CIMICE: Markov Chain Inference Method to Identify Cancer Evolution

Nicolò Rossi Nicola Gigante

Carla Piazza Nicola Vitacolonna

Dip. di Scienze Matematiche, Informatiche e Fisiche - UniUD

Premises

The Context

- Investigation on the mutational history of a cancer cell
- Relying on single cell data at a single time instant
- For the reconstruction a probabilistic model

The Aim

- Identification of a minimal set of assumptions on models/data
- Detection of the sources of uncertainty in the reconstruction
- Provision of suggestions for further experiments

Premises

The Context

- Investigation on the mutational history of a cancer cell
- Relying on single cell data at a single time instant
- For the reconstruction a probabilistic model

The Aim

- Identification of a minimal set of assumptions on models/data
- Detection of the sources of uncertainty in the reconstruction
- Provision of suggestions for further experiments

Our Results

Model Reconstruction

We find a minimal set of assumptions such that:

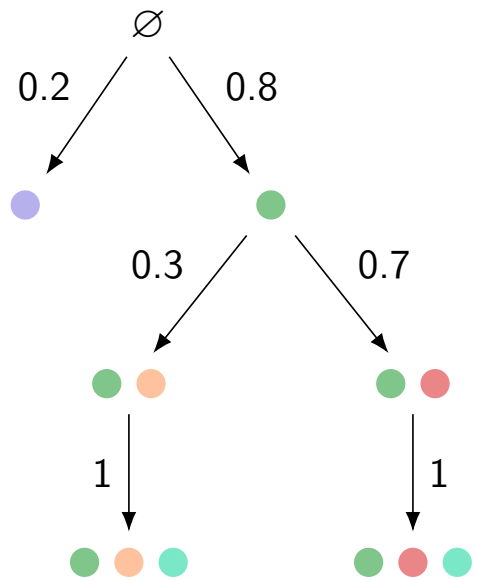
- without convergent evolutionary paths
  - there is one probabilistic underlying model
  - CIMICE infers it in efficiently w.r.t. the data size
- with convergent evolutionary paths
  - there is an infinite set of possible models
  - CIMICE heuristically assigns weights on convergences to pick a preferred one

Our Results

Model Reconstruction

We find a minimal set of assumptions such that:

- without convergent evolutionary paths
  - there is one probabilistic underlying model
  - CIMICE infers it in efficiently w.r.t. the data size
- with convergent evolutionary paths
  - there is an infinite set of possible models
  - CIMICE heuristically assigns weights on convergences to pick a preferred one

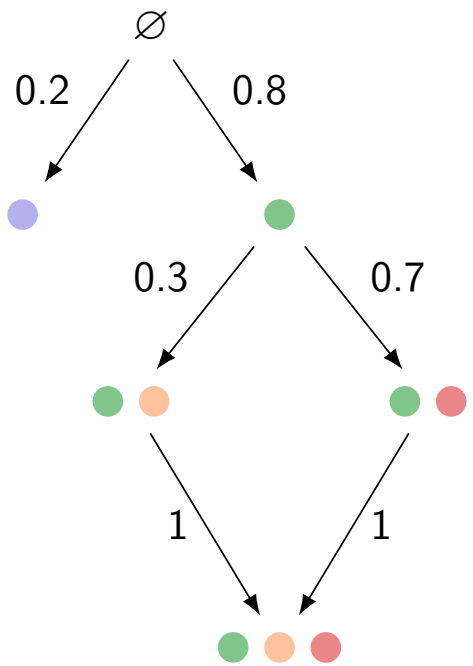


Our Results

Model Reconstruction

We find a minimal set of assumptions such that:

- without convergent evolutionary paths
  - there is one probabilistic underlying model
  - CIMICE infers it in efficiently w.r.t. the data size
- with convergent evolutionary paths
  - there is an infinite set of possible models
  - CIMICE heuristically assigns weights on convergences to pick a preferred one

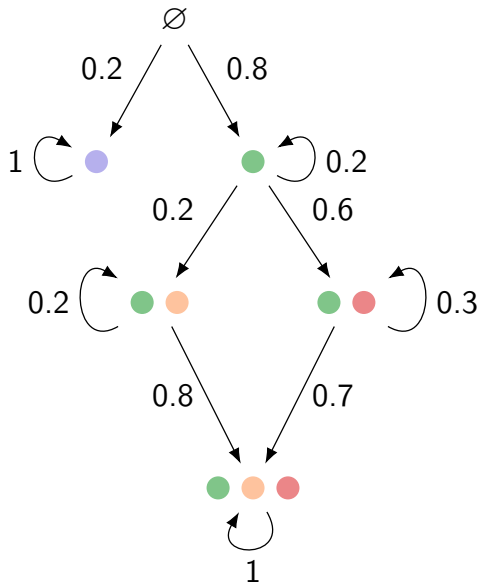


Our Results

Generative Models

Whenever our biological assumptions are reasonable, CIMICE produces synthetic data to

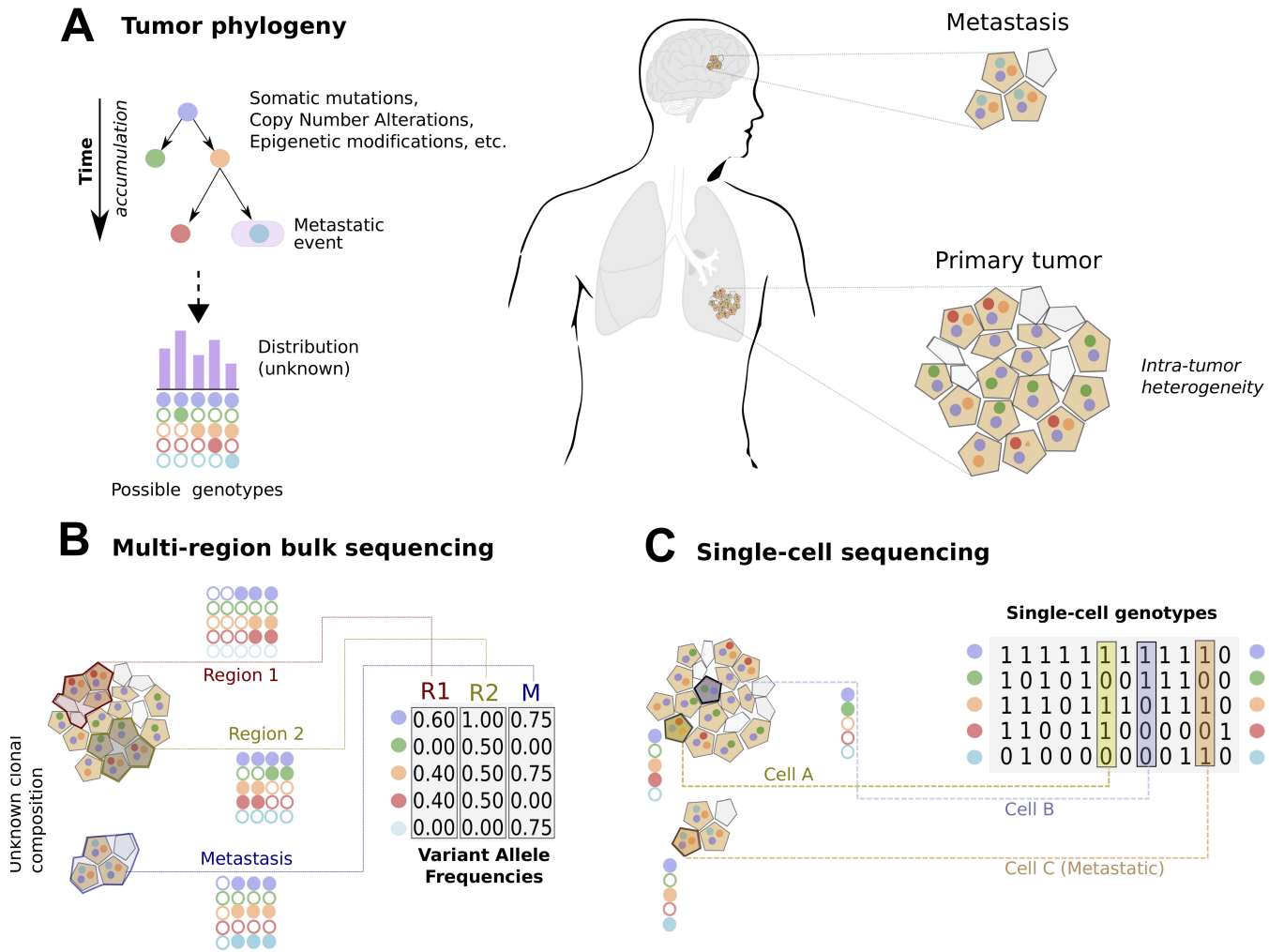
- generate more data from an inferred model
- test different model inference methods



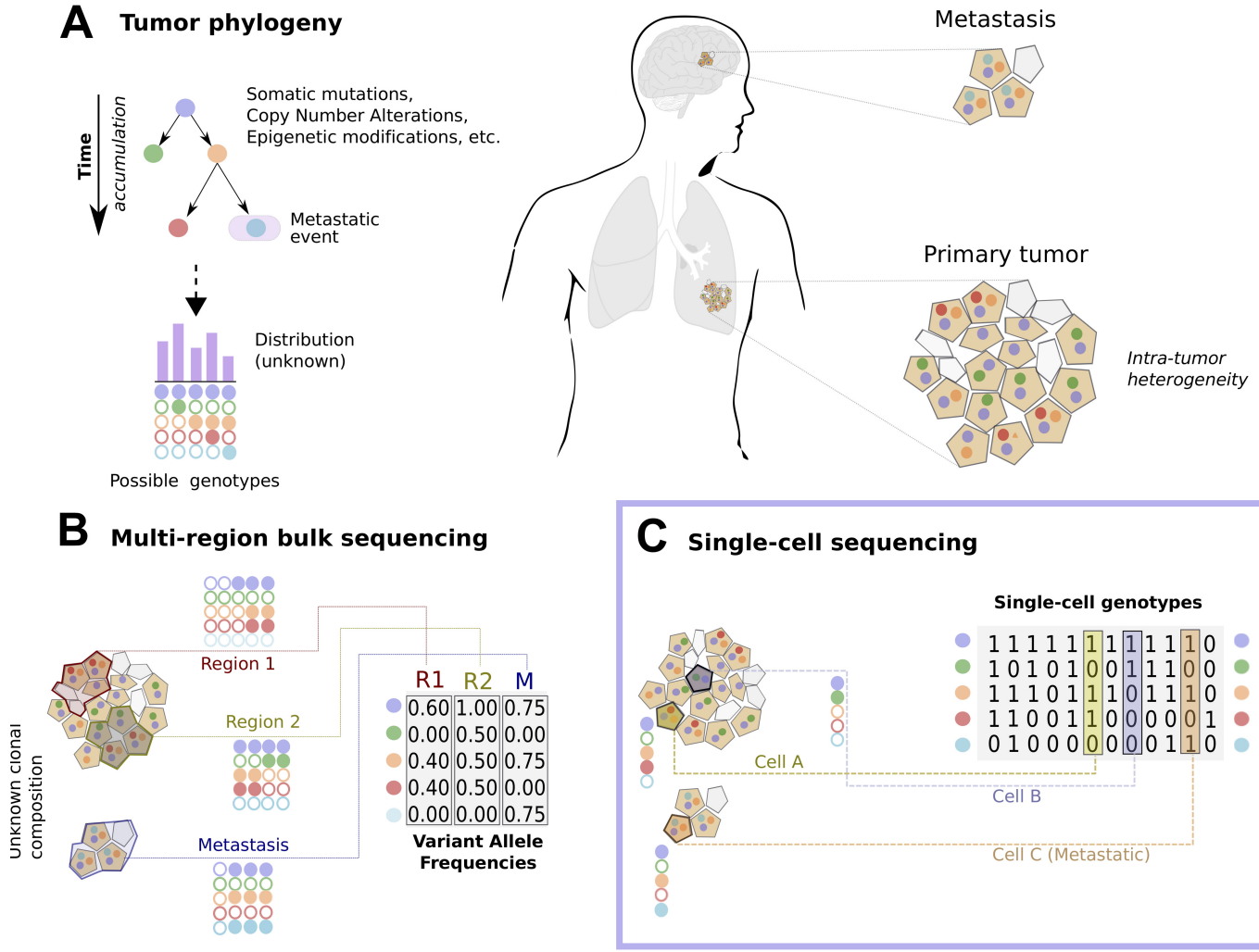
Plan of the Talk

1. Single Cell Data
2. From Biological Assumptions to Models
3. Two Reconstruction Problems
4. Our Inference Algorithm
5. CIMICE Tool
6. Synthetic Models and Tests
7. Real Data
8. Conclusion

Single Cell Data



Single Cell Data



From Biological Assumptions to Models

Cancer Progression Markov Chains (CPMC)

Biological Assumptions

- $\cup$ Along cancer progression cells accumulate gene mutations
- $\emptyset$ Initially there are no mutated genes
- MC New mutations probabilistically depends only on the current genotype
- $\forall$ Each time a minimal number of mutations is acquired

Models: CPMC

- The states are the genotypes
- The healthy genotype is the source
- It is a Discrete Time Markov Chain
- It is acyclic and anti-transitive

From Biological Assumptions to Models

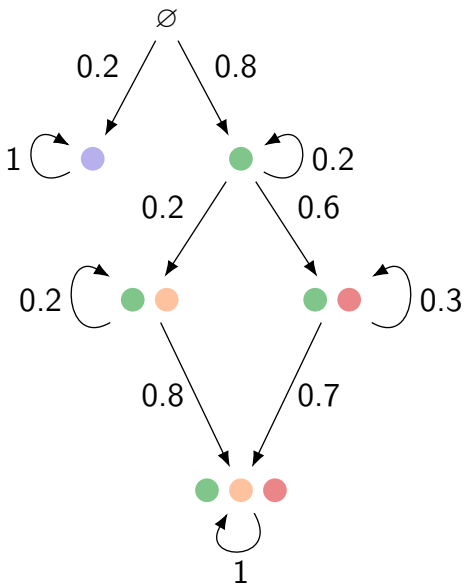
Cancer Progression Markov Chains (CPMC)

Biological Assumptions

- $\cup$ Along cancer progression cells accumulate gene mutations
- $\emptyset$ Initially there are no mutated genes
- MC New mutations probabilistically depends only on the current genotype
- $\forall$ Each time a minimal number of mutations is acquired

Models: CPMC

- The states are the genotypes
- The healthy genotype is the source
- It is a Discrete Time Markov Chain
- It is acyclic and anti-transitive



From Biological Assumptions to Models

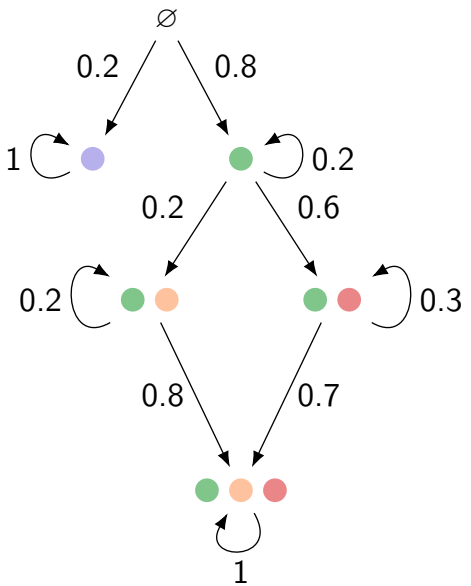
Cancer Progression Markov Chains (CPMC)

Biological Assumptions

- $\cup$ Along cancer progression cells accumulate gene mutations
- $\emptyset$ Initially there are no mutated genes
- MC New mutations probabilistically depends only on the current genotype
- $\forall$ Each time a minimal number of mutations is acquired

Models: CPMC

- The states are the genotypes
- The healthy genotype is the source
- It is a Discrete Time Markov Chain
- It is acyclic and anti-transitive



From Biological Assumptions to Models

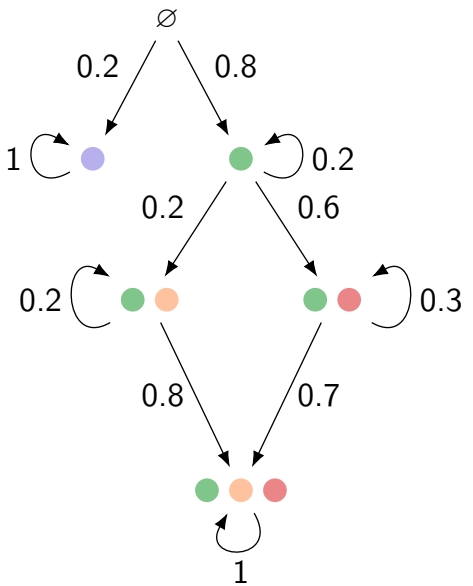
Cancer Progression Markov Chains (CPMC)

Biological Assumptions

- $\cup$ Along cancer progression cells accumulate gene mutations
- $\emptyset$ Initially there are no mutated genes
- MC New mutations probabilistically depends only on the current genotype
- $\forall$ Each time a minimal number of mutations is acquired

Models: CPMC

- The states are the genotypes
- The healthy genotype is the source
- It is a Discrete Time Markov Chain
- It is acyclic and anti-transitive



From Biological Assumptions to Models

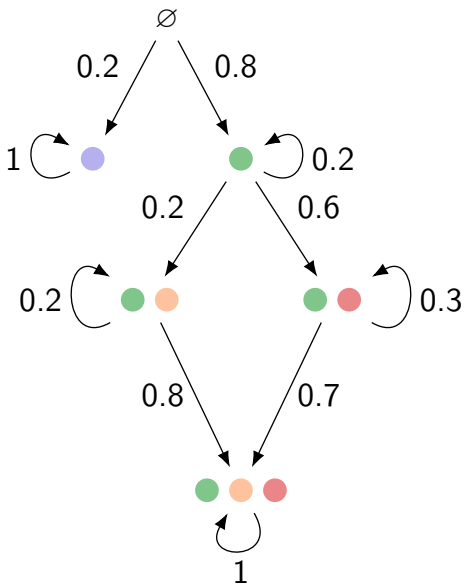
Cancer Progression Markov Chains (CPMC)

Biological Assumptions

- $\cup$ Along cancer progression cells accumulate gene mutations
- $\emptyset$ Initially there are no mutated genes
- MC New mutations probabilistically depends only on the current genotype
- $\forall$ Each time a minimal number of mutations is acquired

Models: CPMC

- The states are the genotypes
- The healthy genotype is the source
- It is a Discrete Time Markov Chain
- It is acyclic and anti-transitive



From Biological Assumptions to Models

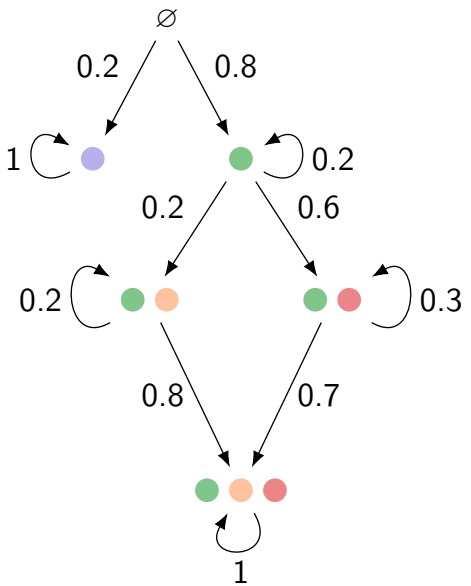
Cancer Progression Markov Chains (CPMC)

Biological Assumptions

- $\cup$ Along cancer progression cells accumulate gene mutations
- $\emptyset$ Initially there are no mutated genes
- MC New mutations probabilistically depends only on the current genotype
- $\forall$ Each time a minimal number of mutations is acquired

Models: CPMC

- The states are the genotypes
- The healthy genotype is the source
- It is a Discrete Time Markov Chain
- It is acyclic and anti-transitive



What's new?

Main differences w.r.t. literature

We propose a method that:

- refers to single cell DNA-seq data
- assumes clean and rich data
- points out where more knowledge is needed
- exploits a deterministic approach, which can be extended to include statistical methods and expert knowledge
- is patient-driven and suitable to study treatment effects

Two Reconstruction Problems

Single Time Input:

D_i

a single cell data matrix

Many Time Input:

$D_0, D_1, \ldots D_t$

a time sequence of matrices

Output:

A CPMC whose simulation would generate the data

In such CPMC time ticks at each mutational event (no self-loops)

Remark

- It is not a standard Markov Chain reconstruction problem
- There can be infinitely many solutions Appendix

Two Reconstruction Problems

Single Time Input:

D_i

a single cell data matrix

Many Time Input:

$D_0, D_1, \ldots D_t$

a time sequence of matrices

Output:

A CPMC whose simulation would generate the data

In such CPMC time ticks at each mutational event (no self-loops)

Remark

- It is not a standard Markov Chain reconstruction problem
- There can be infinitely many solutions Appendix

Our Inference Algorithm

Fundamentals

- the topology is directly induced by $\cup$ and $\triangledown$
- without convergencies, the probabilities can be directly computed thanks to the topology, MC, and Bayes' theorem
- with convergent paths, we exploit heuristics to estimate *backward probabilities* Appendix

Convergency Ambiguities

s_1	1	0
s_2	0	1
s_3	1	1
s_4	1	1

$\Longrightarrow$

Our Inference Algorithm

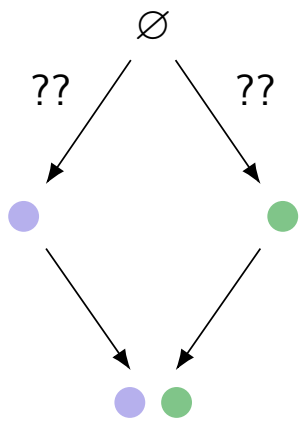
Fundamentals

- the topology is directly induced by $\cup$ and $\triangledown$
- without convergencies, the probabilities can be directly computed thanks to the topology, MC, and Bayes' theorem
- with convergent paths, we exploit heuristics to estimate *backward probabilities* Appendix

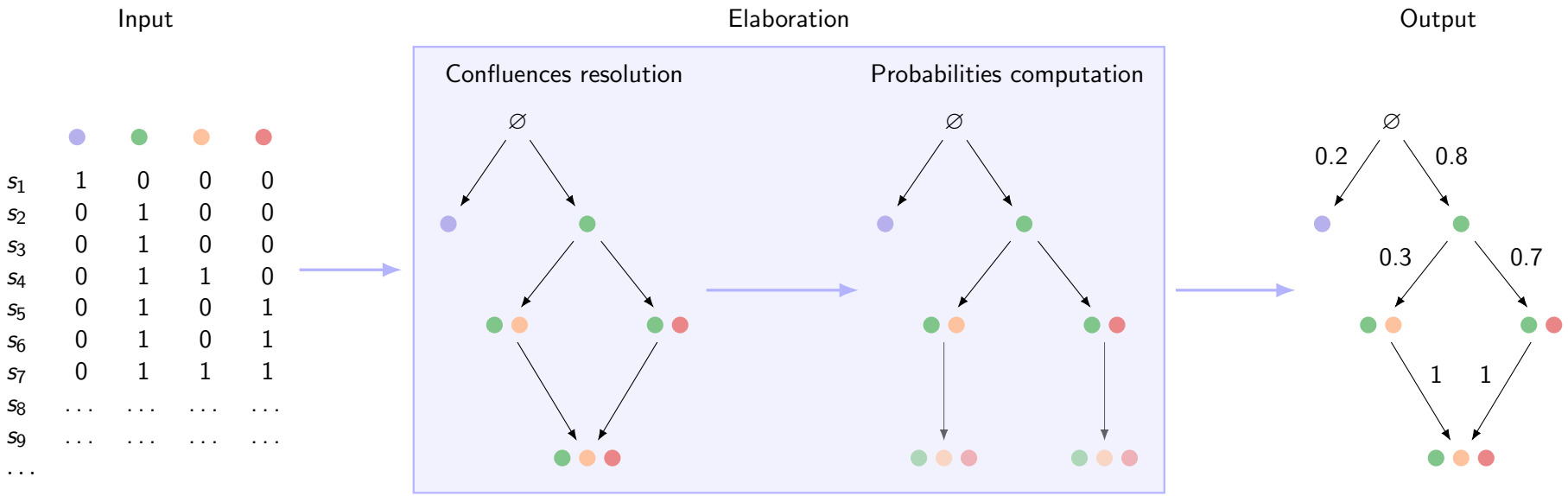
Convergency Ambiguities

s_1	1	0
s_2	0	1
s_3	1	1
s_4	1	1

$\Longrightarrow$



CIMICE Tool

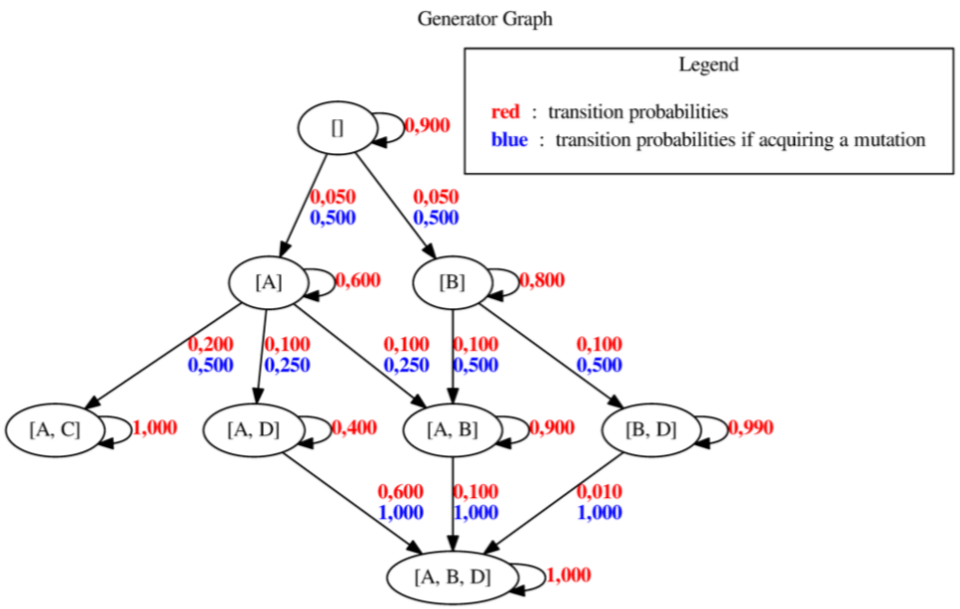


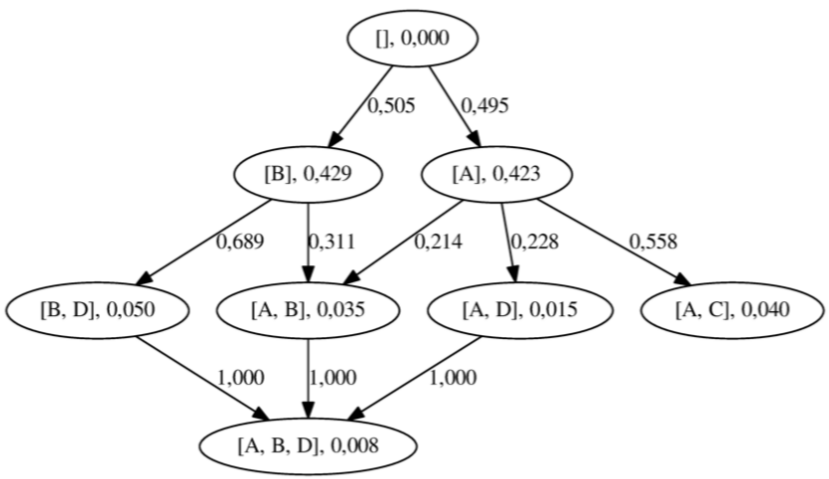
https://github.com/redsnic/tumorEvolutionWithMarkovChains

Synthetic Models and Tests

Random walks on a given CPMC can be used to either:

- generate more data on a specific case or
- generate the genotypes of the "artificial" dataset and then test our methodology



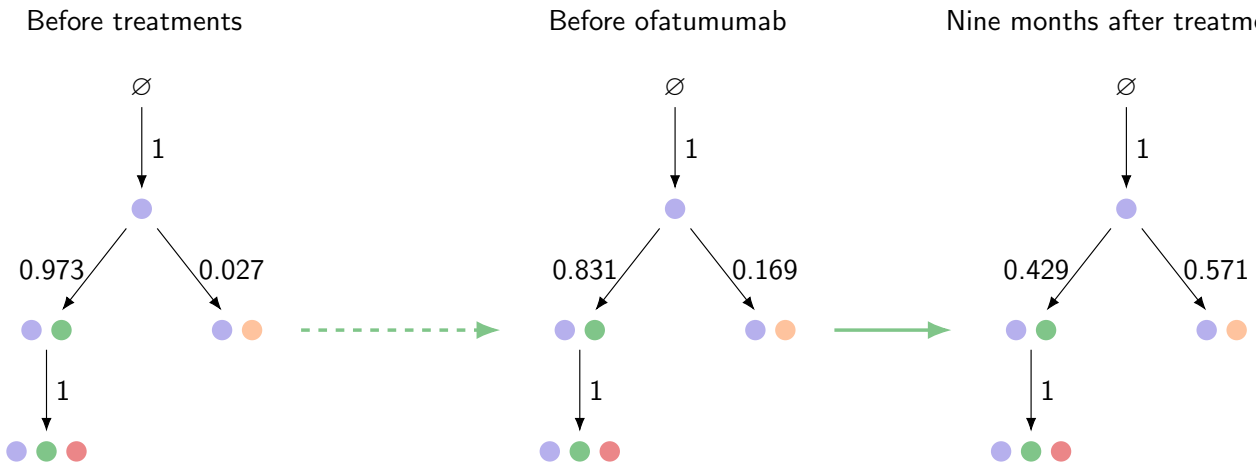


Real Data

Present situation

- small single cell datasets are available
- in Ogundijo and Wang, BMC Bioinf. 20:6(2019) a method for inferring genotype proportions from is described
- their output is a suitable input for us

Chronic Lymphocytic Leukemia - Patient CLL077



Drugs: chlorambuci, fludarabine and cyclophosphamide, ofatumumab

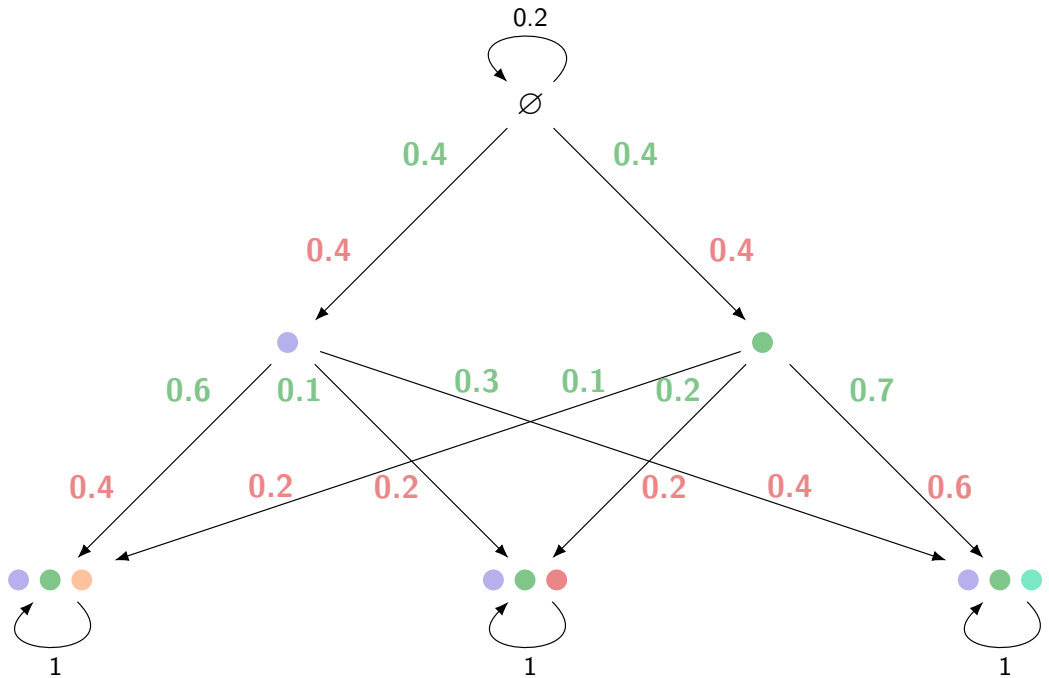
Conclusions

- We proposed a method for tumor phylogeny reconstruction using DTMCs
- The reconstruction algorithm is time efficient, can easily be enhanced with statistical methods, and it is suitable for very large datasets
- We are planning to include treatments information directly in the model
- It is possible to use the artificial data produced by CIMICE to test similar tools

Related Works

- *The evolution of tumour phylogenetics: principles and practice* R. Schwartz and A. A. Schäffer, Nature Reviews Genetics 18:213-229, 2017
- *Learning mutational graphs of individual tumour evolution from single-cell and multi-region sequencing data* D. Ramazzotti, A. Graudenzi, L. De Sano, M. Antoniotti, and G. Caravagna, BMC Bioinformatics 20:210, 2019
- *SeqClone: sequential Monte Carlo based inference of tumor subclones* Oyetunji E. Ogundijo and Xiaodong Wang, BMC Bioinformatics 20:6, 2019

Infinitely Many Solutions



	∅					
D_1	0.2	0.4	0.4	0	0	0
D_2	0.04	0.08	0.08	0.28	0.12	0.4
D_3	0.008	0.016	0.016	0.336	0.144	0.48
D_4	0.0016	0.0032	0.0032	0.3472	0.1488	0.496

Back to Presentation

Technical Details

- the data are $P[X(k) = S]$ for each S genotype
- k is unknown but large enough
- the probability of the edge (S, T) on the chain with self-lops is

$$P(S, T) = P[X(k) = T | X(k-1) = S]$$

  does not depend on k
- we are interested in the probability of the edge (S, T) on the chain without self-loops

$$JP(S, T) = \frac{P(S, T)}{norm(S)}$$

Back to Presentation

Technical Details

- by acyclicity and anti-transitivity of CPMCs

$$P(S,T) = \frac{\sum_{i=1}^{k} P[f_i(T)] * P[X(i-1) = S | f_i(T)]}{\sum_{i=0}^{k-1} P[X(i) = S]}$$

- which without self-loops become

$$JP(S,T) = \frac{\sum_{i=1}^{k} P[f_i(T)] * P[X(i-1) = S | f_i(T)]}{norm(S)}$$

- in the case without convergencies $P[X(i-1) = S | f_i(T)] = 1$ and

$$\sum_{i=1}^{k} P[f_i(T)] = P[X(\leq k) = T] = P[X(k) = T] + \sum_{Y \in Adj^-[T]} P[X(\leq k) = Y]$$

Back to Presentation

Technical Details

- by acyclicity and anti-transitivity of CPMCs

$$P(S,T) = \frac{\sum_{i=1}^{k} P[f_i(T)] * P[X(i-1) = S | f_i(T)]}{\sum_{i=0}^{k-1} P[X(i) = S]}$$

- which without self-loops become

$$JP(S,T) = \frac{\sum_{i=1}^{k} P[f_i(T)] * P[X(i-1) = S | f_i(T)]}{norm(S)}$$

- in the case with convergencies the heuristics extimate

$$P[X(i-1) = S | f_i(T)] \text{ and } P[X(\leq k) = T]$$

Back to Presentation