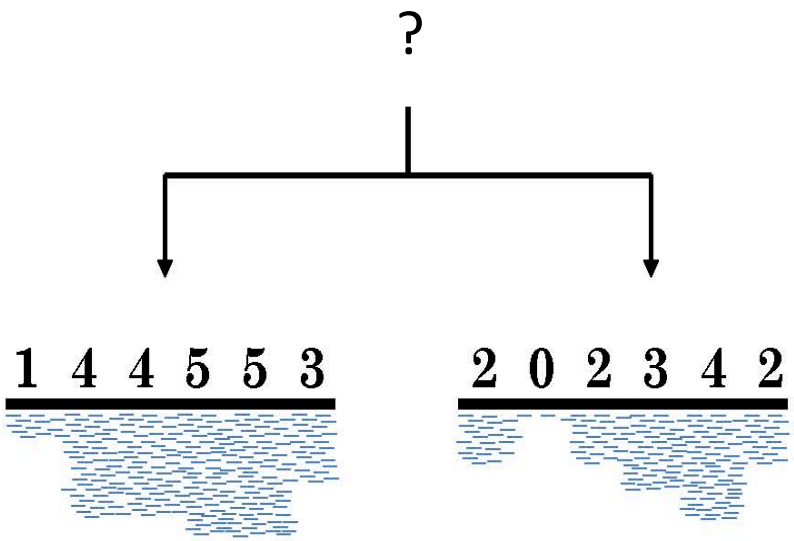
CN3 simulations results

- Comparing DP and an ILP with O(n) variables and constraints

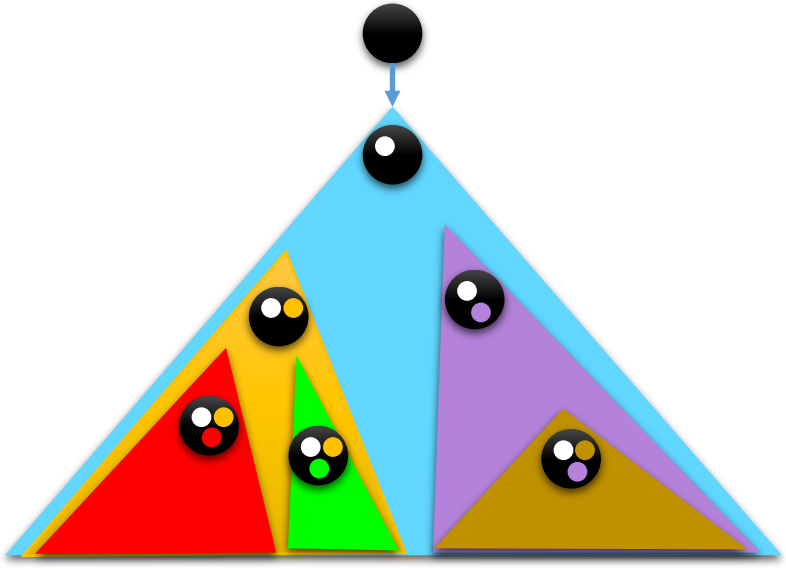


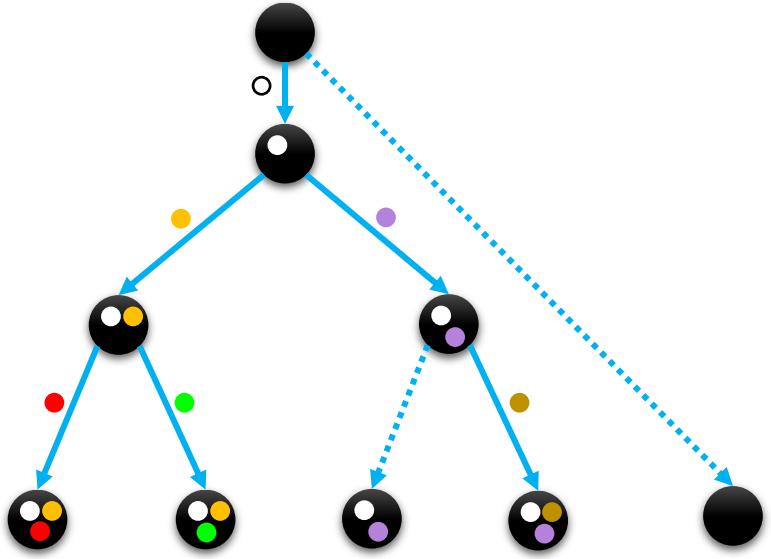
CN3 Conclusions

- DP alg for CN3: linear in n, pseudo-polynomial ($O(nB^7)$)
- An ILP with O(n) variables and constraints
- In simulations: ILP runs faster than DP, obtains the optimum, independent of B
- Complexity of CN3 is still open!

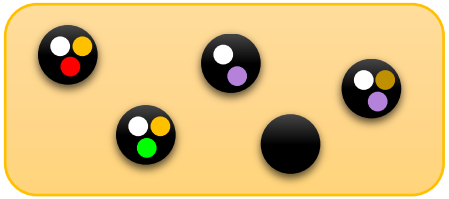


Tree for tumor evolution





Extant clones inferred from multi samples

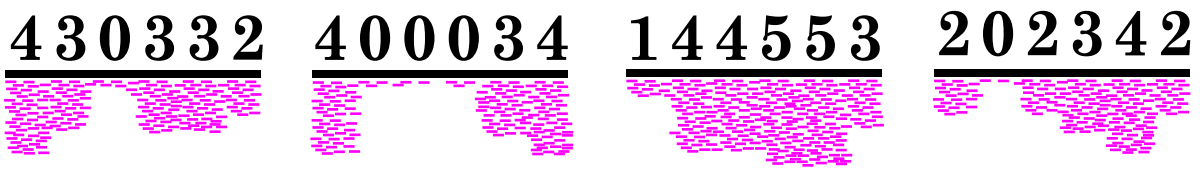


Evolutionary history of clones is modelled by a phylogenetic tree:

1. Leaves correspond to extant clones
2. Edges are labeled by mutations

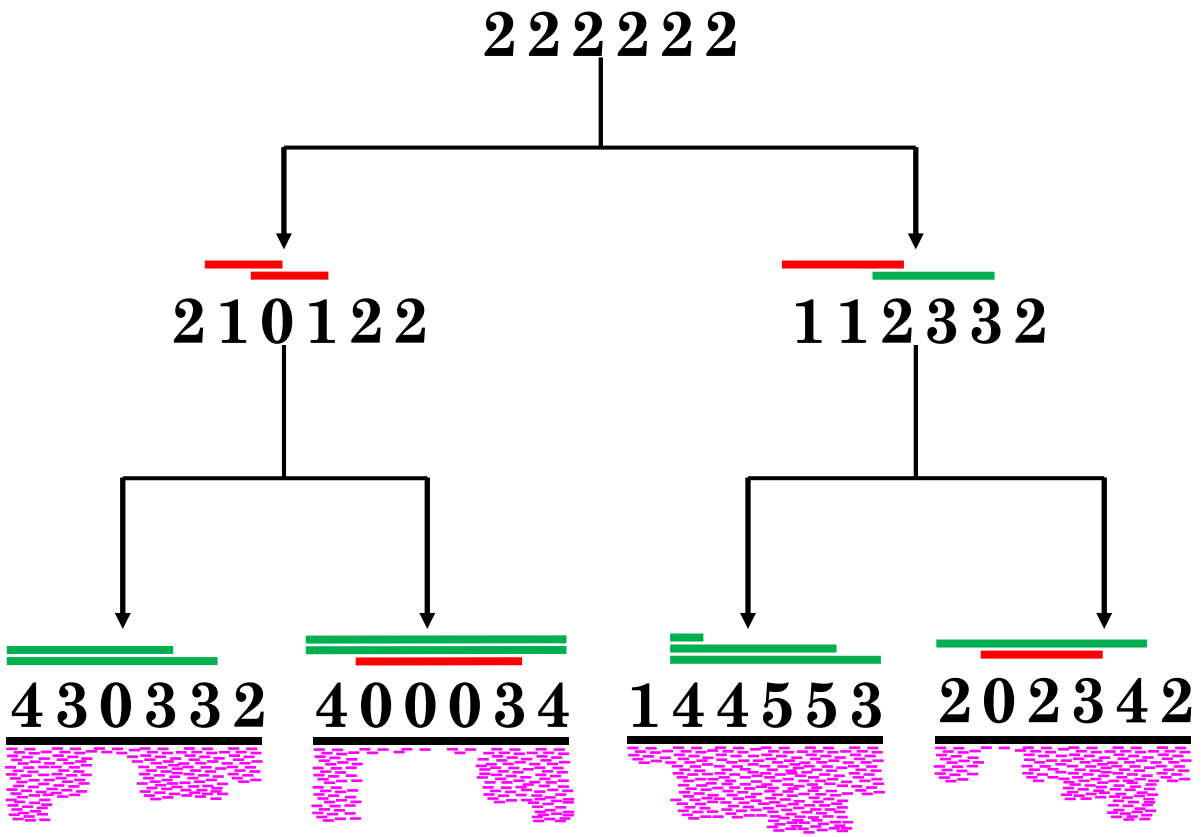
Copy-Number Tree (CNT)

- Input: collection of CN profiles



Copy-Number Tree (CNT)

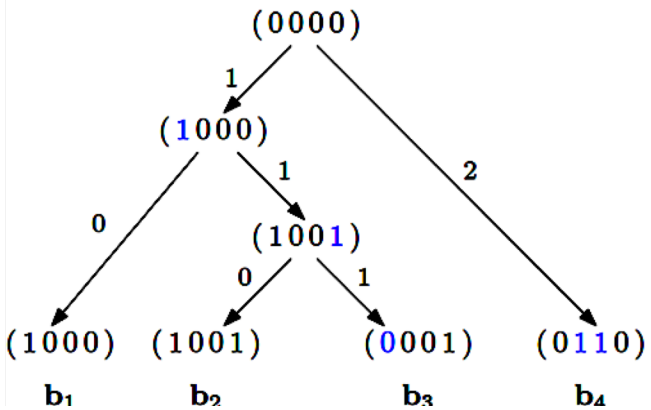
- Input: collection of CN profiles
- Output: copy-number phylogeny:
  - Rooted in the normal diploid profile
  - The leaves are the input profiles
  - Explained by the minimum number of events



Computational Complexity

- CNT is NP-hard
- Reduction from the NP-hard Steiner Problem in Phylogeny (also called Maximum Parsimony Phylogeny):
  - Binary vectors
  - Events are single flips

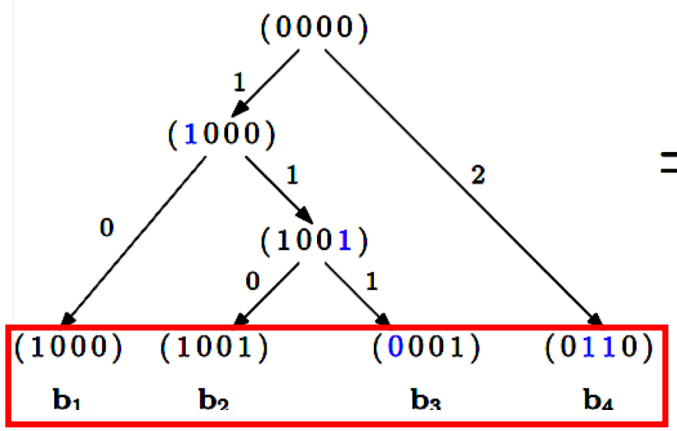
MPP instance and solution T with cost $\Delta(T) = 5$



Computational Complexity

- Transformation:
  - 0 → 2 and 1 → 1 (real copy-numbers)
  - A wall Ω (21⋯12) between consecutive real copy-numbers, large enough

MPP instance and solution T with cost $\Delta(T) = 5$



$\Rightarrow$

$\Omega = (2 \quad 1 \quad \cdots \quad 1 \quad 2)$

$|\Omega| = nk = 20$

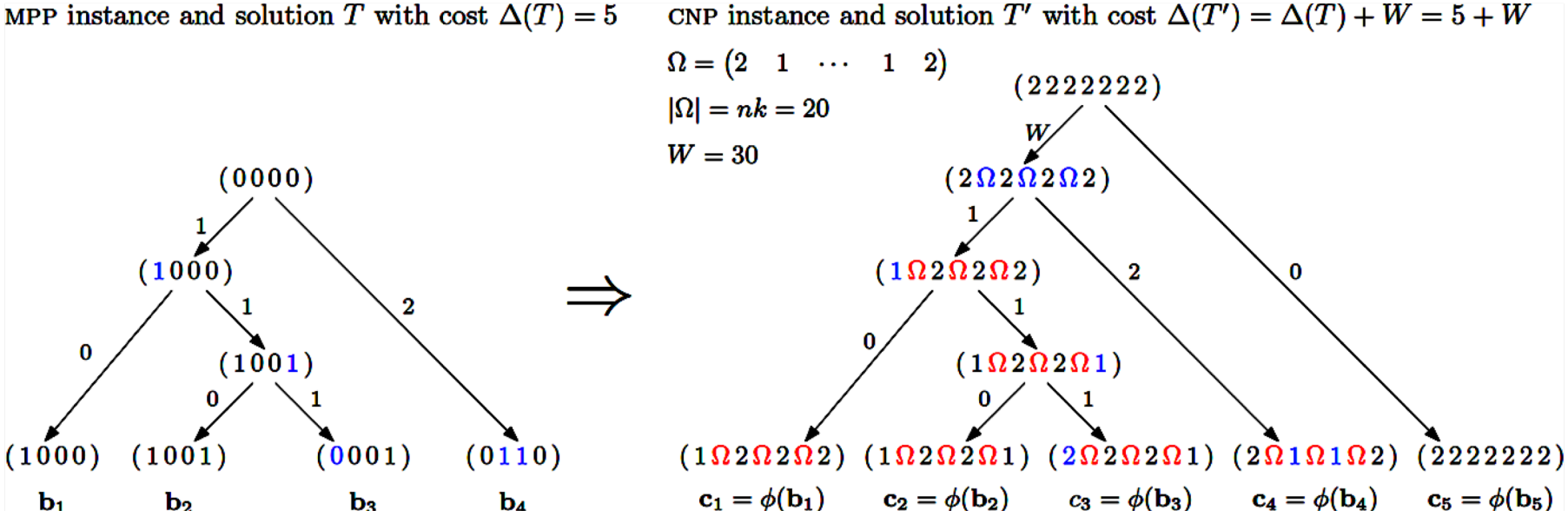
$(1\Omega 2\Omega 2\dot{\Omega}2)$ $(1\Omega 2\Omega\dot{2}\Omega 1)$ $(2\Omega\dot{2}\Omega 2\Omega 1)$ $(2\dot{\Omega}1\Omega 1\Omega 2)$ $(\dot{2}222222)$

$\mathbf{c}_1 = \phi(\mathbf{b}_1)$ $\mathbf{c}_2 = \phi(\mathbf{b}_2)$ $\mathbf{c}_3 = \phi(\mathbf{b}_3)$ $\mathbf{c}_4 = \phi(\mathbf{b}_4)$ $\mathbf{c}_5 = \phi(\mathbf{b}_5)$

Computational Complexity

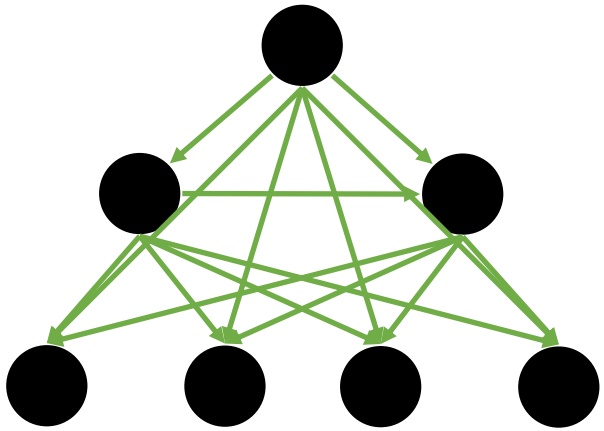
- Key observation: no amplification or deletion breaches the wall Ω

  ➔ Amplifications and deletions correspond to flips



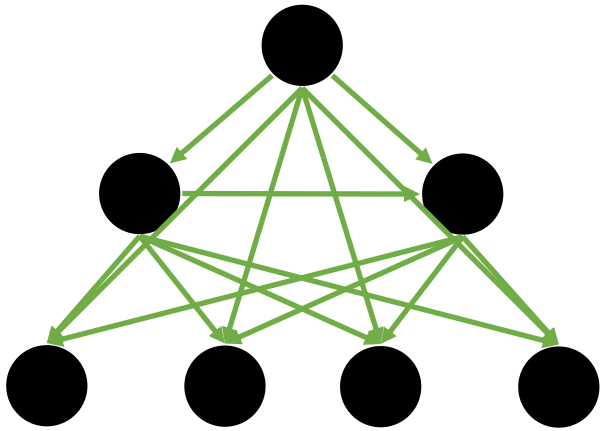
ILP Model

1. Build a spanning tree from a DAG that contains all the possibilities



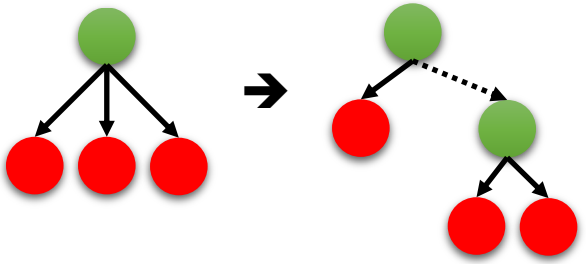
ILP Model

1. Build a spanning tree from a DAG that contains all the possibilities

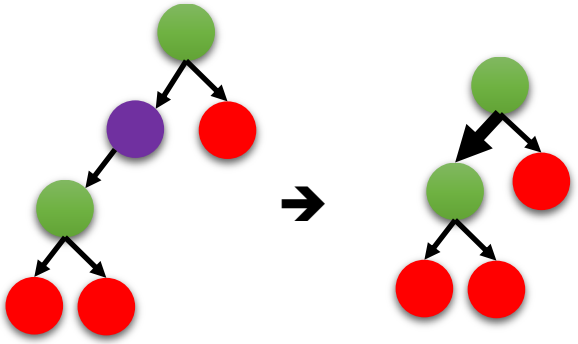


Assumptions

1. Tree is binary, w.l.o.g. by splitting high degree vertices

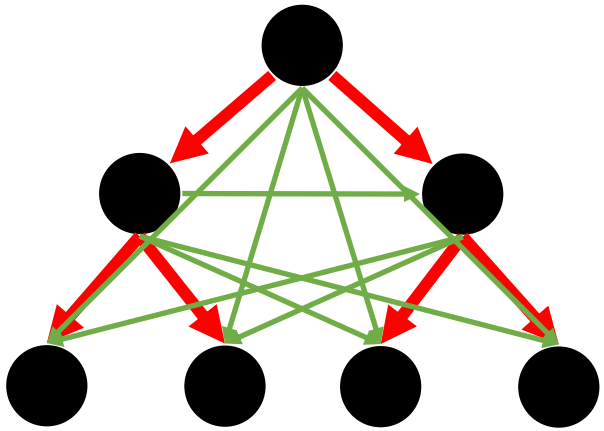


2. Tree is full binary, w.l.o.g. by collapsing outdegree-2 internal non-root vertices and solving an additional instance adding the root to the leaves



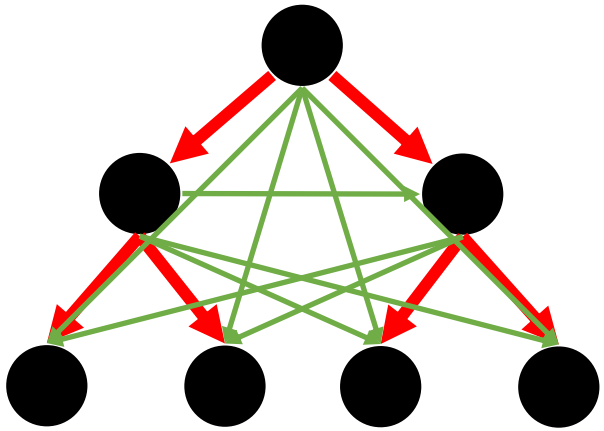
ILP Model

1. Build a spanning tree from a DAG that contains all the possibilities



ILP Model

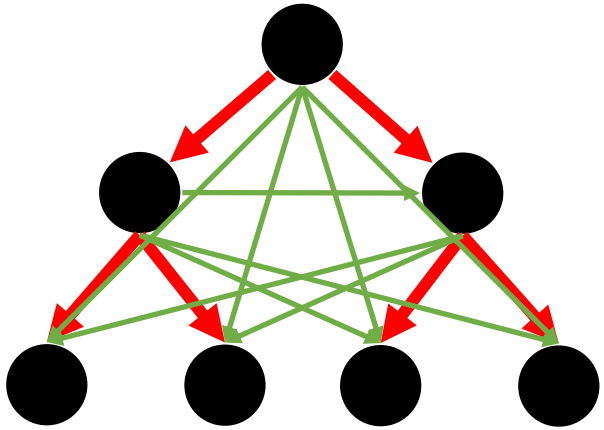
1. Build a spanning tree from a DAG that contains all the possibilities



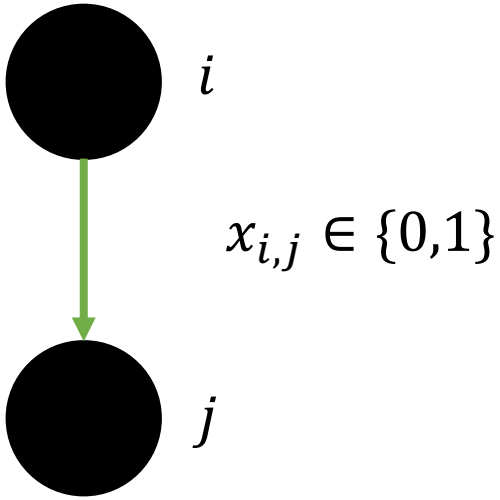
2. Compute the labeling of the internal vertices and the cost of the edges in order to minimize number of events

ILP Model

1. Build a spanning tree from a DAG that contains all the possibilities

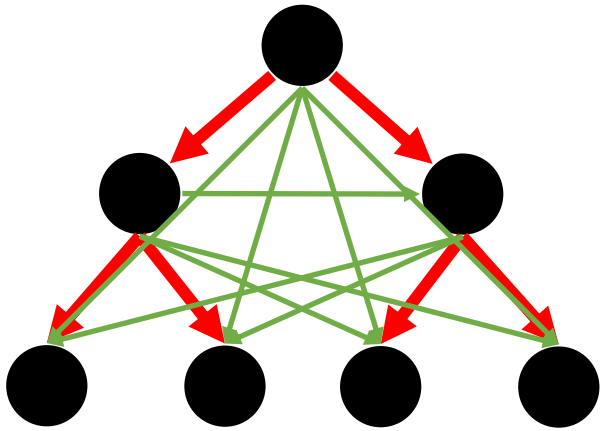


2. Compute the labeling of the internal vertices and the cost of the edges in order to minimize number of events

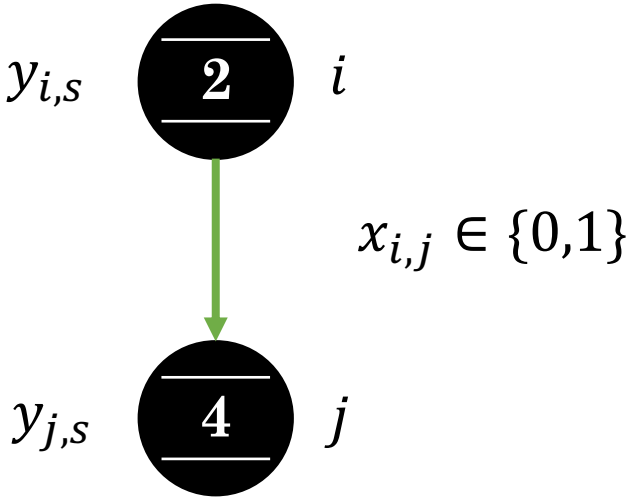


ILP Model

1. Build a spanning tree from a DAG that contains all the possibilities

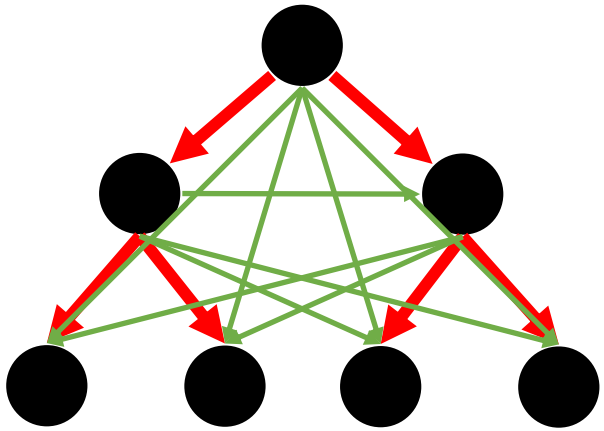


2. Compute the labeling of the internal vertices and the cost of the edges in order to minimize number of events

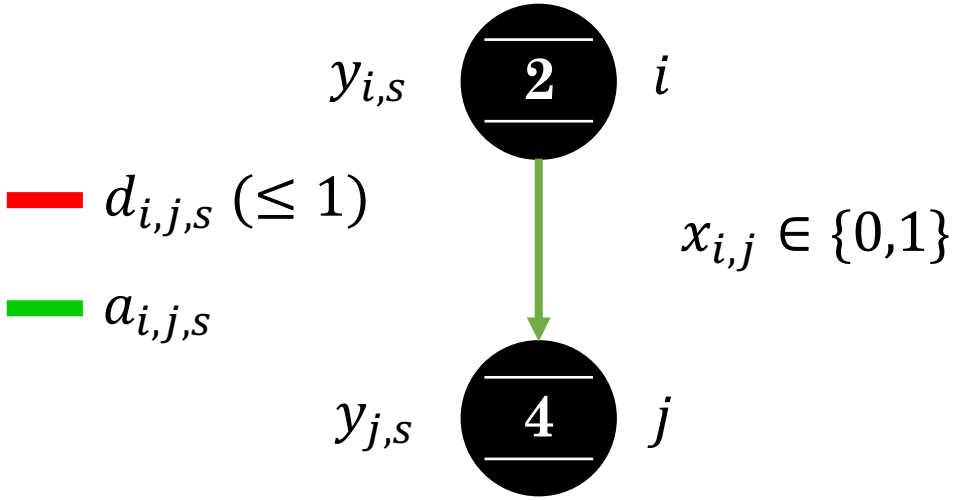


ILP Model

1. Build a spanning tree from a DAG that contains all the possibilities



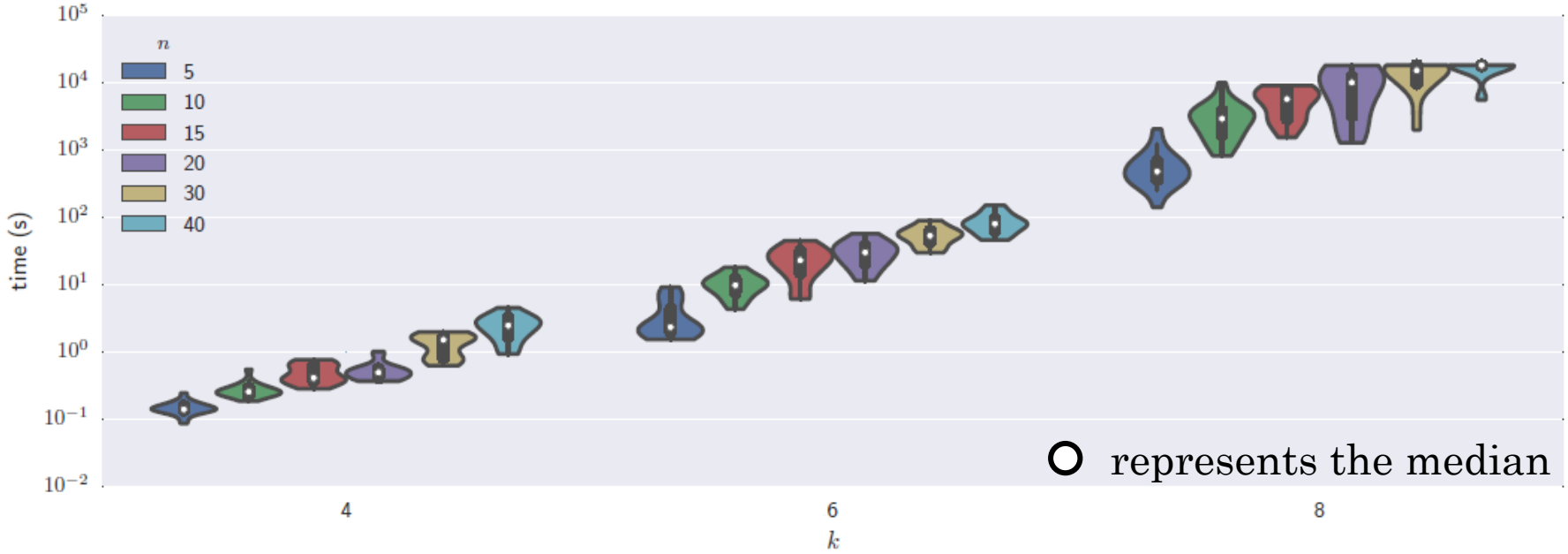
2. Compute the labeling of the internal vertices and the cost of the edges in order to minimize number of events



$O(k^2n + kn \log e)$ variables and constraints

Simulations: Running time

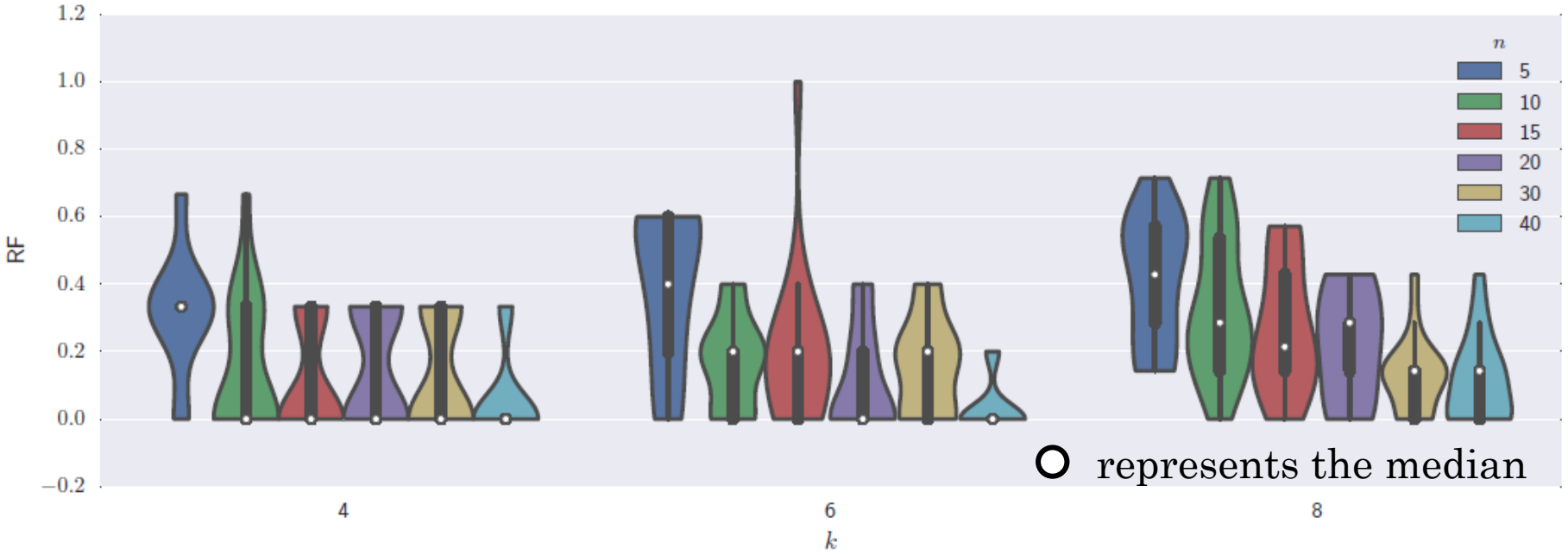
- Implemented in C++ with CPLEX 12.3
- Simulated instances (1 chromosome) were generated varying the profile's length (n), the number of leaves (k), and number of events:
  ➔ $n \in \{5,10,15,20,30,40\}$ and $k \in \{4,6,8\}$ are realistic



(a) Running time in seconds (log scale)

Simulations: Accuracy

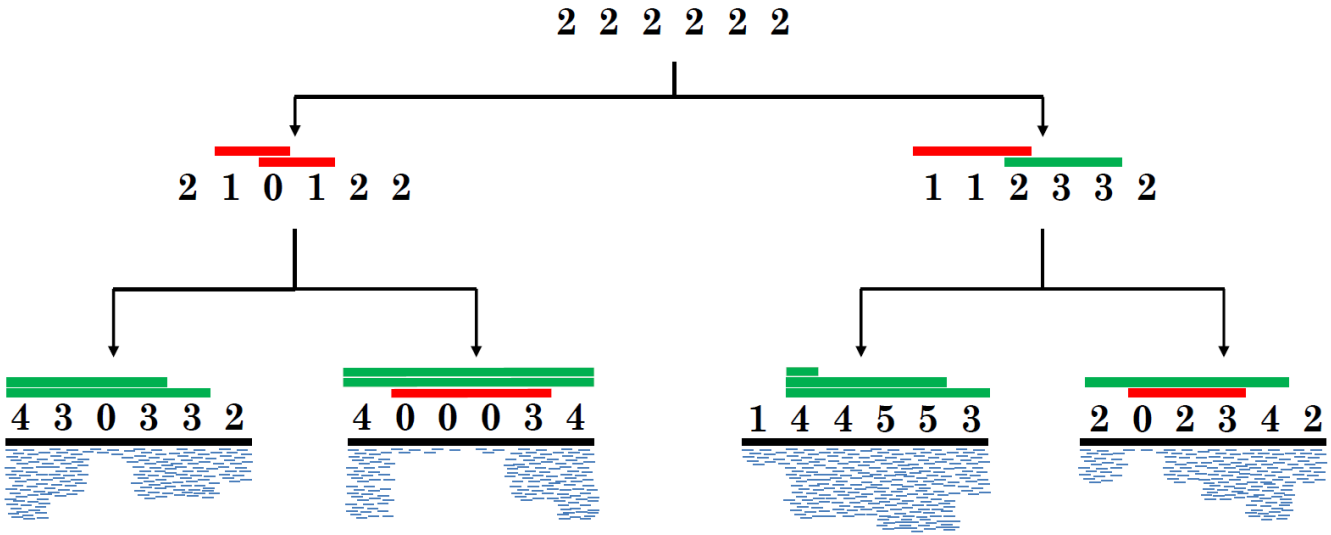
- RF metric – symmetric difference between partitions of two trees – as measure of topological accuracy
- 0 corresponds to identical topologies



(b) Normalized Robinson-Foulds (RF) metric

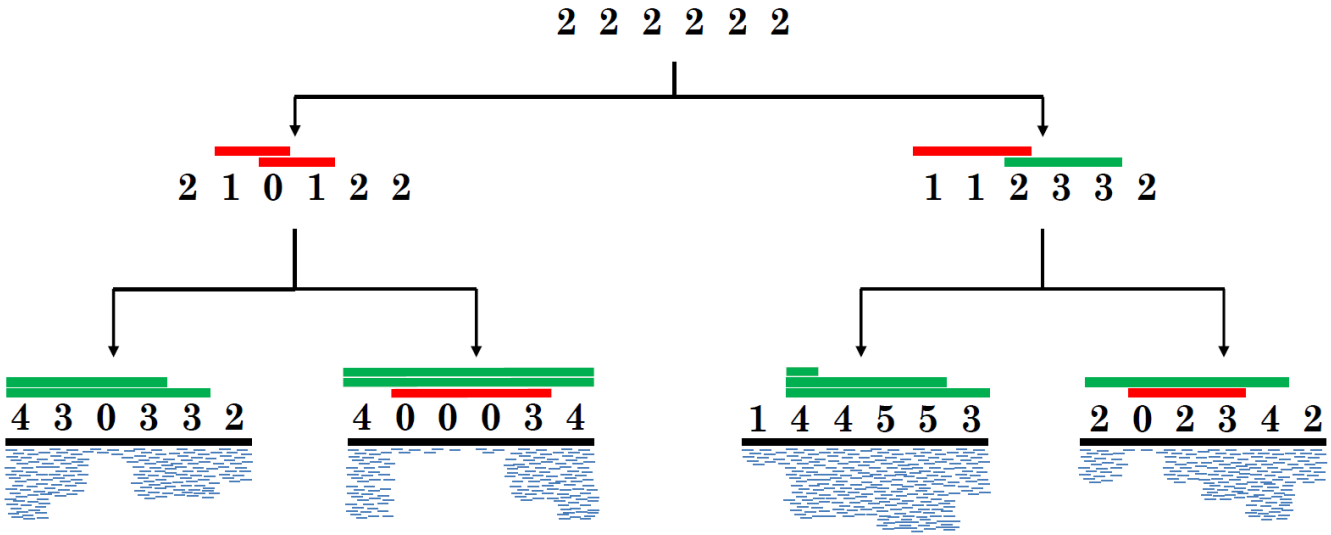
CNT Conclusions

- CNT: NP-hardness, solving algorithm (ILP)
- The results of ILP on simulated data show good accuracy on instances of real size
- Next step: experiments on real data ➔ need of dealing with additional factors as diploid genomes



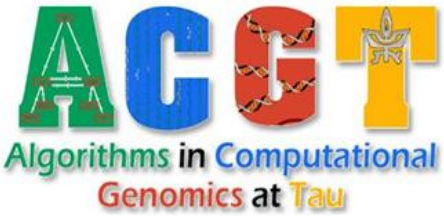
Open Questions

- Computational complexity of CN3
- Computational complexity of the ‘small phylogeny’ for CNT (with fixed topology)



Thanks for the attention.

Questions?



Ron Shamir

Roded Sharan

Meirav Zehavi

Ron Zeira

Mohammed El-Kebir

Ben Raphael

Simone Zaccaria