CIMICE

markov Chain Inference Method to Identify Cancer Evolution

Nicolò Rossi, supervised by Carla Piazza, Alberto Policriti and Bud Mishra
Università degli studi di Udine
rossi.nicolo@spes.uniud.it

UNIVERSITÀ
DEGLI STUDI
DI UDINE
hic sunt futura

Abstract

Cancer is one of the first causes of death in the world, especially in high-income countries. In fact, the World Health Organization has estimated that more than 60 people over 100000 die of cancer involving the respiratory system, making it the third most lethal disease in developed countries [3]. The strength of cancer is that of not being just a single disease but a class of them, where each tumor in each person and even in each tissue is different and evolves accordingly to rules that are still unclear. To cope with this behavior researchers in medicine and biology are gradually developing methods capable of providing **personalized treatments** to the patients. The sequencing techniques are currently in fast development and data about cancer is growing faster every day as the technologies are becoming more precise, reliable and cheaper. **Data analysis** has acquired a crucial role in this field and one of the most challenging topic in this area is the reconstruction of cancer progression from cancer data. This is not an easy task as genomic data is noisy, incomplete and, in almost every case, acquired at a single time in the tumor existence, forcing the reconstruction of temporal information from static data. After many years from the first seminal work of Desper et al. [4], there is still not an accepted method that is actually used in medicine, but the efforts on the topic are growing and this research field got a name of its own: **tumor phylogenetics**. The number of tools available to this end is so high that many reviews are now focusing on understanding and comparing their many complex approaches [5, 1]. Our goal is to prepare another tool for cancer progression inference but focusing on keeping it **simple**, **fast** and **modern** by making very few and reasonable assumptions, managing the analysis even on the largest datasets available, and using single cell sequencing data. Then we will proceed by comparing it with some more complex state-of-the-art tools, in particular the ones from the TRONCO pipeline (CAPRI, CAPRESE, TRaIT) [2], using both synthetic and real data.

Tumor phylogenetics

Tumor phylogenetics is the research field that aims to reconstruct the sequence of mutations that lead to the formation of a tumor from experimental data. The name phylogenetics refers to the phylogeny of species, an older area of study that had already developed many methods to trace the evolutionary history of organisms. There is a wide agreement on considering **cancer progression** evolutionarily when the tumor undergoes a selective pressure, like that of a treatment, while in other circumstances this evolution is less precisely guided. It is in fact more probable that, without selection, the tumor starts acquiring mutations in a random fashion leading to the formation of many **subclones** with different genotypes and survival fitness to treatments [5]. From the cancer point of view this is quite an advantage: the survivability of its cells to external attacks is increased, cell specialization is supported, and functionality acquisition speeds is improved. Knowing in advance what mutations a tumor will acquire would be an huge advantage to prepare better treatments to patients, maximizing success probabilities and minimizing treatment related risks. There are three main kind of experiments for cancer data taking, as it can be seen in figure 1. With CIMICE we focus on **single cell** data, the one with best resolution.

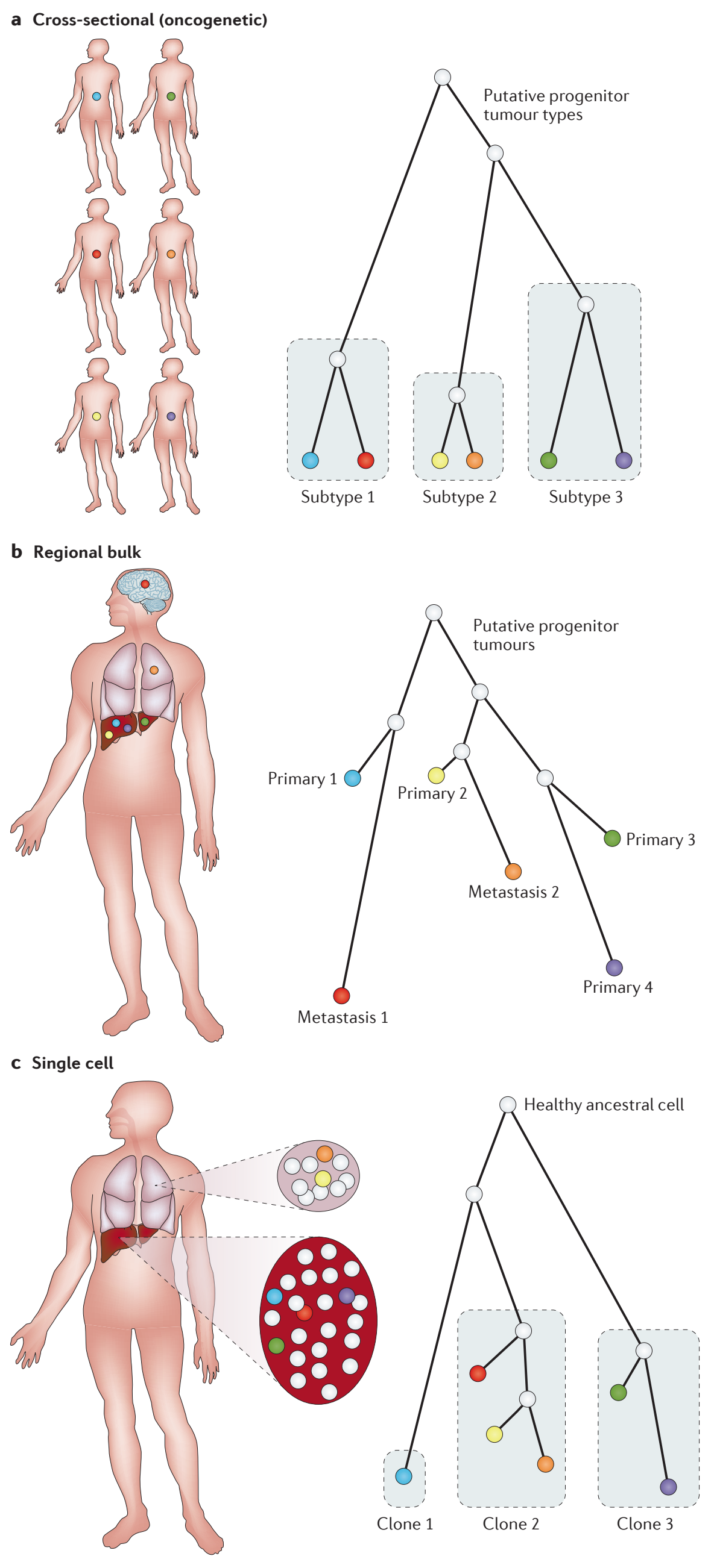

Figure 1: Visual representation of the three kind of experiments and a possible phylogenetic tree of the analyzed samples [5] (license number 4425280760601)

Discrete time Markov chains

A **discrete time Markov chain** (or DTMC) is a Markov process, so a random process with the Markov Property, with finite and discrete states and discrete transitions among them. The **Markov Property** says that the transitions to other states are related only on the current state, so that there is no memory of the previous events that led to that particular state. Typically this is formalized as follows:

$$P(X(t_n) = s_n | X_{n-1}(t_{n-1}) = s_{n-1}, \ldots, X_1(t_1) = s_1) = P(X(t_n) = s_n | X_{n-1}(t_{n-1}) = s_{n-1})$$

Where $s_1, \ldots, s_n$ are states, $X(t_1), \ldots, X(t_n)$ are the states reached after the $t_1, \ldots, t_n$ discrete transitions respectively.

A DTMC can be fully defined by a discrete and finite set of states Q, a transition matrix M with size $|Q| \times |Q|$ where each cell $\langle a, b \rangle$ describes $P(X(t_n) = b | X(t_{n-1} = a))$ and q_I a vector of $|Q|$ elements indicating for each state its probability of being the initial state. Markov chains are also the simplest Markov models and they are used with the assumption of the system being fully observable and autonomous.

Cancer model

We consider cancer as an **evolutionary process** that involves multiple entities, the cells. In our model each cell has its own evolution that is represented by the temporal sequence of all the different genotypes it had during its entire existence, starting from the clonal, or healthy, genotype. We consider the evolutionary process continuous and irreversible for each cell and cancer is seen as the collection of all the evolutionary trajectories of the cells it is made of. Considering the evolution of a single cell, according to our model, if a cell has a certain genotype at a certain moment in its life, after a certain time interval the cell can be either in the same stage as before or in a further one, if it evolved to another genotype acquiring new mutations. If we consider a time interval (Δt) small enough to make negligible the probability of having more than one mutational event occurring in it, the number of possible stages after that time is reduced to two: the current state and the next stage of this cell evolution. These stages are mutually exclusive and have a specific probability of occurrence. By these considerations our cancer model can be interpreted as a Markov chain, an example can be seen in figure 2. Note that each node of the graph is associated with a genotype (a cancer state) and that we can use this graph as an **artificial dataset generator** for our simulations.

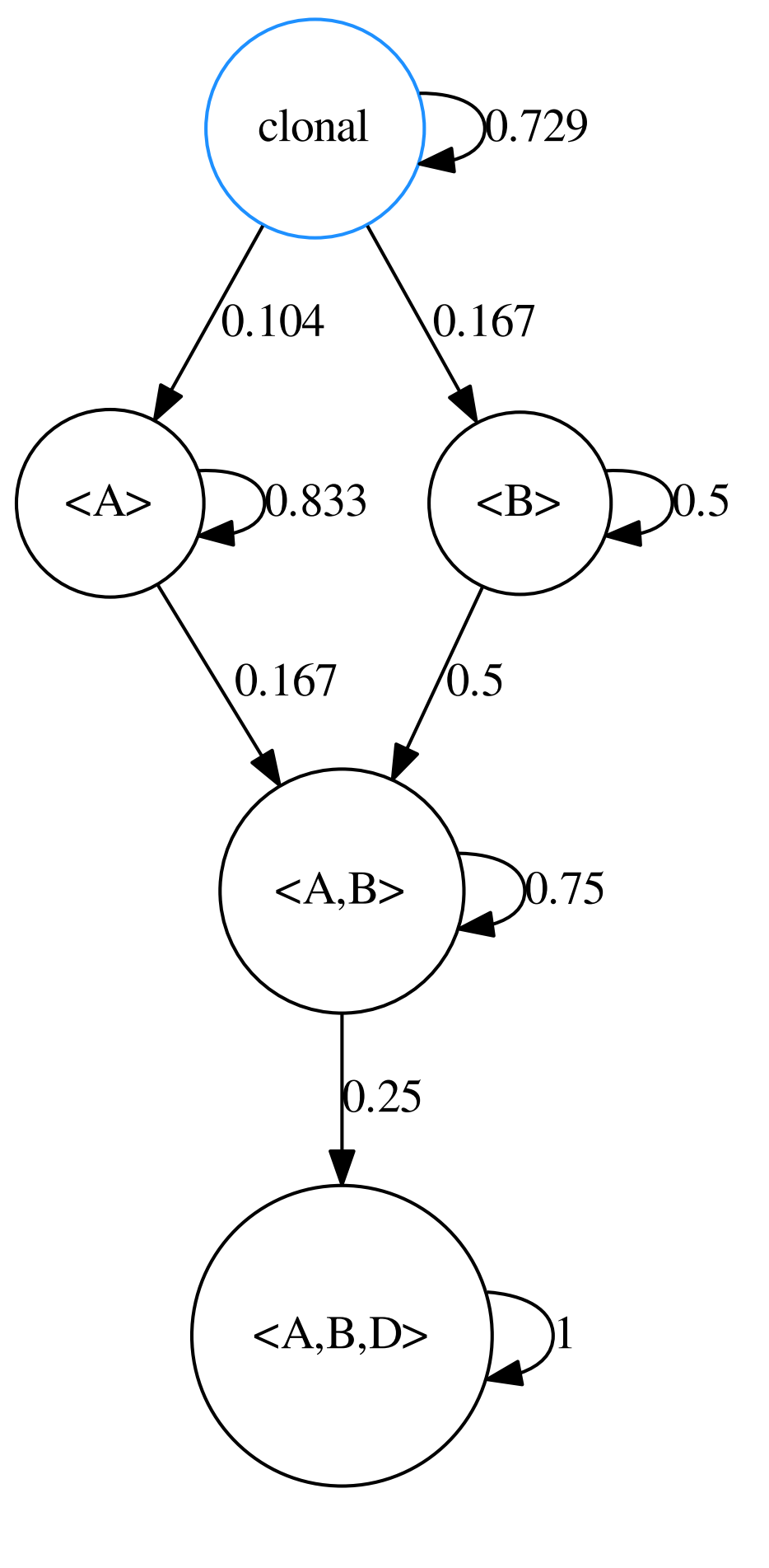

Figure 2: This graph shows the transition probability after Δt time of a really simplified cancer example. "Clonal" represents the healty genotype.

Model inference and theoretical proprieties

The model inference method is based on two phases: **structure reconstruction** and **weight estimation**. In the first we prepare an approximation of the possible evolutionary sequences of the different **genotypes** found in the dataset. The dataset is in fact a boolean matrix in which each position shows if a given gene was found mutated in a given sample; each row of this matrix is a genotype. The edges on the graph are computed by placing an edge $\langle a, b \rangle$ if the set of mutations of genotype b is a superset of that of a and there is not another genotype c that is also a superset of a but has less mutations. The second phase is divided in four steps: **upWeights computation**, **upWeights normalization**, **downWeights computation** and **downWeights normalization.** In figures 3, 4, 5 and 6 we give a full example of the reconstruction made by CIMICE, each node in these graphs is labelled with a unique genotype and its **observed probability**. Note that the resulting Markov chain has no self loop because the notion of time is changed and a transition is no longer triggered by the elapse the Δt interval, but by the acquisition of a mutation.

Dataset:

s\g	A	B	D
S1	1	0	0
S2	1	0	0
S3	1	0	0
S4	0	1	0
S5	0	1	0
S6	0	1	0
S7	0	1	0
S8	1	1	0
S9	1	1	0
S10	1	1	1

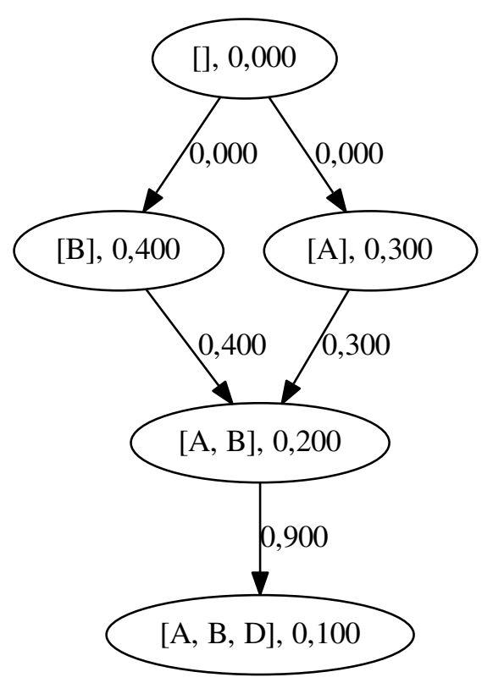

Figure 3: (1) upWeights computation

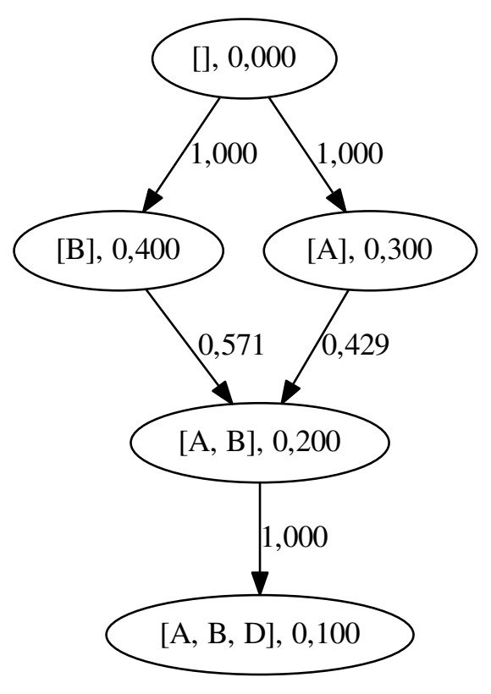

Figure 5: (3) upWeights normalization

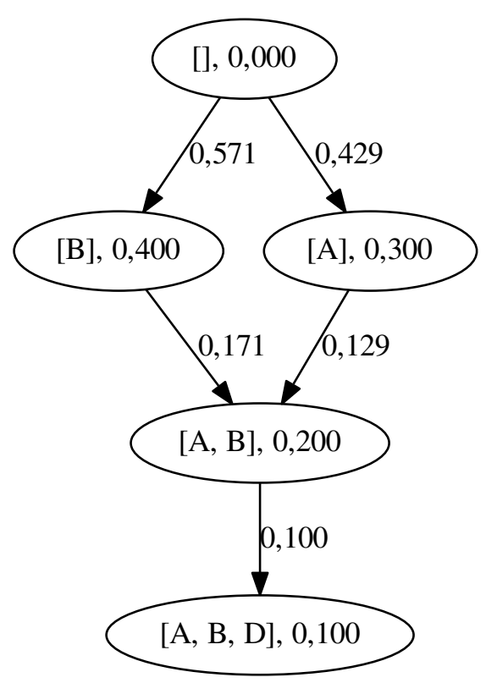

Figure 4: (2): downWeights extraction

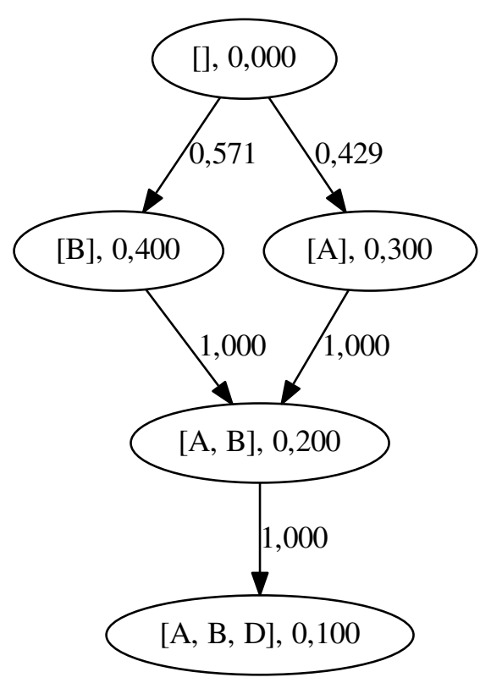

Figure 6: (4) downWeights normalization

We proved mathematically that our method can reconstruct the exact model that generated the dataset seen with this notion of time in the case of tree structured evolution, while, in the more generic case of DAG structured evolution, we proved that we reconstruct one of the models that could have generated the dataset. In fact, in the last case, the inferred model is dependent on the heuristic used for disentangling confluences, that we call **splitting function**. We use the normalized upWeights for this end under the assumption that the most probable history for a genotype is the most observed one, but the choice of the splitting function is actually a parameter of CIMICE.

Conclusion and future plans

In this work we presented CIMICE, a simple, efficient and novel method for cancer progression inference based on Markov chains. The method exploits few commonly used assumption. We proved its correctness with respect to our cancer model and we evaluated its performance on simulated and biological data with promising results (not shown in this poster). It would also be very interesting to enrich this method considering the peculiarities of different mutational types like CNVs and chromosome scale variations that could help in finding the temporal order of events. We could develop many different splitting functions for our tool and compare them to check their performances in order to understand how valid our heuristic is. It would also be quite important to find larger and higher resolution datasets that could help in improving our biological validation. There are many more ways to improve CIMICE, but this was exactly our goal, to create a **simple core** on which gradually grow a more advanced cancer inference method.

References

[1] Niko Beerenwinkel, Chris D Greenman, and Jens Lagergren. Computational cancer biology: an evolutionary perspective. PLoS computational biology, 12(2):e1004717, 2016.

[2] Luca De Sano, Giulio Caravagna, Daniele Ramazzotti, Alex Graudenzi, Giancarlo Mauri, Bud Mishra, and Marco Antoniotti. Tronco: an r package for the inference of cancer progression models from heterogeneous genomic data. Bioinformatics, 32(12):1911–1913, 2016.

[3] WTO Department of Information, Evidenceand Research. Who methods and data sources for country-level causes of death 2000-2016. March 2018.

[4] Richard Desper, Feng Jiang, Olli-P Kallioniemi, Holger Moch, Christos H Papadimitriou, and Alejandro A Schäffer. Inferring tree models for oncogenesis from comparative genome hybridization data. Journal of computational biology, 6(1):37–51, 1999.

[5] Russell Schwartz and Alejandro A Schäffer. The evolution of tumour phylogenetics: principles and practice. Nature Reviews Genetics, 18(4):213, 2017.

Implementation

CIMICE is implemented in Java and its source code is available at: `https://github.com/redsnic/tumorEvolutionWithMarkovChains`.