SAPIENZA
UNIVERSITÀ DI ROMA

TODO Titolo

Faculty of Information Engineering, Computer Science and Statistics
Bachelor's Degree in Computer Science

Alessio Bandiera
ID number 1985878

Advisor
Prof. Ivano Salvo

Academic Year 2023/2024

TODO Titolo
Bachelor's Thesis. Sapienza University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: alessio.bandiera02@gmail.com

TODO. test

Contents

Chapter 1

Introduction

1.1 Context

1.1.1 Cancer

Cancer is a medical condition characterized by uncontrolled cell proliferation, which allows cells to infiltrate into organs and tissues, thereby altering their functions and structure. This exponential growth is driven by mutations in cellular DNA, which encodes the instructions for cell development and multiplication, therefore errors in these instructions can lead to cancerous transformation. In most types of cancer, a single aberration is insufficient for cancer development; instead, multiple mutations are required. Some of these mutations are present since birth, while others occur throughout life due to chance or lifestyle choices. Additionally, for tumor proliferation to occur, mutations in genes that regulate cell growth are necessary [13]. Specifically, proto-oncogenes, which promote mitosis, and tumor suppressor genes, which inhibit cell growth, are involved in this process, known as *oncoevolution* [4].

talk about oncogenes in depth?

1.1.2 Cancer treatment

Research aimed at finding a cure for cancer is continuously evolving due to the tumor's lethality and complexity. Currently, the primary techniques used to remove, control, manage, and delay the effects of cancer include [2]:

- **surgery**, which involves the removal of the cancerous region and is generally reserved for solid tumors;
- **radiotherapy**, which uses x-rays to destroy tumor cells, aiming to target the cancerous region as precisely as possible to preserve healthy tissue; however, radiotherapy can increase the risk of developing secondary tumors, such as leukemia or sarcomas, and may lead to delayed effects like dementia, amnesia, or progressive cognitive difficulties;
- **chemotherapy**, which employs cytotoxic drugs to block cellular division in

both cancerous and healthy cells, but they can also induce side effects in rapidly renewing tissues.

- **hormone therapy**, which alters the balance of specific hormones, potentially leading to side effects such as joint pain or osteoporosis;
- **targeted therapy**, which involves drugs containing antibodies or inhibitory substances that specifically target cancer cells, promoting their destruction by the immune system; however, developing effective targeted therapies can be challenging due to the complexity of the target's structure or function; in addition, this approach may also induce unwanted side effects in various organs, and cancer cells may develop resistance if they find alternative ways to develop that do not rely on the therapy's target [12] .

check this out

In recent years, targeted therapy in particular has been the focus of extensive research due to its potential to precisely affect only the desired target, thereby reducing the side effects that currently characterize most cancer treatments and potentially limiting damage to healthy cells [11].

expand target therapy on how it works? if yes, make subsection

Chapter 2

Driver mutations

2.1 Mutations

2.1.1 Cell signaling and signaling pathways

Cell signaling is the process by which cells interact with each other, themselves, or their environment. This involves the transduction of signals, which can be chemical, or can involve other types such as pressure, temperature, or light signals [5]. **Pathways** are sequences of molecular interactions within a cell that lead to a change in the cell or the production of a specific product [10]. These pathways have a direction in which the actions occur, with the terms *upstream* and *downstream* indicating the initial and final stages of these processes, respectively.

In cancer research, **signaling pathways** are of particular interest because they mediate the transduction of cell signals. Identifying and targeting the signaling pathways responsible for cancer growth could potentially halt the development of the disease.

check this out, also check if what i wrote is actually true, i think i read it somewhere but can't find the source right now; expand on cell signaling? expand of pathways? if yes, make subsections

2.1.2 Passenger and driver mutations

There are two types of mutations in cancer: **passenger mutations** and **driver mutations**. Passenger mutations do not confer direct benefits to tumor growth or development, whereas driver mutations actively contribute to cancer progression by providing an evolutionary advantage and promoting the proliferation of tumor cells. A **driver gene** is a gene that harbors at least one driver mutation, though it may also contain passenger mutations . A driver pathway consists of at least one driver gene. Driver mutations, genes, and pathways are of significant scientific interest due to their crucial role in cancer proliferation.

DO I ADD THIS AS A CITATION???

Driver genes can be classified into 12 signaling pathways, which regulate cellular functions related to survival, fate, and genomic maintenance.

use (and expand) this? same source as prev

2.2 Classifying mutations

2.2.1 Frequency

To classify mutations into the two categories described, assessing their biological function is essential, though this remains a challenging task. Numerous methods exist to predict the functional impact of mutations based on *a priori* knowledge. However, these approaches often fail to integrate information effectively across various mutation types and are limited by their reliance on known proteins, rendering them less effective for less-studied ones [9].

With the decreasing cost of DNA sequencing, it is now possible to categorize mutations by examining their frequency, as driver mutations are typically the most recurrent in patients' genomes [9]. In fact, key driver events, such as TP53 loss-of-function mutations, can be identified by their significantly high frequency of occurrence across a set of tumors [1]. However, in many cases, since driver mutations are predominantly located in genes that are part of cell signaling pathways, different patients may harbor mutations in different pathway loci. Indeed, driver mutations can vary extensively between patient samples, even within the same cancer type [9]; additionally, there is minimal overlap of mutated genes across sample pairs, even from the same patient [14], reducing the statistical power of frequency analyses.

Moreover, multiple alternative driver alterations in different genes may lead to similar downstream effects. In such instances, the selective advantage is distributed among the alterations frequences of these genes. In current cancer genomics studies, where the number of samples is significantly smaller than the number of genes profiled per sample, frequency-based methods lack the statistical efficacy to distinguish passenger and driver mutations [1].

Therefore, studies should be conducted at the pathway level, as it is well established that different mutations can affect the same pathway across multiple samples [9]. However, since each pathway involves multiple genes, numerous possible combinations of driver mutations could impact a crucial cancer pathway, making it computationally unfeasible to test every possible gene permutation [6] — estimates suggest that the human genome contains more than 50,000 genes [8]. Hence, it is necessary to identify a property to leverage in order to conduct the research efficiently.

2.2.2 Mutual exclusivity and coverage

Most techniques developed in recent years for recognizing driver mutations leverage a statistical property observed in cancer patient data: each patient typically has a relatively small number of mutations that affect multiple pathways, thus each pathway will contain *1 driver mutation on average* per sample. This concept of mutual exclusivity among driver mutations within the same pathway, as statistically observed in patient samples, is then axiomatized and employed by research algorithms designed to identify driver mutations [9]. Additionally, mutual exclusivity *does not affect different pathways*; it is a phenomenon that occurs exclusively within a single pathway. While the precise explanation for this occurrence is not yet fully understood,

several hypotheses appear promising [7, 3, 6]:

- one hypothesis is that mutually exclusive genes are functionally connected within a common pathway, acting on the same downstream effectors and creating functional redundancy; consequently, they would share the same selective advantage, meaning that the alteration of one mutually exclusive gene would be sufficient to disrupt their shared pathway, thereby removing the selective pressure to alter the others; this explanation, however, does not fully account for the phenomenon because the co-alteration of mutually exclusive genes should not result in negative effects on the cell.

- an alternative explanation is that the co-occurrence of mutually exclusive alterations is detrimental to cancer survival, leading to the elimination of cells that harbor such co-occurrences; moreover, some pairs of mutually exclusive genes could be *synthetic lethal*, meaning that while the alteration of one gene may be compatible with cell survival, the simultaneous aberration of both genes would be lethal to the cell .

add example from survey paper?; also, use example? (mail "Risposte (parziali) alle questioni, ERG e SPOP")

In addition, another key property of driver pathways is **coverage**, i.e. driver genes constituting a driver pathway are frequently mutated across many samples.

Thus, *a driver pathway consists of genes that are mutated in numerous patients, with mutations being approximately mutually exclusive.* It is also observed that pathways exhibiting these characteristics are generally shorter and comprised of fewer genes on average [9].

2.2.3 *De novo* and *knowledge-based* approaches

Although the true explanation for mutual exclusivity remains unknown, and its therapeutic potential is still uncertain, this phenomenon is frequently observed in data and is thought to potentially lead to discoveries in cancer treatment. Existing approaches can be categorized into two types: *de novo* approaches, which identify mutually exclusive patterns using only genomic data from patients, and *knowledge-based* methods, which integrate the analysis with external *a priori* information [7]. De novo approaches might lack sufficient information as they do not utilize existing databases. Conversely, given that our understanding of gene and protein interactions in humans is still incomplete and many pathway databases fail to accurately represent the specific pathways and interactions present in cancer cells, *knowledge-based* approaches may be limited by their dependence on existing data sources. Consequently, *de novo* methods might yield new but potentially less accurate results, while *knowledge-based* approaches may limit the discovery of novel biological insights [9].

2.3 Mutual exclusivity formalization

2.3.1 Hard and soft mutual exclusivity

In the statistical literature, two types of mutual exclusivity are defined: **hard** and **soft**. Hard mutual exclusivity describes events that are presumed to be strictly mutually exclusive, with the null hypothesis being that any observed overlap is due to random errors. However, in this context it is not feasible to test for hard mutual exclusivity, as this is a property observed statistically from patient data. Therefore, it is necessary to relax the constraint to soft mutual exclusivity, where two otherwise independent events overlap less than expected by chance due to some statistical interaction [1].

2.3.2 Mutual exclusivity of a group

For a pair of genes, soft mutual exclusivity can be assessed using the Fisher's exact test. However, there is no agreed-upon method for analytically testing mutual exclusivity among more than two genes. One approach could involve checking whether each pair of genes within the group exhibits mutual exclusivity; this method, however, may be overly strict, as a gene set can exhibit a strong mutual exclusivity pattern as a whole even if no individual pairs show any [1].

Chapter 3

TODO

3.1 Multi-Dendrix

3.1.1 Dendrix

A very well known paper, which developed two algorithms called "Dendrix" [6], gave the following mathematical formalization to the properties of **mutual exclusivity** and **coverage**. Consider a so called "mutation matrix" A, with m rows and n columns, where each row represents a patient and each column represents a gene; the entry $a_{i,j}$ is equal to 1 if and only if gene j is mutated in patient i .

add table as example?

Given a gene g, let

$$\Gamma(g) = \{i : a_{i,g} = 1\} \tag{3.1}$$

denote the set of patients which have g mutated; futhermore, given a set of M genes, let the **coverage** be

$$\Gamma(M) = \bigcup_{g \in M} \Gamma(g) \tag{3.2}$$

which denotes the set of patients in which at least one of the genes in M is mutated. In accordance with the previous definitions of mutual exclusivity, we say that a set M of genes is **mutually exclusive** *if no patient has more than one mutated gene*, formally

$$\forall g, g' \in M \quad \Gamma(g) \cap \Gamma(g') = \varnothing \tag{3.3}$$

Any gene set can be thought as a $m \times k$ submatrix of a mutation matrix A, up to rearranging A's columns — their order does not matter since they represent genes. Accordingly, such a submatrix is said to be **mutually exclusive** *if each row contains at most one 1.*

Furhermore, given a gene set M, the following properties are formalized:

i) coverage: most patients have at least one mutation in M;

ii) approximate exclusivity: most patients have exactly one mutation in M.

To measure these two attributes, a measure what quantifies the trade-off between coverage and mutual exclusivity is introduced. Given a set M of genes, the *coverage overlap* is defined as follows:

$$\omega(M) = \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)| \tag{3.4}$$

Note that the sum in Eq. (3.4) is the number of 1s in M's corresponding submatrix; hence, the

Acknowledgements

TODO

Bibliography

[1] Özgün Babur, Mithat Gönen, Bülent Arman Aksoy, et al. "Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations". In: *Genome Biology* 16.1 (Feb. 2015). ISSN: 1474-760X. DOI: 10.1186/s13059-015-0612-6. URL: http://dx.doi.org/10.1186/s13059-015-0612-6.

[2] *Cancro: la cura*. URL: https://www.airc.it/cancro/affronta-la-malattia/guida-alle-terapie/cancro-la-cura.

[3] Jaroslaw Cisowski and Martin O. Bergo. "What makes oncogenes mutually exclusive?" In: *Small GTPases* 8.3 (July 2016), 187–192. ISSN: 2154-1256. DOI: 10.1080/21541248.2016.1212689. URL: http://dx.doi.org/10.1080/21541248.2016.1212689.

[4] Wikipedia contributors. *Carcinogenesis*. July 2024. URL: https://en.wikipedia.org/wiki/Carcinogenesis.

[5] Wikipedia contributors. *Cell signaling*. Aug. 2024. URL: https://en.wikipedia.org/wiki/Cell_signaling.

[6] Yulan Deng, Shangyi Luo, Chunyu Deng, et al. "Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability". In: *Briefings in Bioinformatics* 20.1 (Aug. 2017), 254–266. ISSN: 1477-4054. DOI: 10.1093/bib/bbx109. URL: http://dx.doi.org/10.1093/bib/bbx109.

[7] Yulan Deng, Shangyi Luo, Chunyu Deng, et al. "Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability". In: *Briefings in Bioinformatics* 20.1 (Aug. 2017), 254–266. ISSN: 1477-4054. DOI: 10.1093/bib/bbx109. URL: http://dx.doi.org/10.1093/bib/bbx109.

[8] Chris Fields, Mark D. Adams, Owen White, et al. "How many genes in the human genome?" In: *Nature Genetics* 7.3 (July 1994), 345–346. ISSN: 1546-1718. DOI: 10.1038/ng0794-345. URL: http://dx.doi.org/10.1038/ng0794-345.

[9] Mark D. M. Leiserson, Dima Blokh, Roded Sharan, et al. "Simultaneous Identification of Multiple Driver Pathways in Cancer". In: *PLoS Computational Biology* 9.5 (May 2013). Ed. by Niko Beerenwinkel, e1003054. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003054. URL: http://dx.doi.org/10.1371/journal.pcbi.1003054.

[10] Nhgri. *Biological Pathways Fact sheet*. Mar. 2019. URL: https://www.genome.gov/about-genomics/fact-sheets/Biological-Pathways-Fact-Sheet.

[11] Cleveland Clinic Medical Professional. *Targeted therapy*. May 2024. URL: https://my.clevelandclinic.org/health/treatments/22733-targeted-therapy.

[12] *Targeted therapy for cancer*. May 2022. URL: https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies.

[13] Bert Vogelstein and Kenneth W Kinzler. "Cancer genes and the pathways they control". In: *Nature Medicine* 10.8 (July 2004), 789–799. ISSN: 1546-170X. DOI: 10.1038/nm1087. URL: http://dx.doi.org/10.1038/nm1087.

[14] Junfei Zhao, Shihua Zhang, Ling-Yun Wu, et al. "Efficient methods for identifying mutated driver pathways in cancer". In: *Bioinformatics* 28.22 (Sept. 2012), 2940–2947. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts564. URL: http://dx.doi.org/10.1093/bioinformatics/bts564.