Modeling dynamics of circulating tumor DNA for detecting resistance to targeted therapies

A phylogenetic approach

Ulysse Herbach

École MMB – Aussois

20 mai 2019

Context

Context

- **Chemotherapies** are toxic even to non-tumor cells
- **Targeted therapies** can focus on a specific mutation
- Unfortunately, emergence of **resistance** is very common

Traditional biopsies

Sequencing DNA of tumor cells from **extracted tissues**:

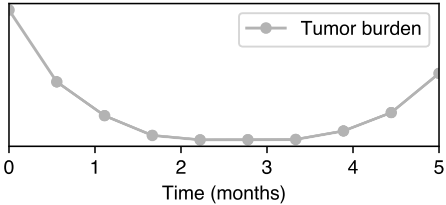
- Cannot be performed very often
- Only reflects some part of the tumor

Liquid biopsies

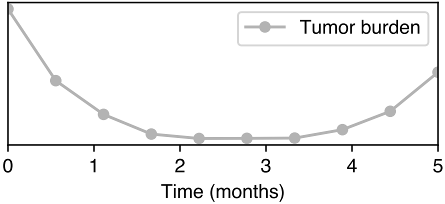
Sequencing plasma cell-free DNA (cfDNA) from **blood samples**:

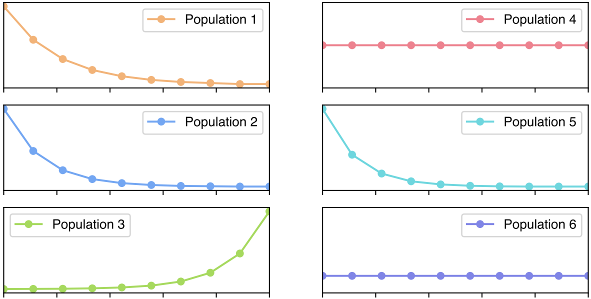
- Can be performed much more often
- Potentially reflects the full heterogeneity of the tumor

Emergence of resistance

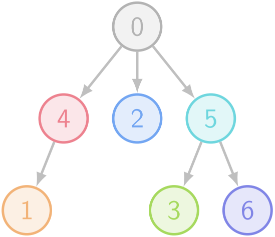


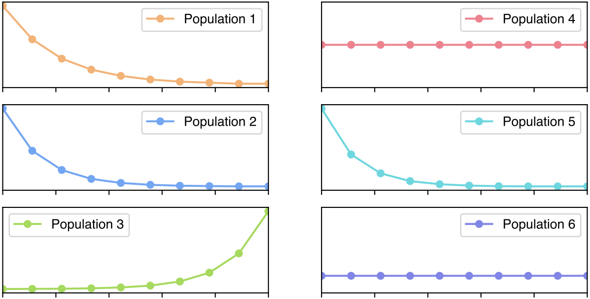
Emergence of resistance





Emergence of resistance

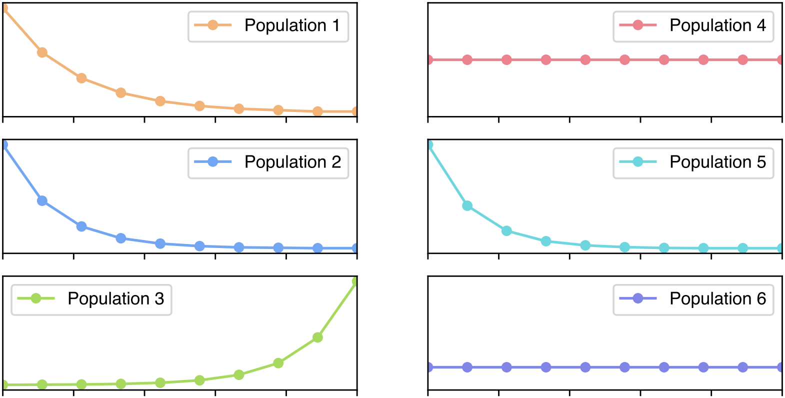




The lost genome

Liquid biopsy data

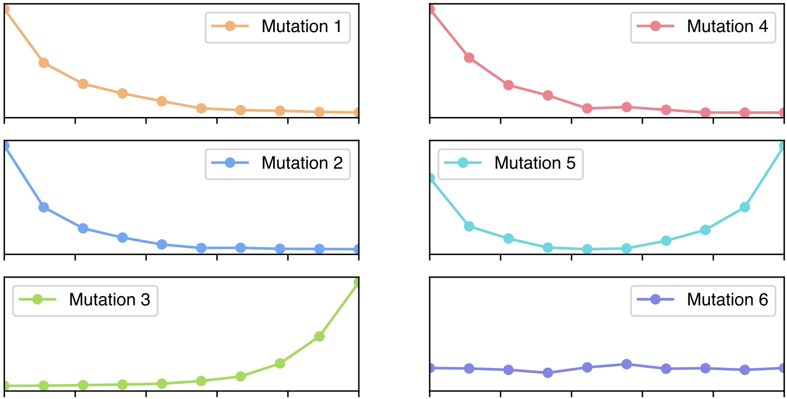
Very short DNA fragments (≈ 150 bp) aligned with a reference sequence: each genome is *lost* and must be *inferred* from a model.



The lost genome

Liquid biopsy data

Very short DNA fragments (≈ 150 bp) aligned with a reference sequence: each genome is *lost* and must be *inferred* from a model.



“Poor man’s phylogenetics”

Traditional phylogenies

Clustering **cells** by **mutational composition**

New wave phylogenies

Clustering **mutations** by **cellular frequencies**

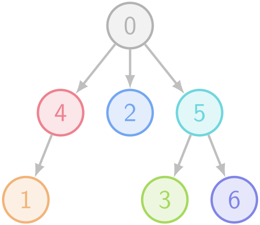
Objectives

- Directly reconstruct a phylogenetic tree
- Exploit the time structure using a dynamical model
- Detect and characterize resistant populations

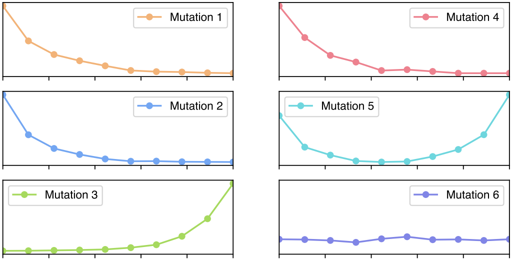
Hypotheses

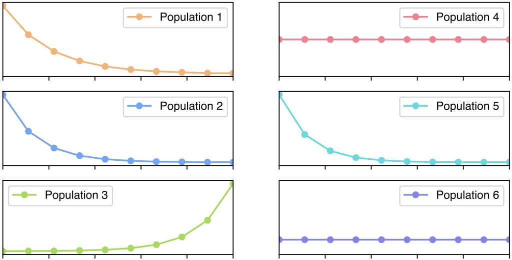
1. All observed mutations have already appeared at $t = 0$
2. A mutation only appears once and then never disappear
3. All cfDNA fragments are degraded faster than cell death

Recap



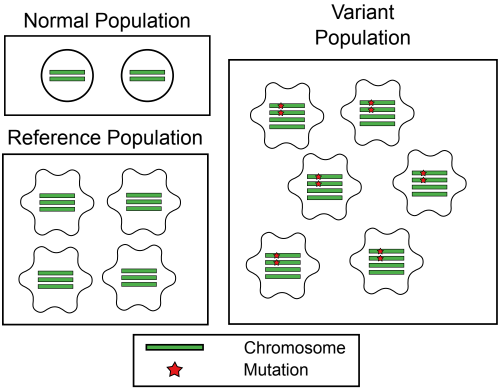
$$\mathcal{P}_1 = \{1, 4\}$$
$$\mathcal{P}_2 = \{2\}$$
$$\mathcal{P}_3 = \{3, 5\}$$
$$\mathcal{P}_4 = \{4\}$$
$$\mathcal{P}_5 = \{5\}$$
$$\mathcal{P}_6 = \{5, 6\}$$





Getting even worse

Cellular frequencies are latent variables:



Roth *et al.* (2014). PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398.

Basic model for cfDNA dynamics

Mutations $\{1, \ldots, m\}$ and corresponding populations $\mathcal{P}_1, \ldots, \mathcal{P}_m$.

- The size $C_i(t)$ of $\mathcal{P}_i$ at time t is described by

$$\dot{C}_i(t) = \lambda_i C_i(t) - \mu_i C_i(t)$$

- The amount of mutation j circulating in the blood is given by

$$\dot{M}_j(t) = \sum_{i=1}^{m} [a_{ij} \mu_i C_i(t)] \; - \; d M_j(t)$$

where $a_{ij} > 0$ if and only if $j \in \mathcal{P}_i$, that is, $\mathcal{P}_i$ is a *subclone* of $\mathcal{P}_j$.

Quasi-steady-state approximation for $d \ll \mu_i$

$$M_j(t) = \frac{1}{d} \sum_{i=1}^{m} [a_{ij} \mu_i C_i(t)] = a_{1j} c_1 e^{b_1 t} + \cdots + a_{mj} c_m e^{b_m t}$$

Statistical model

Parameters:

- $a = (a_{ij}) \in \mathbb{N}^{m \times m}$ scaled tree-structure matrix
- $b = (b_i) \in \mathbb{R}^m$ birth rate of each population
- $c = (c_i) \in (\mathbb{R}_+^*)^m$ initial size of each population

Random variables for $t = (t_k) \in \mathbb{R}^N$ observation times:

- $Y = (Y_{ik})$ hidden: size of population i at time t_k
- $X = (X_{jk})$ observed: amount of mutation j at time t_k

$$\mathcal{L}(Y) = \bigotimes_{i,k} \mathrm{Gamma}(c_i \exp(b_i t_k), 1)$$

$$\mathcal{L}(X|Y) = \bigotimes_{j,k} \mathrm{Poisson}(a_{1j} Y_{1k} + \cdots + a_{mj} Y_{mk})$$

Inference

- We need to infer $\theta = (b, c)$ **and** the nonzero structure of a
- By Cayley's formula there are $(m+1)^{m-1}$ possible models...

Variational trick

Idea:

- each a corresponds to a **rooted tree** $z \in \mathcal{T}_0$ on $\{0, 1, \ldots, m\}$
- we can **regularize** the tree structure by making z random

For **any** distribution $q(z) > 0$, Jensen inequality gives:

$$\begin{aligned}
\log p_\theta(x) = \log \left[\sum_z p_\theta(x, z)\right] &= \log \left[\sum_z \frac{p_\theta(x, z)}{q(z)} q(z)\right] \\
&\geqslant \sum_z \log \left[\frac{p_\theta(x, z)}{q(z)}\right] q(z) \\
&= \sum_z [\log p_\theta(x, z) - \log q(z)] q(z)
\end{aligned}$$

Remarks

- [left] $-$ [right] $= \mathrm{KL}[q(\cdot) \| p_\theta(\cdot|x)] \geqslant \frac{1}{2} \| q(\cdot) - p_\theta(\cdot|x) \|_1^2$
- Hence the **optimal** choice is $q(z) = p_\theta(z|x)$
- This choice is nothing more than the **EM algorithm**

Variational distribution over trees

An interesting distribution for $z \in \mathcal{T}_0$ is given by

$$q_\alpha(z) = \frac{1}{C(\alpha)} \prod_{(i,j)\in z} \alpha_{ij} \quad \text{where} \quad C(\alpha) = \sum_{z' \in \mathcal{T}_0} \prod_{(i,j)\in z'} \alpha_{ij}$$

Matrix-tree theorem

Let $n \in \mathbb{N}$ and $A = (A_{ij})$ be the $n \times n$ matrix defined by

$$A_{ij} = \begin{cases} -\alpha_{ij} & \text{if } i \neq j \\ \sum_{k=1}^{n} \alpha_{kj} & \text{if } i = j \end{cases}$$

Then for any $v \in \{1, \ldots, n\}$ we have

$$\det(A_{\{v\}}) = \sum_{T \in \mathcal{T}_v} \prod_{(i,j)\in T} \alpha_{ij}$$

where $\mathcal{T}_v$ is the set of all trees on $\{1, \ldots, n\}$ that are rooted at v.

Thanks!

The ITMO project:

Maths

- Nicolas Champagnat
- Anne Gégout-Petit
- Pierre Vallois
- Coralie Fritsch
- Aurélie Muller-Gueudin
- Aline Kurtzmann

Medicine

- Alexandre Harlé
- Jean-Louis Merlin
- Erwan Pencreac'h