MOCATHE

MOdel for CAncer THErapies

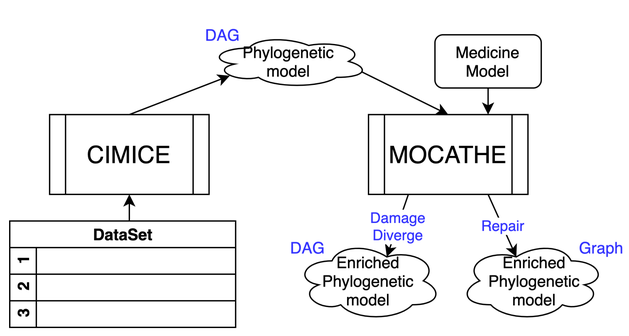
Cristian Stocco, supervised by Carla Piazza
Università degli studi di Udine
cristian.stocco001@spes.uniud.it

UNIVERSITÀ DEGLI STUDI DI UDINE
hic sunt futura

Abstract

We present the tool **MOCATHE** (**MO**del for **CA**ncer **THE**rapies) which is an extension of **CIMICE**, a software which infers a phylogenetic model of cancer progression. **CIMICE** is a tool based on **Discrete-Time Markov Chain** (DTMC), including estimators of cancer regression when treatments (e.g., drugs) are prescribed. The integration is made with a **Medicine** model in order to predict how cancer reacts to the treatment.

Tumors are one of the principal causes of death in modern society and understanding how a cancer progresses within a patient could be helpful in the development of therapy strategies. Because of the complexity of tumors, many progression models have been proposed in the literature exploiting different mathematical and logical frameworks.



This is an interdisciplinary work which involves **Biology**, and in particular tumor phylogenetics and evolutions, **Mathematics**, which provides formal models for describing progressions, **Computer Science**, with algorithms for the inference of models from data.

We consider discrete models based on **DTMC** for modelling progressions and **Propositional Logic** formulae for describing treatments. Some properties of our models are formally analysed and proved and a Java implementation of the framework is provided.

Molecular Biology & Genetics

Biology is a natural science which examines life and living organisms. The **molecular biology** defines molecular basis of biological activities and **genetics** which concerns with the study on **genes** and their role in **inheritance**.

Our interests of **molecular biology** are focused on macromolecules such as DNA, RNA, with their mutations between them, changes which involves genes to produce different types of **genotypes**. There are two main factors in which a change occurs: 1) **mutation**, that is a process for transferring informations altered by radiations or chemicals; 2) **evolution**, simply variation over time for the organism's change.

Our data are arranged in **Mutational Arrays** D, matrices with dimension $m \times n$ whose m is the number of samples and n is the number of genes. Therefore for each row, the representation of an **active gene** for the current **genotype** is expressed by an 1, intuitively 0 represents the genotype not containing that particular gene.

The input dataset D is a matrix $M^{m \times n}$:

$$D = \{gnt_i \mid i \in [m], n = |gnt_i|\}$$

Discrete-Time Markov Chain

A **Discrete-Time Markov Chain** (or DTMC) is a Markov process, therefore a random process based on the Markov Property, with finite and discrete states and discrete transitions among them. It could also be summarized as a graph $G = \langle V, E \rangle$ with probability distribution p over weights.

$$p : V \times V \to [0, 1]$$
$$(\forall u \in V)\left(\sum_{v \in V} p(\langle u, v \rangle) = 1\right)$$

Moreover, the **Markov Property** enunciates that transitions amid stated are triggered only by the current state, giving no memory from previous triggered events while choosing the next state.

$$(i) \quad \mathbb{P}(X_0 = i_1) = \lambda_{i_1}$$
$$(ii) \quad (X_{n+1} = i_{n+1} | X_0 = i_0, \ldots, X_n = i_n) = p_{i_n i_{n+1}}$$

Where $\lambda \in \mathbb{R}^k_{\geq 0}$ defines the **initial distribution**, with $k = |V|$.

Cancer Progression Markov Chain

The **Cancer Progression Markov Chain** is an enhancement of the **DTMC** to evaluate and to infer the progression of a cancer. Therefore, some properties given from that structure type are extended in this chapter.

Given a set of genes $G = \{g_1, \ldots, g_n\}$. The set of genotypes $S = \{S_0, \ldots, S_m\}$ is a set of S_i which is a subset of genes, where S_0 defines the "clonal cell", substantially $S_0 = \emptyset$.

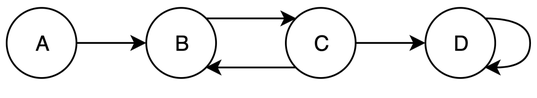
The **Cancer Progression Markov Chain** is a pair $C = (S, p)$ in which

- $S = \{S_1, \ldots, S_m\}$ is a finite set of genotypes over a set of genes $G = \{g_1, \ldots, g_n\}$
- for each $i, j \in [m]$ $i \neq j \to p(S_i, S_j) > 0$ if and only if $S_i \subseteq S_j$ and for each $k \neq i, j$ such that $(S_i \subseteq S_k \subseteq S_j) \to p(S_i, S_j) = 0$

The **initial distribution** λ is considered implicitly as the starting state, so the clonal cell $\lambda_{i_1} = 1$ and for all other vertices, its probability is 0: $\forall j \in [2, \ldots, |S|]$ $\lambda_{i_j} = 0$.

Example

This is an example of a simple Markov Chain.



Models

The disease context is managed by already-implemented **CIMICE** model based on **CPMC**. The weights on transaction edges are describing the probability of cell mutation, without mentioning times because it is difficult to predict as state-of-art.

Given a set of genotype $S = \{S_1, \ldots, S_m\}$ the observed genotypes is a probability function for the genotypes S in the dataset D

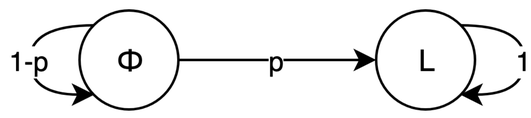
$$P_D(gnt) = \frac{\#(gnt, D)}{|D|}$$

Where the **Cancer Model** is a graph $A = \langle V_A, E_A, W_A \rangle$ defined as following

$$V_A = \{\langle gnt, P_D(gnt) \rangle \mid gnt \in D \wedge P_D(gnt) > 0\} \cup \{\langle \emptyset, P_D(\emptyset) \rangle\}$$
$$E_A = \{\langle u, v \rangle \mid u, v \in V_A \wedge u[0] \subset v[0] \wedge |v[0]| - 1 = |u[0]|\}$$
$$W_A = \text{see Thesis for more details}$$

The **Medicine Model** represents a logic formula, its probability of application, and the type of the treatment of the DNA: DNA Damage and DNA Diverge, both with suffix D and DNA Repair with suffix R. The **Medicine Model** is a graph $M_L = \langle V_L, E_L, W_L \rangle$ with the following attributes

- $\varphi = \phi_1 \vee \cdots \vee \phi_n$: logic formula of the medicine M_L
- p: precision of the medicine M_L
- suffix $L \in \{R, D\}$: label of the treatment type



The **Therapy Model** faces the same types of the **Medicine Model** explained before and it is built up by the parallel composition of the previous two models. The **Therapy Model** is a graph $C_L = \langle V_L, E_L, W_L \rangle$ with the following attribute

- suffix $L \in \{R, D\}$: label of the treatment type

defined as following

$$V_L = \begin{cases} V_A \cup V_{exp} & \text{if } L = R \\ V_A \cup \{F\} & \text{otherwise} \end{cases}$$

$$E_L = E_A \cup E_{AM} \qquad E_{AM} = \begin{cases} \{e = \langle u, v \rangle \mid u, v \in V_L \wedge v = u \setminus \text{positiveLiteral}(\varphi_u)\} & \text{if } L = R \\ \{e = \langle u, F \rangle \mid u \in V_L \wedge literal(\varphi_u) \neq \emptyset\} & \text{otherwise} \end{cases}$$

$$W_L(e = \langle u, v \rangle) = \begin{cases} p & \text{if } e \in E_{AM} \\ W_A(e) * (1 - p) & \text{if } e \in E_A \wedge literal(\varphi_u) \neq \emptyset \\ W_A(e) & \text{if } e \in E_A \wedge literal(\varphi_u) = \emptyset \end{cases}$$

Theoretical proprieties

We further demonstrate some theorems in order to maintain the properties of **MOCATHE** inherited by **CIMICE** or to confute them, also some theorems show some properties belonging both of them.

- after parallel composition, when C_D is a Therapy Model the **DAG property is kept**
- after parallel composition, when C_R is a Therapy Model the **DAG property is dropped**
- after the application of balancing function, the **probability distribution** is kept
- properties of **Markov Chains** such as transient, recurrent, absorbing state, and comm. classes
- dimensional bounds of **MOCATHE** based on number of genes and vertices expansion
- diameter of **CIMICE** and **MOCATHE**

Implementation

The development of the **MOCATHE** software is made in Java programming language such as **CIMICE** does. We added a simple example to understand how to use the command line and the parameters in input, calculating also the complexity of the algorithm based on the dimension of V and E, given $G = \langle V, E, W \rangle$. A simple improvement calculating the $G^{-1} = \langle V, E^{-1}, W \rangle$ is proposed.

Conclusion

In this work we presented **MOCATHE**, a tool which extends **CIMICE**, adding functionalities for representation of treatment effects in cancer progression.

CIMICE builds inference models of cancer phylogenetics based on **DTMC** and our extension **MOCATHE** allows the user to analyze the effect of different type of treatments. In **MOCATHE**, therapies are represented as the **Propositional logic** formulae which states the combination of mutation on which a therapy has effect, with a probability of success. Together, the computations provides a **DTMC** representing the evolution of the tumor when the therapy is prescribed.

Finally, we propose some further extensions such as the estimation of time instants, the parallel composition of different drugs, the strength of therapy by Markov Chain properties, and the last but not the least trasformations from a normal logical formula to a Disjoint Normal Form.

References

[1] Navnath S.Gavande et al. DNA repair targeted therapy: The past or future of cancer treatment? Pharmacology & Therapeutics, 2016.

[2] Nicolò Rossi CIMICE: Markov Chain Inference Method to Identify Cancer Evolution. Bioinformatics, 2018.

[3] Computational Biology and Bioinformatics. CIMICE: Markov Chain Inference Method to Identify Cancer Evolution Bioinformatics, 2019.