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.

Contents

Chapter 1

Introduction

placeholder.

abstract di tutta la tesi

1.1 Cancer

1.1.1 Overview

ho l'impressione che qui manca qualcosa, poco, ma qualcosa

Cancer is a group of diseases characterized by uncontrolled cell proliferation, which allows cells to infiltrate into organs and tissues, thereby altering their functions and structure. There are more than 100 types of cancer, which can be grouped into broader categories such as carcinomas, sarcomas, leukemias and others [27].

Estimates predicted more than 2 million new cases of cancer of any site in the United States in 2024, with over 600,000 deaths [19]. According to SEER [9] 22, cancer is most frequently diagnosed among individuals aged 65-74, who also represent the age group with the highest cancer-related mortality, accounting for approximately 30% of cases in both diagnoses and deaths [3].

Cancer is often preceded by a range of symptoms, some of which may be subtle or easily overlooked. Possible signs and symptoms include persistent coughing, changes in bowel habits, unexplained bleeding, lumps, unexplained weight loss, persistent pain, yellowing or itchy skin, and feeling tired or unwell without a clear reason [26].

The exponential growth of cancer 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. These genetic mutations can arise from several factors, including random chance or exposure to carcinogens [21]. These and other causes will be discussed in the following sections.

1.1.2 Causes

The development of cancer is a complex, *multistep process* influenced by various factors, making it too simplistic to attribute cancer to a single cause. While genetic mutations can occur randomly through errors in DNA during cell division, or be

inherited from a parent [21], many agents — such as radiation, chemicals, and viruses — have been found to induce cancer.

Radiation and many chemical carcinogens work by damaging DNA and *causing mutations*. These are known as **initiating agents** because they trigger genetic changes that lead to cancer. For example, solar ultraviolet radiation, chemicals in tobacco smoke, and *aflatoxin* are well-documented carcinogens. Tobacco smoke, in particular, is a major cause of lung cancer and is also linked to cancers of the oral cavity, throat, larynx, esophagus, and other areas. It is estimated that smoking contributes to a significant portion of all cancer deaths.

In contrast, some carcinogens — known as **tumor promoters** — facilitate cancer development by *stimulating cell proliferation* rather than by inducing mutations. Tumor formation in animal models typically require both an initiating agent and a promoter to facilitate the growth of mutated cells. For instance, hormones (especially estrogens) play a role as tumor promoters in certain cancers.

Additionally, some viruses are known to cause cancer in both animals and humans, such as those linked to liver cancer and cervical carcinoma. These viral-induced cancers highlight the broader impact of carcinogens and underscore their role in both viral and non-viral cancer development [7].

In summary, the various ways in which different factors contribute to cancer emphasize the complexity of the disease and underscore the importance of developing effective treatment approaches, which will be explored in later sections.

1.1.3   Mutations in cancer development

The fundamental feature of cancer development is **tumor clonality**, meaning tumors often develop from single cells that start to proliferate abnormally. However, the clonal origin of tumors does not mean that the initial progenitor cell had all the features of a cancer cell from the start. Instead, cancer evolves through a multistep process in which cells *gradually acquire malignant characteristics* through a series of **alterations**. This multistep nature is indicated by the fact that most cancers develop later in life. For example, the incidence of colon cancer increases markedly with age, showing a dramatic rise as individuals grow older. This steep age-related increase suggests that cancer typically results from **multiple abnormalities** accumulated over many years.

At the cellular level, cancer development is viewed as a process of mutation and selection for cells with progressively greater abilities to proliferate, survive, invade, and metastasize. The first stage, known as **tumor initiation**, involves a genetic alteration that triggers abnormal growth in a single cell, leading to the expansion of a population of clonally derived tumor cells. **Tumor progression**, continues as *additional mutations* arise within this cell population, with some mutations providing a selective advantage. As a result, cells bearing these advantageous mutations become *dominant* within the tumor, a process known as **clonal selection**. This selection continues throughout the tumor's evolution, causing it to grow more rapidly and become increasingly malignant [7].

Undoubtedly, mutations are fundamental to the development of cancer and to its progression. Therefore, to effectively combat this disease, it is essential to gain a comprehensive understanding of how these genetic alterations occur and contribute to tumor development.

1.2 Targeted therapy

1.2.1 Current cancer treatment

Research aimed at finding cancer treatment is continuously evolving due to the disease's lethality and complexity. Currently, the primary techniques used to remove, control, manage, and delay the effects of cancer include [4]:

- *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.

Recent advancements in traditional cancer treatments like chemotherapy, radiotherapy, and surgery have contributed to a decline in cancer mortality rates over the years. However, these methods still face significant limitations, often resulting in tumor recurrence and mortality, due to their various side effects. This has prompted a shift toward **mutation-targeted therapies**, as a result of their potential to precisely target cancer cells and minimize damage to healthy cells and tissue [20, 25].

1.2.2 Overview and origin

Targeted therapy is a form of cancer treatment that targets proteins responsible for the growth, division, and spread of cancer cells, and it forms the basis of precision medicine. The targets include growth factor receptors, signaling molecules, cell-cycle proteins, and other molecules crucial for normal tissue development and homeostasis, which often become overexpressed or altered in cancer cells, leading to their aberrant function [28].

Unlike standard chemotherapy, which indiscriminately destroys both rapidly dividing cancerous and normal cells, targeted therapies specifically attack abnormal proteins produced by mutated genes. Because normal cells lack these tumor-specific mutations, targeted therapies often show a higher degree of selectivity, causing fewer off-target effects and achieving more rapid and substantial tumor reduction [25].

The concept of targeted therapy originates from the german Nobel Prize Paul Ehrlich's idea of a *"magic bullet"* [10], when he envisioned a chemical capable of specifically targeting microorganisms. Over a century later, advances in molecular biology enhanced our understanding of the mechanisms behind cancer initiation, promotion, and progression. This progress led to the development of treatments that can interfere with specific molecular targets, typically proteins, linked to tumor growth and progression [28].

1.2.3 Therapy types

Most types of targeted therapy consist of **small-molecule drugs**, which are used for targets located inside cells because their small size allows them to enter cells easily, and **monoclonal antibodies**, which are laboratory-produced proteins engineered to bind to specific targets on cancer cells. Some monoclonal antibodies help the immune system identify and destroy cancer cells by marking them, while others directly inhibit the growth of cancer cells or induce their self-destruction, and still others deliver toxins directly to cancer cells [20].

Most targeted therapies treat cancer by interfering with specific proteins that promote tumor growth and spread. This approach differs from chemotherapy, which often kills all rapidly dividing cells. The following are the different approaches that targeted therapy employs [20].

- *Immunotherapy.* Cancer cells can often evade detection by the immune system. Certain targeted therapies mark cancer cells, making them easier for the immune system to identify and destroy, while others enhance the immune system's ability to fight cancer more effectively.

- *Signal interruption.* Targeted therapies can interrupt signals that cause cancer cells to grow and divide uncontrollably. Cells normally divide in response to specific signals binding to proteins on their surface. However, some cancer cells present changes in the proteins that tell them to divide without the signals. Targeted therapies can block these proteins, slowing the uncontrolled growth of cancer.

- *Angiogenesis inhibition.* The process through which new blood vessels form is called angiogenesis; beyond a certain size tumors need new blood vessels, thus the tumor sends signals to start angiogenesis. Some targeted therapies can disrupt the signals that trigger this process, preventing the formation of a blood supply, and restricting the tumor's size.

- *Cell-killing agents delivery.* Some monoclonal antibodies are combined with substances like toxins, chemotherapy drugs, or radiation. These antibodies

bind to targets on the surface of cancer cells, delivering the cell-killing agents directly into the cells, causing them to die. Most importantly, cells without these targets remain unharmed.

- *Apoptosis activation.* Cancer cells often evade the natural process of cell death, known as apoptosis, which initiates when cells become damaged or are no longer needed. Some targeted therapies can trigger apoptosis in cancer cells, leading to their death.

- *Hormone therapy.* Some types of breast and prostate cancer require specific hormones to grow. Hormone therapies block the body's production of growth hormones or prevent them from acting on cells, including cancer cells.

The diverse strategies employed by targeted therapies highlight the innovative approaches being developed to treat cancer more precisely. As research advances, these methods will continue to evolve, potentially improving outcomes and reducing side effects compared to traditional treatments.

1.2.4 Drawbacks and side effects

Like all cancer treatment, targeted therapy also has limitations, and often works best when combined with other types of targeted therapies or additional cancer treatments like chemotherapy and radiation [20].

In particular, developing drugs for certain targets can be challenging due to factors including the target's structural complexity, its function within the cell, or a combination of both. Moreover, cancer cells can develop resistance to targeted therapy, which may occur if the target itself mutates, rendering the therapy unable to interact with it effectively. Alternatively, resistance can arise if cancer cells adapt and find new growth mechanisms that do not rely on the target [20].

As for side effects, in general, targeted molecular therapies have good toxicity profiles. However, side effects differ from person to person, even among those undergoing the same cancer treatment [18], and some patients may be highly sensitive to these drugs and may develop specific and severe toxicities [28].

The most common side effects of targeted therapy are diarrhea and liver issues, but they may also include problems with blood clotting and wound healing, high blood pressure, fatigue, mouth sores, nail changes, loss of hair color, and skin problems. In rare cases, a perforation may occur in the wall of the esophagus, stomach, small intestine, colon, rectum, or gallbladder. Medications are available to manage many of these side effects, either by preventing them or treating them once they arise. Additionally, most side effects of targeted therapy subside after the treatment is completed [20].

In conclusion, although targeted therapy shows promise with generally manageable side effects, it has limitations such as potential drug resistance and varying individual responses. Effective management of these side effects and ongoing research are essential to improving treatment outcomes and patient care.

1.2.5 Drugs targeting mutations

As mentioned earlier, mutations play a crucial role in the growth and development of cancer. Targeted therapy allows for precise targeting of the mutations that enable cancer to continue its progression. In particular, oncogenic gene mutations may be druggable in several ways [25]:

- some oncogenic gene mutations encode proteins that are structurally or functionally different from the wild-type (WT), normal version of the protein; these differences create an opportunity for developing targeted therapies, because a drug can be designed specifically to bind to these unique features, and inhibit the protein's activity, without affecting the WT protein in healthy cells;

- gene mutations often result in the abnormal activation of some protein, through mechanisms like a *gain-of-function mutation* or *gene amplification*; although these proteins are considered druggable, the mutation does not necessarily change the protein in a way that allows for mutant-specific targeting, i.e. drugs may also target the WT version of the protein present in healthy cells, potentially leading to more side effects;

- some oncogenic mutations create novel molecular dependencies or vulnerabilities in cancer cells, which can be exploited by targeted therapies; these are called *actionable mutations* because they provide new targets for drug development that are specific to cancer cells and do not exist in normal cells.

While truly druggable mutations in the first category are relatively rare, many overactive or amplified targets still offer effective therapeutic opportunities due to their elevated expression levels or the significant dependence of cancer cells on these specific proteins. Additionally, mutations that currently lack targeted therapy options can still function as biomarkers to guide other therapeutic decisions [25].

Advances in targeted therapies have been significantly driven by technological progress in sequencing over the past two decades, particularly with the development of next-generation sequencing (NGS). The identification of both common and rare genetic mutations has launched research into targeted therapies against mutant proteins and aberrant molecular signaling pathways. Moreover, the discovery of the BCR-ABL fusion gene and the development of the BCR-ABL inhibitor *imatinib* marked a breakthrough in targeted cancer therapies, leading to numerous FDA-approved drugs [25]. However, the challenge of developing targeted therapies remains difficult, particularly for mutations that affect normal and cancerous proteins alike, or those for which no targeted therapies currently exist. The complexities of druggable mutations and their effects on treatment underscore the need for ongoing research and refinement in this area. Given the importance of fully understanding the role of mutations in cancer development, in order to improve targeted therapies and cancer treatment overall, research must focus on genomic mutations and their classification. The next chapter will discuss the existence of different types of mutations and the current techniques used to classify them.

Chapter 2

Classifying mutations

placeholder

2.1 Mutations and pathways

2.1.1 Passenger and driver mutations

In the previous sections, it was discussed how mutations play a critical role in cancer development, but not every aberration that occurs in a tumor is relevant to its proliferation. In fact, cancer mutations are generally divided into two categories: **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. *Driver genes* are genes that harbor at least one driver mutation, though it may also contain passenger ones [24].

Although the general consensus is that mutations are divided into these two categories, some studies do not fully agree with this dichotomous model [16]. Further complicating matters, the term *driver gene* has two distinct meanings in cancer research. Originally, the *driver-versus-passenger* concept was used to differentiate mutations that provide a selective growth advantage from those that do not. According to this definition, genes without driver mutations cannot be classified as driver genes. However, many genes that have few or no driver mutations are still referred to as driver genes in the literature. This includes genes that are overexpressed, underexpressed, or epigenetically altered in tumors. Although some of these genes might contribute significantly to cancer development, classifying all of them as *driver genes* may be misleading [24].

Despite these complexities, it is generally accepted that mutations can be categorized into the two described types. This classification is essential, as identifying driver mutations can significantly advance the development of targeted therapies, which may specifically target these driver mutations directly.

2.1.2   Problems with frequency analyses

To classify mutations into these two *drivers* and *passengers*, 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 [17].

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. For instance, some methods focus on comparing the mutation frequency of an individual gene to that of others within the same or related tumors, while accounting for sequence context and gene size. For genes with a very high number of mutations, such as TP53 or KRAS, most statistical methods will strongly suggest that these genes are drivers; these highly mutated genes are often referred to as *mountains* [17, 24].

However, genes with more than one, but still relatively few mutations, termed *hills*, are more common in cancer genome landscapes. In these cases, mutation frequency and context alone are insufficient to reliably identify driver genes, as background mutation rates can vary significantly among different patients and regions of the genome. Recent research on normal cells has shown that the mutation rate can vary more than 100-fold within the genome. In tumor cells, this variation can be even greater, affecting entire genomic regions in a seemingly random manner [24].

Moreover, because driver mutations are primarily found in genes involved in cell signaling pathways, in many cases different patients harbor mutations in different pathway loci. Consequently, driver mutations can vary widely between patient samples, even within the same cancer type, resulting in minimal overlap of mutated genes across sample pairs, even from the same patient, which further reduces the statistical power of frequency-based analyses [17, 29].

Therefore, methods based solely on mutation frequency can only prioritize genes for expanded investigation and cannot definitively identify driver genes that are mutated at relatively low frequencies [24].

2.1.3   Focus on pathways

An alternative method to assessing the recurrence of individual mutations or genes is to examine mutations within the context of cellular signaling and regulatory pathways, and biological considerations further support this approach. In particular, multiple alternative driver mutations in different genes can lead to similar downstream effects, hence the selective advantage is distributed across the frequencies of these gene alterations, which means that different mutations can affect the same pathway across various samples [1, 17]. This suggests that the focus should be on driver pathways rather than on individual driver mutations. Indeed, most recent cancer genome sequencing studies analyze known pathways for enrichment of somatic mutations, and methods have been developed to identify pathways that are significantly mutated

across multiple patients. Additionally, new algorithms have extended pathway analysis to genome-scale gene interaction networks [23].

However, pathway analysis relies on the prior identification of gene groups within known pathways. While some of them are well-documented and cataloged in multiple databases, our understanding remains incomplete. In particular, many databases aggregate all elements of pathways but often lack details on which specific components are active in particular cell types.

These limitations, along with the increasing number of sequenced cancer genomes, raise the question of whether it is possible to automatically identify groups of genes with driver mutations or mutated driver pathways directly from somatic mutation data, collected from a large number of patients [23]. This topic will be explored in the following sections, which will discuss the various techniques developed for identifying driver pathways based on mutation data.

2.1.4   Searching for driver pathways

Finding mutated driver pathways may seem implausible, because of the enormous number of possible gene sets to test, e.g. there are more than 10^{26} sets of 7 human genes. This makes it necessary to find specific properties or characteristics to guide the search efficiently. Fortunately, our current understanding of the somatic mutational processes in cancer suggests constraints on the expected patterns of mutations, which considerably narrow down the number of gene sets that need to be considered [23].

First, studies suggest that a major cancer pathway should be disrupted in a substantial number of patients, thus it is expected that most patients will exhibit aberrations in some gene within this pathway. Therefore, it is assumed that driver genes constituting a driver pathway are frequently mutated across many samples, a property that is referred to as **coverage** [23].

Second, while this feature is useful for identifying driver pathways, most techniques developed in recent years for recognizing driver pathways leverage a much stronger 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 and pathways [17]. Note that mutual exclusivity *does not affect different pathways*, as it occurs exclusively within a single pathway.

Therefore, *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 [17].

While the precise explanation for this phenomenon is not yet fully understood, several hypotheses appear plausible [8, 5, 23]:

- 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.

An example of the latter is provided by the gene pair ERG and SPOP, which are commonly overexpressed in patients with prostate cancer, but they are mutually exclusive due to their *synthetic lethality*. Wild-type SPOP facilitates the degradation of various proteins, including ZMYND11, which regulates androgen receptor (AR) signaling. Tumors with mutant ERG require reduced AR signaling to sustain their cancerous effects; therefore, mutant ERG upregulates WT SPOP to enhance the degradation of ZMYND11 and lower AR signaling. In contrast, when SPOP is mutated, it loses the ability to degrade ZMYND11, leading to its accumulation and increased AR signaling. This amplified AR signaling is incompatible with the function of mutant ERG, which relies on low AR signaling. Consequently, while ERG and SPOP mutations can each support oncogenic activity individually, their simultaneous aberration is not viable due to the conflicting requirements for AR signaling [2].

The next sections will focus on the algorithms and techniques developed to quantify levels of mutual exclusivity within gene groups.

2.2   Assessing mutual exclusivity

2.2.1   Challenges in quantifying mutual exclusivity

Finding an effective method to appropriately quantify the level of mutual exclusivity is not straightforward. 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.

Moreover, while soft mutual exclusivity of a pair of genes can be assessed using the Fisher's exact test, there is no agreed-upon method for analytically testing mutual exclusivity among more than two genes. For instance, one intuitive 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].

Due to the complexity of measuring mutual exclusivity, recent papers have proposed various approaches, based on different assumptions, which will be discussed in later sections.

2.2.2 A deterministic formalization of mutual exclusivity

One of the earliest [8] and most widely used mathematical formalizations for modeling and quantifying mutual exclusivity was introduced by Vandin et al. [23], the authors of an algorithm called Dendrix. But, before discussing it, some preliminary definitions are needed to provide context. In fact, all papers explored in this work will reference the following definitions.

Definition 2.1 (Mutation matrix). A **mutation matrix** is a matrix with m rows and n columns, where each row represents a patient and each column represents a gene, and the entry $a_{i,j}$ is equal to 1 if and only if gene j is mutated in patient i.

Example 2.1 (Mutation matrix). An example of a mutation matrix is the following:

	g_1	g_2	g_3
p_1	0	1	0
p_2	1	1	0
p_3	0	0	1

Table 2.1. A mutation matrix.

Definition 2.2 (Coverage of a gene). Given a gene g, the **coverage of** g denotes the set of patients which have g mutated, and it is defined as follows

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

Definition 2.3 (Mutual exclusivity). A set M of genes is **mutually exclusive** if no patient has more than one mutated gene of M, formally

$$\forall g, g' \in M \quad \Gamma(g) \cap \Gamma(g') = \varnothing$$

Definition 2.4 (Coverage of a set). Given a set M of genes, the **coverage of** M denotes the set of patients in which at least one of the genes of M is mutated, and it is defined as follows

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

Note that any gene set M of size k can be thought of as an $m \times k$ column submatrix of a mutation matrix A of size $m \times n$, 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. These two representations will be used interchangeably throughout this work.

To define an equation that mathematically assesses mutual exclusivity and coverage of a given gene set M, the formalization should reflect the following properties (discussed in previous sections):

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

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

To evaluate these two properties, Vandin et al. [23] introduced a measure that quantifies the trade-off between the two. First, they define the following formula, which measures M's *coverage overlap.*

Definition 2.5 (Coverage overlap)**.** Given a set M of genes, the **coverage overlap of** M is defined as follows:

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

Note that the sum in Definition 2.5 is the number of 1s in M's corresponding submatrix.

Example 2.2 (Coverage overlap)**.** Considering the mutation matrix in Example 2.1; if $M = \{g_1, g_2\}$, then

$$\omega(M) = |\Gamma(g_1)| + |\Gamma(g_2)| - |\Gamma(\{g_1, g_2\})| = |\{p_2\}| + |\{p_1, p_2\}| - |\{p_1, p_2\}| = 1 + 2 - 2 = 1$$

Indeed, $\omega(M)$ is the number of patients that are counted more than once in the sum, i.e. *the number of patients that have more than one mutation in M*. Note that $\omega(M) \geq 0$, with equality holding only if no patient has more than one mutated gene of M.

Definition 2.6 (Mutual exclusivity)**.** A gene set M is considered to be **mutually exclusive** if $\omega(M) = 0$.

Note that this definition matches the one given in Definition 2.3.

Finally, the equation developed by Vandin et al. [23] can be described.

Definition 2.7 (Weight of gene set)**.** Given a set of genes M, to take into account both coverage and coverage overlap, the following measure is introduced:

$$W(M) := |\Gamma(M)| - \omega(M) = 2\,|\Gamma(M)| - \sum_{g \in M} |\Gamma(g)|$$

This weight assesses the degree of mutual exclusivity among M's genes, and the extent to which their mutations cover the patient data. It does this by calculating

M's coverage and subtracting M's coverage overlap. Indeed, $W(M) = \Gamma(M)$ when M is mutually exclusive, because it has no coverage overlap. Therefore, the *optimal gene set* is the one that **maximizes** its weight, as a higher weight value indicates greater levels of coverage and mutual exclusivity.

As previously mentioned, a gene set M can be represented as a column submatrix of a mutation matrix A. Thus, finding the *optimal gene set* is equivalent to identifying the *optimal A's column submatrix*, which means that the following problem has to be solved.

> **Maximum Weight Submatrix Problem** (MWSP): Given an $m \times n$ mutation matrix A, and an integer $k > 0$, find a $m \times k$ submatrix of A that maximizes $W(M)$.

Finding the solution to this problem is computationally difficult, even for small values of k (e.g. there are $\approx 10^{23}$ subsets of size $k = 6$ of 20,000 genes). In fact, the following proof (provided by Vandin et al. [23]) shows that this problem is NP-Complete.

Theorem 2.1 (NP-Completeness of the MWSP)**.** The Maximum Weight Submatrix Problem is NP-Complete.

la prova della polinomialità della costruzione della matrice è giusta?

Proof. The proof of NP-Hardness is by reduction from the Independent Set Problem (ISP), which is known to be NP-Hard [12]. In the ISP, it is asked whether there is an independent set of size k in a given graph G. An independent set for $G = (V, E)$ is a set of vertices $I \subseteq V(G)$ such that there is no edge among the vertices of I, i.e.

$$\forall u, v \in I \mid u \neq v \quad (u, v) \notin E(G)$$

Given an instance of the ISP, a mutation matrix representing an instance of the MWSP is built in polynomial time as follows:

- let $\Delta := \max_{v \in G} \deg(v)$, and for each $v \in V(G)$ let $\mathcal{S}_v := \left\{ s_v^{(1)}, \ldots, s_v^{(\Delta - \deg(v))} \right\}$ be a set of variables; also, consider the following set

$$\mathcal{S} := \{ s_e \mid e \in E(G) \} \cup \left(\bigcup_{v \in V(G)} \mathcal{S}_v \right)$$

- build a matrix A of size $|\mathcal{S}| \times |V(G)|$, as illustrated below

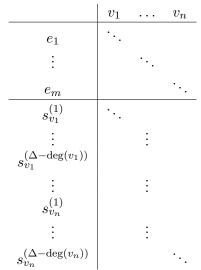|	v_1	$\ldots$	v_n
e_1	$\ddots$		
$\vdots$		$\ddots$	
			$\ddots$
e_m			
$s_{v_1}^{(1)}$	$\ddots$		
$\vdots$		$\vdots$	
$s_{v_1}^{(\Delta-\deg(v_1))}$			
$\vdots$		$\vdots$	
$s_{v_n}^{(1)}$		$\vdots$	
$\vdots$		$\vdots$	
$s_{v_n}^{(\Delta-\deg(v_n))}$			$\ddots$

Table 2.2. The described matrix.

- define A's cells as follows:

$$a_{s,v} = 1 \iff s = s_{(u,v)}, u \in V(G) \vee s \in \mathcal{S}_v$$

which means that $a_{s,v}$ will be 1 if and only if s is either a variable from the set $\{s_e \mid e \in E(G)\}$ where the edge e has v as endpoint, or s is a variable defined in $\mathcal{S}_v$.

Note that this matrix can be built in polynomial time, because:

- its first half contains $m \cdot n$ cells

- the vertex $\hat{v}$ that maximizes $|\mathcal{S}_{\hat{v}}|$ is such that

$$\deg(\hat{v}) = 1 \implies |\mathcal{S}_{\hat{v}}| = \Delta - \deg(\hat{v}) = \Delta - 1 = O(\Delta) = O(n)$$

and since there are n sets, the matrix's second half contains $(n \cdot n) \cdot n = n^3$ cells

therefore the time complexity to create it is $O\left(n^2 + nm\right)$. Moreover, note that:

i) $\forall v \in V(G) \quad |\Gamma(v)| = \Delta$ due to the added variables at the end of each column;

ii) $\forall u, v \in V(G) \quad \Gamma(u) \cap \Gamma(v) \neq \varnothing \iff (u,v) \in E$, since no pair of columns can have a 1 in the same row in the second half of A by definition of the sets $\mathcal{S}_{v_1}, \ldots, \mathcal{S}_{v_n}$, therefore $\Gamma(u)$ and $\Gamma(v)$ can have an intersection if and only if there is an edge $(u,v) \in E(G)$.

Hence, consider a set $M = \{v_1, \ldots, v_k\}$ of k columns of A. Note that:

- from (i) it follows that

$$\sum_{i=1}^{k} |\Gamma(v_i)| = k\Delta$$

and consequently $|\Gamma(M)| \leq k\Delta$, meaning that the largest value $|\Gamma(M)|$ can have is $k\Delta$; hence, from the equation in Definition 2.7, it follows that the maximum value $W(M)$ can reach is

$$W(M) = 2|\Gamma(M)| - \sum_{i=1}^{k} |\Gamma(v_i)|$$
$$= 2k\Delta - k\Delta$$
$$= k\Delta$$

- from (ii) it follows that $|\Gamma(M)| = k\Delta \iff \forall u, v \in V(G) \quad \Gamma(u) \cap \Gamma(v) = \varnothing \iff \forall u, v \in V(G) \quad (u, v) \notin E(G) \iff M$ is an independent set, by definition.

This means that $W(M)$ is maximized if and only if M is an independent set; therefore, the MWSP can be solved on A if and only if the ISP can be solved on G. $\qquad\square$

The approach developed by Vandin et al. [23] will be discussed in the next chapter.

2.2.3   A statistical approach

Babur et al. [1] criticize the metric developed by Vandin et al. [23] because it has a strong bias toward highly mutated genes, and in some instances, the excessive emphasis on coverage leads to false positives and negatives. . They propose a metric that extends Fisher's exact test — also known as *hypergeometric test* — to quantify the mutual exclusivity between multiple measurements.

ci sono esempi in file supplementari, li guardo?

Specifically, the alteration of a pair of genes is defined to be **mutually exclusive** *if their overlap in samples is significantly less than expected by chance*, and this can be assessed through a *hypergeometric test*. It is important to note that a uniform alteration frequency across may not always hold, particularly for hyper-mutated samples often resulting from prior mutations in DNA repair mechanisms. Addressing this heterogeneity is challenging, as each overlap in the null model has a different probability. This remains an open problem, and to partially mitigate it, albeit at the cost of statistical power, hyper-altered samples are excluded from the analysis.

Babur et al. [1] also developed a metric to assess the mutual exclusivity of a group of genes. Consider the following null hypothesis:

> H_0: *The specific member gene in the group is altered independently from the union of other alterations in the group.*

Using Dendrix's notation, H_0 states that for a given gene set M, for every gene $j \in M$, mutations in $\Gamma(j)$ are independent of alterations in $\Gamma(M - \{j\})$. In brief, H_0

states that any observed pattern of mutual exclusivity among gene alterations is due to random chance, not due to any underlying biological or oncogenic mechanism.

H_0 is then tested for each $j \in M$ by evaluating the co-distribution of i with the union of the others through Fisher's exact test, generating $|M|$ p-values. These p-values represent the probabilities for the independent distribution of each member gene. To ensure that every group member contributes to the pattern, the least significant — i.e. the largest — p-value of the group is used as the initial score of the group. With Dendrix's notation

$$s_0 := \max_{j \in M} H \left\langle \Gamma(j), \Gamma(M - \{j\}) \right\rangle \tag{2.1}$$

where s_0 is the initial score, and H is the hypergeometric test.

Since multiple groups are being tested, s_0 is affected by *multiple hypothesis testing*, because the probability of finding statistically significant results may increase by chance. To account for it,

finish this section

From this second set of p-values, the least significant one is selected as the multiple hypothesis testing corrected final score.

talk about the last paragraph, which is even less comprehensible

2.2.4   Extending the deterministic equation

While identifying individual driver pathways is crucial for cancer research and treatment, most cancer patients are likely to have driver mutations across multiple pathways. The metrics discussed so far do not account for multiple pathways simultaneously. As will be shown in the next chapter, these formalizations can be applied iteratively to identify multiple driver pathways, though this may not be the most precise approach. Leiserson et al. [17] refine the weight function of Vandin et al. [23] to extend their metric, enabling the assessment of mutual exclusivity across multiple driver pathways.

To effectively identify multiple pathways, it is necessary to establish criteria for evaluating potential *collections of gene sets*. Based on the same biological reasoning mentioned earlier, it is expected that each pathway will contain approximately one driver mutation. Furthermore, since each driver pathway is crucial for cancer development, it is expected that most patients will harbor a driver mutation in most driver pathways. Consequently, high exclusivity is predicted within the genes of each pathway, along with high coverage of each pathway individually. One metric that meets these criteria is the sum of individual weights of a given collection of gene sets, as defined below.

Definition 2.8 (Weight of a collection)**.** Given a collection of t gene sets $M = \{M_1, \ldots, M_t\}$, the **weight of** M is defined as follows:

$$W'(M) := \sum_{\rho=1}^{t} W(M_\rho)$$

This metric is employed by Leiserson et al. [17] in their algorithm, which will be explored in the next chapter.

2.2.5   A clustering approach

Another notable approach used in several studies to find mutually exclusive modules involves constructing gene graphs and identifying clusters based on specific criteria. This method is demonstrated by Hou et al. [13], who propose an algorithm designed to address the limitations of previous techniques. They argue that earlier approaches are generally inefficient for large datasets, lack consistency in results due to their randomized nature, and can only identify a few small modules. Additionally, these methods require restructuring whenever new biological information is added, whereas their approach has a notable degree of flexibility, as its objective function does not need to change with the addition of new data sources. Moreover, it has low computational cost, an important consideration in this context, as previously discussed. The following equations will describe how Hou et al. [13] formalized biological assumptions to define a technique able to achieve these results.

First, the structure of the graph will be described. Let $G = (V, E)$ be a *complete graph* of genes, thus an edge exists between any pair of vertices. Each edge $(u, v) \in E(G)$ is assigned two weights:

- a **positive weight** w_{uv}^+, which represents *the cost of placing u and v in different clusters*; thus, by making w_{uv}^+ large, placing u and v in different clusters is discouraged, and viceversa;

- a **negative weight** w_{uv}^-, which represents *the cost of placing u and v in the same cluster*; thus, by making w_{uv}^- large, placing u and v in the same cluster is discouraged, and viceversa.

The clustering algorithm aims to identify clusters of vertices (i.e. genes) that exhibit both *high coverage and mutual exclusivity within clusters.*

As mentioned earlier, this algorithm is quite flexible and allows the integration of weight values with information derived from external data. Specifically, Hou et al. [13] present multiple versions of their algorithm, depending on the type of information used to define the weights between edges. The following labels will be used to distinguish the components of the weights:

- the (e) label refers to *exclusivity*;

- the (c) label refers to *coverage*;

- the (n) label refers to *network information*;

- the (x) label refers to *expression data.*

To define the weights, linear combinations of these components are utilized. The different versions of this algorithm, and the various definitions of the weights, will be discussed in the following chapter. This secion will specifically focus on how Hou et al. [13] defined mutual exclusivity and coverage.

Let A be an $m \times n$ mutation matrix, as described in Definition 2.1. In addition, let C be an $m \times n$ matrix representing the CNV data, where $c_{i,j} = 0$ means that there is no change in the copy number of gene j in sample i, otherwise, the corresponding number reflects the deviation of the CNV number from its baseline — hence, C contains both positive and negative values. Following this, a binary matrix M is constructed combining A and C, as follows:

$$m_{i,j} = 0 \iff \begin{cases} a_{i,j} = 0 \\ l_{\text{cnv}} < c_{i,j} < h_{\text{cnv}} \end{cases} \tag{2.2}$$

where l_{cnv} and h_{cnv} are lower and upper bounds on copy numbers that determine the significance level. Thus, if $m_{i,j} = 0$, no mutation of gene j is recorded in sample i, otherwise gene j is *deemed mutated*.

Definition 2.9 (Coverage of a vertex). Given a vertex $u \in V(G)$, i.e. a gene, the **coverage of** u is defined as follows

$$\mathscr{S}(u) := \{i \mid m_{i,u} = 1\}$$

and it denotes the set of patients in which u is altered.

Note that $\mathscr{S}(u)$ corresponds to $\Gamma(u)$ under Dendrix's notation, but is defined through the M matrix respectively.

Now that the preliminaries have been covered, the following definitions will outline how mutual exclsuivity and coverage are defined.

Definition 2.10 (Mutual exclusivity component). The **mutual exclusivity component** between two genes $u, v \in V(G)$ is defined as follows:

$$w_{uv}^-(\text{e}) := a \cdot \frac{|\mathscr{S}(u) \cap \mathscr{S}(v)|}{\min(|\mathscr{S}(u)|, |\mathscr{S}(v)|)}$$

where a is a user-defined scaling parameter.

This ratio is often referred to as *Intersection over Minimum* (IoM), and suits the criteria of mutual exclusivity because the fewer patients who have both u and v mutated, the smaller the weight, making it more plausible that u and v are mutually exclusive, therefore the cost of placing them in the same cluster should be low. Note that

$$\forall u, v \in V(G) \quad a = 1 \implies 0 \leq w_{uv}^-(\text{e}) \leq 1 \tag{2.3}$$

Focusing on **coverage**, if two genes u and v increase the coverage of the set significantly, $w_{uv}^+(\text{c})$ should be large such that they are encouraged to be placed in the same cluster. Let

$$D(u, v) := |\mathscr{S}(u) \Delta \mathscr{S}(v)| \tag{2.4}$$

where Δ denotes the symmetric difference of two sets; a large value of $D(u,v)$ suggests that u and v should be placed in the same cluster. Also, let

$$\mathscr{D} := \{D(u,v) \mid u,v \in V(G)\} \tag{2.5}$$

and let $T(J)$ be the J-th percentile of the values in $\mathscr{D}$.

Definition 2.11 (Coverage component)**.** The **coverage component** is defined as follows:

$$w_{uv}^{+}(\text{c}) := \begin{cases} 1 & D(u,v) > T(J) \\ \dfrac{D(u,v)}{T(J)} & D(u,v) \leq T(J) \end{cases}$$

Note that, similar to Equation 2.3

$$\forall u,v \in V(G) \quad 0 \leq w_{uv}^{+}(\text{c}) \leq 1 \tag{2.6}$$

The next chapter will illustrate the various versions of the algorithm developed by Hou et al. [13].

Chapter 3

Finding driver mutations

In the previous chapter, various studies were discussed in terms of how they formalized biological assumptions, with particular emphasis on the metrics developed to assess *coverage* and *mutual exclusivity* within gene groups. This chapter will delve deeper into the algorithms employed by these studies to identify driver pathways using their respective metrics and hypotheses.

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. *De novo* approaches might lack sufficient information as they do not utilize pre-existing pathway databases, protein-protein interaction (PPI) networks or phenotype data. 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 [8, 17].

3.1 Dendrix

3.1.1 A greedy approach

Vandin et al. [23] introduced the most widely adopted metric in pathway discovery research, namely $W(M)$ (presented in Definition 2.7). In addition to this, they defined the Maximum Weight Submatrix Problem (MWSP), discussed in Section 2.2.2, and proposed the following *greedy algorithm*, called Dendrix (*de novo* [8]), to solve it.

Algorithm 3.1 *Greedy Dendrix*: given the set of all genes $\mathcal{G}$, and an integer k, the algorithm finds the set of genes M of size k that maximizes $W(M)$.

1: **function** GREEDYDENDRIX($\mathcal{G}$, k)
2:     $M := \{g_1, g_2\}$ such that M maximizes $W(M)$ ▷ pick the best gene pair
3:     **for** $i \in [3, k]$ **do**
4:       $\hat{g} := \arg\max_{g \in \mathcal{G}} W(M \cup \{g\})$ ▷ TODO SOURCE PER PARIM?
5:       $M = M \cup \{\hat{g}\}$
6:     **end for**
7:     **return** M
8: **end function**

Clearly, the time complexity of the algorithm is $O\left(n^2 + kn\right) = O\left(n^2\right)$ — where $n = |\mathcal{G}|$, therefore $k \leq n$ — because finding $\{g_1, g_2\}$ in line 2 requires $O\left(n^2\right)$ and the $\arg\max$ in line 4 has cost $O(n)$.

While this algorithm is efficient, there is generally no guarantee that it will identify the optimal set $\hat{M}$ that maximizes $W(\hat{M})$. However, Vandin et al. [23] prove that Algorithm 3.1 can correctly identify $\hat{M}$ with high probability when the mutation data come from the *Gene Independence Model* (GIM), which is described below.

Definition 3.1 (Gene Independence Model). Let A be an $m \times n$ mutation matrix, such that $\hat{M}$ is the *maximum weight submatrix* of A, and $|\hat{M}| = k$; the matrix A satisfies the **Gene Independence Model** (GIM) if and only if:

    i) each gene $g \notin \hat{M}$ is mutated in each patient with probability $p_g \in [p_L, p_U]$, independently of all other events;

    ii) $W(\hat{M}) = \Omega(m)$;

    iii) for all $M \subset \hat{M}$ of cardinality $l := |M|$, it exists $0 \leq d < 1$ such that

$$W(M) \leq \frac{l + d}{k} W(\hat{M})$$

Note that:

- condition (*i*) reflects the independence of mutations for genes outside the mutated pathway, a standard assumption for somatic single-nucleotide mutations, according to Vandin et al. [23];

- condition (*ii*) ensures that mutations in $\hat{M}$ cover a large number of patients and are mostly exclusive;

- condition (*iii*) means that each gene in $\hat{M}$ is important, so there are no subset of $\hat{M}$ that predominantly contributes to $W(\hat{M})$.

Although it is possible to efficiently obtain accurate results with high probability under the GIM, genes within $\hat{M}$ may exhibit *observed mutation frequencies* similar

ne forniscono una dimostrazione nel materiale supplementare, ma è lunga 3 pagine, e non credo sia oppurtuno inserire una cosa così lunga qui, penso che sia "beyond the scope"

capire bene questo modello

WHAT ARE THESE??

qua scrivono che la definizione di Omega è che $W(\hat{M}) = rm$ per $0 < r \leq 1$???????

to those of genes outside $\hat{M}$. This similarity can make it challenging to distinguish between them based solely on mutation frequency, regardless of the number of patients.

To illustrate this, consider the following scenario: observed gene mutation frequencies fall within the range of $[3 \times 10^{-5}, 0.13]$ — based on a background mutation rate of $\approx 10^{-6}$ [6]. If somatic mutations are measured in $n = 20,000$ human genes, and the size of $\hat{M}$ is 10, then approximately 2,400 patients are needed for the greedy algorithm to identify $\hat{M}$ with a probability of at least $1 - 10^{-4}$. While this number of patients is expected to be available from large-scale cancer sequencing projects [14], it exceeds current availability.

Therefore, in practical applications, the effectivness of this greedy algorithm depends on having mutation data from a sufficiently large number of patients. Moreover, the GIM model may be appropriate for certain types of somatic mutations, such as single-nucleotide aberrations, but may not be suitable for others. To address these limitations, Vandin et al. [23] developed an alternative approach, which will be detailed in the following section.

3.1.2 Using MCMC

To overcome the drawbacks of the aforementioned greedy algorithm, Vandin et al. [23] developed a *Markov Chain Monte Carlo* (MCMC) approach, which does not rely on assumptions about the distribution of mutation data or the number of patients. In particular, this MCMC method does not assume independence among mutations in different genes, making it particularly useful for analyzing copy-number aberrations (CNAs), which often involve correlated mutations due to amplification or deletion of adjacent genes.

Vandin et al. [23] employed a Metropolis-Hastings algorithm to sample sets $M \subseteq \mathcal{G}$ of k genes, with the following stationary distribution, proportional to $e^{cW(M)}$, for some constant $c > 0$

$$\pi(M) = \frac{e^{cW(M)}}{\sum_{R \in \mathcal{M}_k} e^{cW(R)}}$$

where $\mathcal{M}_k := \{M \subset \mathcal{G} : |M| = k\}$. While there are no guarantees on the rate of convergence of the Metropolis-Hasting algorithm to the stationary distribution, Vandin et al. [23] prove that their MCMC is rapidly mixing, therefore the stationary distribution is effectively reached in a practical number of steps.

The main idea of the MCMC algorithm involves constructing a Markov chain, where each state represents a collection of k columns from a given mutation matrix A, and transitions between these states occur by swapping one gene. Further details of the algorithm are discussed below.

Definition 3.2 (Dendrix's MCMC)**.** The **MCMC's procedure** of Dendrix is defined through the following steps:

1. *initialization*: given the set of all genes $\mathcal{G}$, choose an arbitrary subset $M_0 \subseteq \mathcal{G}$ of k genes;

2. *iteration*: for $t = 1, 2, \ldots$ derive M_{t+1} from M_t as follows:

   (a) choose a gene w uniformly, at random, from $\mathcal{G}$;

   (b) choose a gene v uniformly, at random, from M_t;

   (c) let $M'_t := (M_t - \{v\}) \cup \{w\}$;

   (d) let $\Delta_W := W(M'_t) - W(M_t)$;

   (e) let $P(M_t, w, v) := \min\left[1, e^{c\Delta_W}\right]$, where $c > 0$;

   (f) set $M_{t+1} := M'_t$ with probability $P(M_t, w, v)$, else $M_{t+1} := M_t$.

Note that:

- w is choosen from $\mathcal{G}$, thus if $w \in M_t$ then

$$M'_t = (M_t - \{v\}) \cup \{w\} = M_t - \{v\}$$

  which means that no genes were added to W'_t; this must be allowed because maximizing the weight may require removing genes already present in the current set, without adding new ones;

- Δ_W measures the change in the weight function with the new set, and the constant c is a scaling factor that adjusts the importance of this difference; note that

$$\Delta_W \geq 0 \implies e^{c\Delta_W} \geq 1 \implies P(M_t, w, v) = 1 \implies M_{t+1} = M'_t$$

  in fact when $\Delta_W > 0$ the weight has improved, therefore the next iteration should start from M'_t; conversely

$$\Delta_W < 0 \implies e^{c\Delta_W} < 1 \implies P(M_t, w, v) = e^{c\Delta_W}$$

  which means that when $\Delta_W < 0$ (i.e., the weight has decreased) the change will be perfomed with probability $e^{c\Delta_W}$ — note that this value is still close to 1 if the weight has not decreased significantly.

Vandin et al. [23] prove that their MCMC is rapidly mixing for some $c > 0$, but details of this proof are beyond the scope of this work. The following sections will briefly discuss some of the results obtained with their algorithm.

3.1.3 Results

This section will briefly discuss the results reported by Vandin et al. [23] from running the MCMC algorithm on real data. To improve the efficiency of the algorithm, the mutation matrix was optimized by merging genes $T = \{g_1, \ldots, g_h\}$ that were mutated in the same patients into larger *metagenes* g_T, where each metagene represents the combined mutations occurring in those same patients. The MCMC algorithm samples gene sets with a frequency proportional to their weights, thus in order to focus on high-weight sets, only those with a frequency of at least 1% were reported.

Vandin et al. [23] applied their MCMC algorithm (with the constant c set to $c = 0.5$) to analyze somatic mutations obtained from high-throughput genotyping of 238 oncogenes across 1,000 patients, spanning 17 cancer types (a study conducted by Thomas et al. [22]). A mutation matrix was constructed with 298 patients and 18 mutation groups, based on the groupings from Thomas et al. [22]. They ran the MCMC algorithm on gene sets ranging in size from 2 to 10, sampling every 10,000 iterations after running the algorithm for 10 million iterations. The analysis identified a set of 8 mutation groups, altered in 94% of patients with at least one mutation, totaling 295 mutations ($p < 0.01$). They state that these mutated genes are linked to well-known cancer pathways. Additionally, two sets of size 10, which included the initial 8 mutation groups, were found in 95% of patients, accounting for 302 mutations ($p < 0.01$).

They also applied their algorithm to somatic mutations in lung adenocarcinoma, using data from The Cancer Genome Atlas (TCGA) [6], which included 188 patients and 623 genes, of which 356 were found to be mutated in at least one patient. For gene sets of size $k = 2$, the pair (EGFR, KRAS) was identified in 99% of samples, covering 90 patients, with no overlap, indicating high mutual exclusivity; additionally, when analyzing sets of size $k = 3$, the algorithm uncovered a novel triplet (EGFR, KRAS, STK11), appearing in 8.4% of samples; the significance of both the pair and the triplet was confirmed through permutation tests. Vandin et al. [23] highlight that all three genes are involved in regulating the mTOR pathway, which is known to be crucial in lung adenocarcinoma. To identify additional sets, the MCMC algorithm was rerun after removing the triplet, revealing the pair (ATM, TP53) with a sampling frequency of 56%, covering 76 patients. Although these reported sets showed high exclusivity, their relatively low coverage suggests that they may not represent complete driver pathways. Indeed, Vandin et al. [23] propose that this could be due to the limited number of genes analyzed or the focus on specific mutation types. Moreover, there was no significant overlap between patients with mutations in (ATM, TP53) and those with mutations in (EGFR, KRAS, STK11), suggesting that these gene sets likely belong to distinct biological pathways.

The MCMC algorithm was also applied to mutation data from 84 patients with glioblastoma multiforme (GBM), analyzing somatic mutations across 601 genes, resulting in a total of 453 mutations, with 223 genes found to be mutated in at least one patient (filtered using CNAs). For gene sets of size $k = 2$, the most frequently sampled pair was (CDKN2B, CYP27B1), appearing 18% of the time, and for $k = 3$, the triplet (CDKN2B, RB1, CYP27B1) was sampled 10% of the time — a permutation test confirmed the significance of this triplet. The analysis revealed that CYP27B1 had a nearly identical mutation profile to a metagene composed of six adjacent genes, but was excluded due to an additional mutation in a single patient. The metagene's amplification likely targeted CDK4, suggesting that the key triplet of interest was (CDKN2B, RB1, CDK4), which is part of the RB1 signaling pathway, associated with shorter survival in GBM patients. After removing this triplet, the pair (TP53, CDKN2A) was sampled with 30% frequency, linked to the p53 tumor suppression pathway. Further analysis, after removing both sets, revealed the pair (NF1, EGFR), which was sampled 44% of the time and is part of the RTK pathway, crucial for cell proliferation and survival.

All these findings highlight the ability of the MCMC algorithm to identify significant gene sets related to known cancer pathways. The following section will introduce an alternative approach to solving the MWSP, and will provide an extension of the problem.

3.2 Multi-Dendrix

3.2.1 An alternative approach to the MWSP

Vandin et al. [23] propose a solution to the MWSP (described in Section 2.2.2) that may not appear immediately intuitive, given that the problem itself resembles an optimization problem. Indeed, Leiserson et al. [17], the authors of Multi-Dendrix (*de novo* [8]), present an alternative approach by formulating the problem as an *Integer Linear Program* (ILP), called $\text{Dendrix}_{ILP}(k)$, which is described below.

To begin, it is necessary to define two sets of indicator variables: consider a gene set M, described by a set of variables, one for each gene $j \in M$, defined as follows

$$I_M(j) = 1 \iff j \in M \tag{3.1}$$

and a set of indicator variables, one for each patient i, expressed in this form

$$C_i(M) = 1 \iff \exists g \in M \mid i \in \Gamma(g) \tag{3.2}$$

therefore $C_i(M)$ is equal to 1 if and only if M *covers* the i-th patient.

The ILP formulation provided by Leiserson et al. [17] is illustrated below.

Definition 3.3 ($\text{Dendrix}_{ILP}(k)$)**.** $\text{Dendrix}_{ILP}(k)$ is defined by the following ILP:

$$\text{maximize} \sum_{i=1}^{m} \left(2 \cdot C_i(M) - \sum_{j=1}^{n} I_M(j) \cdot a_{i,j} \right), \tag{3.3}$$

$$\text{subject to} \sum_{j=1}^{n} I_M(j) = k, \tag{3.4}$$

$$\sum_{j=1}^{n} I_M(j) \cdot a_{i,j} \geq C_i(M), \tag{3.5}$$

$$\text{for } 1 \leq i \leq m.$$

Note that Equation 3.3 uses the second version of the definition provided in Definition 2.7, and Equation 3.4 limits the size of M to be exactly k. Moreover, note that Equation 3.5 only forces $C_i(M) = 0$ when the i-th patient has no mutated genes in M, but does not force $C_i(M) = 1$ when the patient has at least one, as required by Equation 3.2. However, the objective function will be maximized when $C_i(M) = 1$, thus Equation 3.2 is satisfied.

Lemma 3.1 (Correctness of Dendrix$_{ILP}(k)$)**.** *Given a gene set M, the sum in Equation 3.3 correctly evaluates $W(M)$.*

Proof. Rearranging the terms in Equation 3.3

$$\sum_{i=1}^{m}\left(2\cdot C_i(M) - \sum_{j=1}^{n} I_M(j)\cdot a_{i,j}\right) = 2\sum_{i=1}^{m} C_i(M) - \sum_{i=1}^{m}\sum_{j=1}^{n} I_M(j)\cdot a_{i,j}$$

and it is trivial to check that

$$|\Gamma(M)| = \sum_{i=1}^{m} C_i(M)$$

since it it true by definition, and

$$\sum_{g\in M}|\Gamma(g)| = \sum_{i=1}^{m}\sum_{j=1}^{n} I_M(j)\cdot a_{i,j}$$

because the RHS counts the number of cells of A such that $a_{i,j} = 1$ for every $j \in M$. $\qquad\square$

The next section will discuss how Leiserson et al. [17] extended this ILP formulation to enable the search for multiple driver pathways.

3.2.2   The ILP

As outlined in Section 2.2.4, Leiserson et al. [17] propose that the most effective approach for this research is to identify multiple driver pathways rather than focusing on a single one. To accomplish this, they extended the weight metric introduced by Vandin et al. [23] to find a collection of gene sets that maximizes the sum of their individual weights. Specifically, they extended the MWSP as follows:

> **Multiple Maximum Weight Submatrices Problem** (MMWSP):
> Given an $m \times n$ mutation matrix A, and integer $t > 0$, and two integers $k_{\min}, k_{\max} \geq 0$, find a collection $M = \{M_1, \ldots, M_t\}$ of column submatrices that maximizes $W'(M)$, where each submatrix M_ρ — for $1 \leq \rho \leq t$ — has size $m \times k_\rho$ for some $k_{\min} \leq k_\rho \leq k_{\max}$.

Note that the sets in the optimal collection may vary in size, as different pathways are likely to have different lengths; additionally, note that this problem is NP-Hard, as for the case where $t = 1$ (proof provided in Theorem 2.1) . Furthermore, Leiserson et al. [17] state that collections M with a large value of $W'(M)$ are also likely to exhibit higher coverage $\Gamma(M_\rho)$, for each individual gene set M_ρ. Consequently, optimal solutions tend to produce collections where many patients have mutations in more than one gene set, which may involve pairs or larger groups of co-occurring mutations — a phenomenon observed in real cancer data.

fix this probably

The ILP developed by Leiserson et al. [17] to simultaneously search for multiple driver pathways is described below.

Definition 3.4 (Multi-Dendrix)**.** Multi-Dendrix is defined by the following ILP:

$$\text{maximize} \sum_{\rho=1}^{t} \sum_{i=1}^{m} \left(2 \cdot C_i(M_\rho) - \sum_{j=1}^{n} I_{M_\rho}(j) \cdot a_{i,j} \right), \tag{3.6}$$

$$\text{subject to} \sum_{j=1}^{n} I_{M_\rho}(j) \cdot a_{i,j} \geq C_i(M_\rho), \tag{3.7}$$

$$k_{\min} \leq \sum_{j=1}^{n} I_{M_\rho}(j) \leq k_{\max}, \tag{3.8}$$

$$\text{for } 1 \leq i \leq m, \ 1 \leq \rho \leq t,$$

$$\sum_{\rho=1}^{t} I_{M_\rho}(j) \leq 1, \ 1 \leq j \leq n. \tag{3.9}$$

Note that:

- Equation 3.6 and Equation 3.7 extend Equation 3.3 and Equation 3.4 respectively;

- Equation 3.8 allows each gene group to have a size between $k_{\min}$ and $k_{\max}$;

- Equation 3.9 forces each gene to appear in *at most 1 set* within the collection.

Moreover, Leiserson et al. [17] state that this ILP can be extended to allow the gene sets of the collection to overlap, since the genes in the intersection may be involved in multiple biological processes. Hence, Equation 3.9 is replaced with the following equation:

$$\sum_{\rho=1}^{t} I_{M_\rho}(j) \leq \Delta, \quad 1 \leq j \leq n \tag{3.10}$$

where Δ is the maximum number of gene sets a gene can be a member of, and the following constraint is added:

$$\sum_{j=1}^{n} \sum_{\rho=1}^{t} \sum_{\substack{\rho'=1 \\ \rho \neq \rho'}}^{t} I_{M_\rho}(j) \cdot I_{M_{\rho'}}(j) \leq \tau, \quad 1 \leq \rho \leq t \tag{3.11}$$

where τ is the maximum size of the intersection between two gene sets.

3.2.3 Comparing Multi-Dendrix with Iter-Dendrix

A normal cell has to acquire multiples capabilities in order to become a cancer cell, and studies identified six such capabilities. Consequently, it is anticipated that only a few pathways will be mutated, with mutations in the genes of each pathway exhibiting both high exclusivity and high coverage. Our goal is to identify sets of genes within these pathways. If the gene sets within each pathway are disjoint, an iterative approach can be effective: after identifying a high-weight set M, we remove the genes in M from the analysis and search for high-weight sets in the reduced mutation matrix. This approach will successfully identify disjoint sets M_1 and M_2 with high weight, as exclusivity is assessed within sets, not between them. However, if M_1 and M_2 share genes, removing one set might also remove part of the other. If the overlap is minimal, we can still identify the remaining portion of the other set. The challenge of identifying two sets M_1 and M_2 that both exhibit high exclusivity and coverage, while allowing for some overlap, remains an intriguing open problem [23].

rephrase this para-graph

Leiserson et al. [17] compare the outputs of their ILP with an iterative version of the Dendrix greedy algorithm (discussed in Algorithm 3.1), which they refer to as Iter-Dendrix, running Dendrix multiple times to produce a collection of gene sets.

Algorithm 3.2 *Iter-Dendrix*: given the set of all genes $\mathcal{G}$, an integer k, and an integer t, the algorithm finds the collection M of t gene sets of size k that maximizes $W'(M)$.

1: **function** IterDendrix($\mathcal{G}$, k, t)
2:     $M := \varnothing$
3:     **for** $i \in [1, t]$ **do**
4:         $M_i := \texttt{greedyDendrix}(\mathcal{G}, k)$ ▷ procedure defined in Algorithm 3.1
5:         $M = M \cup \{M_i\}$
6:         $\mathcal{G} = \mathcal{G} - M_i$
7:     **end for**
8:     **return** M
9: **end function**

Denoting with M and I the collections of gene sets obtained from Multi-Dendrix and Iter-Dendrix respectively, Leiserson et al. [17] state that $W'(M) \geq W'(I)$. They also argue that M could contain sets with strictly greater weight than comprising I due to several factors:

loro dicono che sia ovvio ma non capisco perché dovrebbe essere così ovvio

- there may be multiple gene sets I_ρ that maximize $W(I_\rho)$ on the ρ-th iteration of Iter-Dendrix, and this version of Dendrix can only extend one of these sets;

- the gene set I_ρ that maximizes $W(I_\rho)$ selected by Iter-Dendrix in the ρ-th iteration may not be a member of M, since M could include gene sets that are suboptimal when considered in isolation;

- when $k_{\min} < k_{\max}$ Multi-Dendrix may choose gene sets with fewer than $k_{\max}$ genes if doing so maximizes the overall weight $W'(M)$.

Leiserson et al. [17] state that all of these scenarios occur when analyzing real mutation data.

3.3 MDPFinder

3.3.1 The genetic algorithm

placeholder.

The approach used by Zhao et al. [29] involves a *Genetic Algorithm* (GA), which is versatile and flexible, and can be used to optimize a wide variety of scoring functions. In addition, they state that the GA approach is particularly relevant to the current problem due to its conceptual alignment with the notions of *gene* and *mutation*. It models genetic variation within a population, evolving through a process of random selection, thereby avoiding the need to enumerate all possible solutions.

sezione in cui scrivo che mpdfinder critica la soluzione esatta trovata da multi-dendrix, che nonostante sia esatta potrebbe non essere reale; in questa sezione va menzionato che l'algoritmo si chiama MDPFinder, e va scritto che è knowledge-based

Definition 3.5 (Hypothesis space)**.** A **member** of the population is defined by a binary string of length n, i.e. the number of genes. Given a gene set M, the value of the i-th position of an individual represents the membership of the i-th gene in M.

Therefore, the **hypothesis space** is constituted by all the possible binary strings with length n that have k 1s, namely

$$\mathcal{H} := \left\{ (x_1, \ldots, x_n) \mid x_i \in \{0,1\}, i \in [1,n], \sum_{j=1}^{n} x_j = k \right\}$$

Definition 3.6 (Fitness function)**.** The **fitness** f_i of each individual h_i (its corresponding gene set is M_i) of the population is defined as the rank r_i of the score $W(M_i)$, in the ascending order :

$$\forall h_i \in \mathcal{H} \quad f_i := r_i$$

SECONDO ME È DESCENDING ALTRIMENTI PRENDI QUELLO COL PESO MINORE CHE NON HA SENSO, probabilmente non sto capendo il senso della frase

Definition 3.7 (Selection probability)**.** Given the rank r_i of an individual h_i based on its score, the **selection probability** is defined as follows:

$$p_i = \frac{2r_i}{P(P+1)}$$

where P is the population size. The individual with the highest fitness value is most likely to be transferred into the next generation.

This selection operator is based on roulette wheel selection, which states that the probability of choosing an individual is equal to

$$p_i = \frac{f_i}{\sum_{j=1}^{P} f_j} = \frac{r_i}{\frac{P(P+1)}{2}} = \frac{2r_i}{P(P+1)}$$

which is precisely the equation in Definition 3.7.

Definition 3.8 (Crossover operator)**.** The **crossover operator** specifies the breeding process as follows: the offspring inherits the variables shared by both parents, while the non-shared ones are selected from the symmetric difference of the parents' genetic makeup.

Definition 3.9 (Mutation operator)**.** The **mutation operator** randomly sets the value of one variable from 1 to 0, and changes another variable value from 0 to 1, ensuring the feasibility of every offspring.

<div style="float:left">*fixa questa definizione*</div>

Definition 3.10 (Local search)**.** To prevent premature convergence and enhance the accuracy of the algorithm, a local search strategy is employed to improve search performance. In particular, the values of two variables are randomly altered, as the mutation operator. If this adjustment improves the current solution, it is accepted; the search is terminated once all variables have been tested with this routine.

Definition 3.11 (GA procedure)**.** The following are the details of the **GA procedure**:

1. *population generation*: a random population of size P and mutation rate p_m is generated, where $P = n$ (i.e. the number of available genes);

2. *breeding*: for each iteration, P couples are selected from the current population, based on p_i, and each couple generates an offspring;

3. *mutation*: each offspring may optionally receive a mutation with probability p_m;

4. *selection*: all parents and offspring are ranked based on their scoring values, and the top P individuals are selected to form the next generation (this is referred to as truncation selection);

5. *local search*: verify if the iteration is stuck in a local solution (e.g. if the maximum scoring value does not improve over two consecutive iterations); if this is the case, perform a local search;

6. *termination*: proceed as such until the termination criterion is met (e.g. if the current maximum scoring value does not improve over 10 consecutive iterations); if this occurs, then end the procedure.

3.3.2 The integration procedure

In practical applications, multiple optimal solutions may exist. Additionally, due to data noise and other factors, the solutions considered optimal — i.e. the ones with the highest $W(M)$ — may not necessarily be the most relevant in a biological context. To identify the most biologically meaningful solutions, other types of data are integrated to refine the results. Specifically, the GA procedure is extended by incorporating gene expression data to enhance its performance. The integrative model is developed based on the observation that genes within the same pathway typically collaborate to perform a specific function. Consequently, the expression

profiles of gene pairs within the same pathway often exhibit higher correlations than those in different pathways. This characteristic can be leveraged to distinguish between gene sets that have the same score: the model focuses on detecting gene sets whose scores $W(M)$ are close to the optimal solution, but whose member genes display stronger correlations with each other.

Definition 3.12 (Integrative measure)**.** Given an $m \times n$ mutation matrix A, an expression matrix E with the same dimensions , and an A's submatrix M of size $m \times k$, the integrative model is defined by the following **measure**:

$$F_{ME} := W(M) + \lambda \cdot R(E_M)$$

where E_M is E's expression submatrix that corresponds to M, and $R(E_M)$ is described by the following equation:

$$R(E_M) = \sum_{j_1=1}^{n} \sum_{\substack{j_2=1 \\ j_1 \neq j_2}}^{n} \frac{|\mathrm{pcc}(x_{j_1}, x_{j_2})|}{\frac{k(k-1)}{2}}$$

where $\mathrm{pcc}(\cdot)$ is the Pearson correlation coefficient, and x_j is the expression profile of gene j .

Note that

$$0 \leq R(E_m) \leq 1$$

and $W(M)$ is an integer, therefore when $\lambda = 1$ the value of F_{ME} can be used to discriminate the gene sets with the same $W(M)$. Moreover, for values of $\lambda \geq 1$, the gene set with strong correlation and approximate exclusivity can be identified.

3.4 Mutex

3.4.1 A different greedy method

To identify the most mutually exclusive group, Babur et al. [1] employ a greedy algorithm called Mutex (*knowledge-based* [8]), which is applied to a directed graph constructed from databases containing information about biological pathways .

The search begins by initializing a set with an altered gene as the seed and then expanding the group greedily with the next best candidate gene. Candidate genes are selected such that, after their addition, the group still has a common downstream gene that can be accessed without passing through any non-member genes (the common downstream gene may also be a member of the group) . The group is expanded with the candidate that improves the group score the most. The process continues until no candidates remain or the group reaches a preset size threshold. The algorithm outputs a group and its score for each seed gene .

placeholder.

Definition 3.13 (Proximity)**.** Given a gene graph, the **proximity** of a gene G includes not only the genes directly adjacent to g but also those that share downstream targets in the pathway with g.

nel paper non è menzionato cosa questa "expression matrix" contenga all'interno, c'è una foto con dei colori e basta, però ho visto che "expression matrix" è un termine che si usa per indicare un tipo di matrici solamente che contengono una cosa ben specifica, posso assumere che si sappia già di cosa si parla o devo spiegare cosa sono?

non è spiegato cosa questo voglia dire ma io assumo sia un vettore colonna di E; inoltre, non ho idea di cosa sia il valore $R(E_M)$

menziono i database che hanno usato?

ci sono delle figure con un grafo e l'insieme che viene proressivamente espanso dall'algoritmo, forse potrei riprodurle e inserirle per chiarezza?

qui menzionano una cosa inerente al controllo dell'FDR ma per ora la ometto perché non so che significa, credo andrebbe inserita

una cosa carina sarebbe scrivere lo pseudocodice dell'algoritmo greedy (sarebbe anche più carino a quel punto dimostrarne la correttezza ma non credo sia possibile vista la natura statistica della ricerca effettuata), loro non lo forniscono e lo trattano solo a pa-

3.5   C3

3.5.1   Multiple versions

Hou et al. [13] define three methods for assigning weights to the graph to perform the vertex clustering algorithm called C3 (*knowledge-based* [8]):

1. **ME-CO**, where w^- depends on *mutual exclusivity* and w^+ depends on *coverage*;

2. **NI-ME-CO**, where w^- depends on *mutual exclusivity* and w^+ depends on *coverage* and *network information*;

3. **EX-ME-CO**, where w^- depends on *mutual exclusivity* and w^+ depends on *coverage*, *network information* and *expression data*.

As mentioned earlier in **??**, w^- depends solely on the mutual exclusivity component, while the value of w^+ depends on the version of the algorithm chosen.

3.5.2   The standard version

Definition 3.14 (ME-CO)**.** In the **ME-CO** version of the algorithm, the following definitions apply:

$$\forall u, v \in V(G) \quad w^-_{uv} := w^-_{uv}(\mathrm{e}) \tag{3.12}$$

$$\forall u, v \in V(G) \quad w^+_{uv} := w^+_{uv}(\mathrm{c}) \tag{3.13}$$

The definitions for $w^-_{uv}(\mathrm{e})$ and $w^+_{uv}(\mathrm{c})$ are provided in Definition 2.10 and Definition 2.11 respectively.

qui menzionano che se succede una certa cosa fanno un rescaling, lo menziono?

3.5.3   Integrating network information

Pan-cancer studies, as reported in multiple papers, have demonstrated a significant relationship between network topology and the distribution patterns of cancer drivers. Specifically, the impact of deleterious mutations on the phenotype can be mitigated by certain configurations of the corresponding protein complexes, while other arrangements can amplify their effect. For example, most variants found in healthy individuals tend to be located at the periphery of the interactome, where they do not affect network connectivity. In contrast, cancer-driver somatic mutations are more likely to occur in central, internal regions of the interactome and within highly integrated components.

To precisely quantify the network distances between driver variants, Hou et al. [13] computed the pairwise network distances between genes within a large pathway,

comprising 8726 genes, by using an implementation of the standard Dijkstra algorithm. To reduce the computational cost of running Dijkstra's algorithm $O\left(8726^2\right)$ times, 1000 pairs were randomly selected for this test. Using the most comprehensive known driver list from the Cancer Gene Census (CGC) [11], the same distances were calculated for driver genes, this time for all gene pairs. The resulting distribution of shortest paths is shown in Figure 1 , revealing that the average shortest distance between drivers is significantly smaller than that between two randomly selected genes.

placeholder.

These findings indicate that network distance and connectivity information should be considered when identifying potential driver mutations. This can be achieved by adjusting the positive weight of edges connecting two genes: if both endpoint genes are drivers, they should be sufficiently central within a given pathway, close to other known drivers, or to each other.

From the KEGG Database [15], a (rather sparse) undirected graph G' was retrieved, where each vertex represents a gene and the edges describe interactions between them. Note that $|V(G)| = |V(G')| = n$. For each vertex $u \in V(G')$, let $\mathscr{N}(u)$ denote the set of neighbors of u, and let $\mathscr{N}'(u) = \mathscr{N}(u) \cup \{u\}$. Also, let

$$f(u,v) := \frac{|\mathscr{N}'(u) \cap \mathscr{N}'(v)|}{|\mathscr{N}'(u) \cup \mathscr{N}'(v)|} \tag{3.14}$$

which is known as the Jaccard similarity coefficient; a large value of $f(u,v)$ indicates that u and v are well connected in G' and are likely involved in the same pathway, suggesting that they should be clustered together. Furthermore, let

$$\mathscr{F} := \{f(u,v) \mid u,v \in V(G')\} \tag{3.15}$$

and let $T'(J')$ be the J'-the percentile of the values in $\mathscr{F}$.

Definition 3.15 (Network information component)**.** The **network information component** is defined as follows:

$$w_{uv}^+(\mathrm{n}) := \begin{cases} 1 & f(u,v) > T'(J') \\ \dfrac{f(u,v)}{T'(J')} & f(u,v) \le T'(J') \end{cases}$$

Definition 3.16 (NI-ME-CO)**.** The **NI-ME-CO** version of the algorithm is defined by the following equations:

$$\forall u,v \in V(G) \quad w_{uv}^- := w_{uv}^-(\mathrm{e}) \tag{3.16}$$

$$\forall u,v \in V(G) \quad w_{uv}^+ := w_1 w_{uv}^+(\mathrm{c}) + w_2 w_{uv}^+(\mathrm{n}) \tag{3.17}$$

where $w_1, w_2 \ge 0$ and $w_1 + w_2 = 1$.

placeholder.

3.5.4   Integrating expression data

Expression data may be integrated through the positive weights, based on the assumption that co-expressed genes are likely to be involved in the same function or cancer pathway. Therefore, genes with strong positive or negative co-expression should be encouraged to cluster together.

Given a vertex $u \in V(G)$, let $\mathbf{z}(u)$ be the vector of the time-evolving expression values of u. Thus, let

$$g(u,v) := \frac{|\langle \mathbf{z}(u), \mathbf{z}(v) \rangle|}{||\mathbf{z}(u)|| \, ||\mathbf{z}(v)||} \tag{3.18}$$

where $\langle \mathbf{a}, \mathbf{b} \rangle$ denotes the inner product of the vectors $\mathbf{a}$ and $\mathbf{b}$, while $||\mathbf{a}||$ stands for its L^2 norm. This equation is known as the cosine similarity, since the ratio that defines $g(u,v)$ is equal to the cosine of the angle between $\mathbf{z}(u)$ and $\mathbf{z}(v)$ — the only difference being the absolute value in the numerator to capture both positive and negative correlations. A large value of $g(u,v)$ suggests that the expression vectors of u and v are highly correlated, hence they should be clustered together. Note that

$$\forall u, v \in V(G) \quad 0 \leq g(u,v) \leq 1$$

Moreover, let

$$\mathscr{G} := \{ g(u,v) \mid u,v \in V(G) \} \tag{3.19}$$

and let $T''(J'')$ be the J''-th percentile of the values in $\mathscr{G}$.

Definition 3.17 (Expression data component)**.** The **expression data component** is defined as follows:

$$w_{uv}^+(\mathrm{x}) := \left\{ \begin{array}{ll} 1 & g(u,v) > T''(J'') \\ \dfrac{g(u,v)}{T''(J'')} & g(u,v) \leq T''(J'') \end{array} \right.$$

Definition 3.18 (EX-ME-CO)**.** The **EX-ME-CO** version of the algorithm is defined by the following equations:

$$\forall u, v \in V(G) \quad w_{uv}^- := w_{uv}^-(\mathrm{e}) \tag{3.20}$$

$$\forall w_{uv}^+ := w_1 w_{uv}^+(\mathrm{c}) + w_2 w_{uv}^+(\mathrm{x}) \tag{3.21}$$

where $w_1, w_2 \geq 0$ and $w_1 + w_2 = 1$.

placeholder.

Margin notes:

la prima volta che lessi questo paper non trovai niente sul dove presero queste informazioni, e tutt'ora non mi pare che lo menzionino da nessuna parte; scrivono solo che le informazioni le prendono dal TCGA e dal KEGG, suppongo a questo punto che queste info siano ottenibili dal TCGA ma dovrei controllare manualmente

questa frase l'ho copiata da loro ma non capisco in che senso

parlare del rescaling?

3.5.5 Other versions

Hou et al. [13] also mention that other combinations can be used, with appropriate adjustments to the weights, such as the following version, which will be referred to as NI-EX-ME-CO .

loro non danno un nome ma glielo sto dando io, è un problema?

Definition 3.19 (NI-EX-ME-CO)**.** The **NI-EX-ME-CO** version of the algorithm is defined by the following equations:

$$\forall u, v \in V(G) \quad w_{uv}^- := w_{uv}^-(\text{e}) \tag{3.22}$$

$$\forall w_{uv}^+ := w_1 w_{uv}^+(\text{c}) + w_2 w_{uv}^+(\text{n}) + w_3 w_{uv}^+(\text{x}) \tag{3.23}$$

where $w_1, w_2, w_3 \geq 0$ and $w_1 + w_2 + w_3 = 1$.

3.5.6 The clustering ILP

The classical formulation of correlation clustering does not impose any restrictions on cluster sizes. However, all known driver identification methods inherently include cluster size limits, as these sizes directly affect the computational complexity of the algorithms . Therefore, Hou et al. [13] introduce a cluster size constraint by assuming that all clusters are of size k at most; clearly, setting k equal to the total number of vertices effectively removes this constraint, allowing flexibility in cluster size selection.

portano l'esempio di un altro algoritmo chiamato CoMEt che oltre size 10-12 "fails to operate", ha senso inserirlo? forse si

Another reason for imposing a cluster size limit is the expectation that driver genes of specific cancer types will be grouped together, and recent findings indicate that only a small number of drivers are typically present in any given cancer type . If clusters are too large, they may include drivers from multiple cancer types, hiding the detailed separation of the drivers. Furthermore, introducing cluster size constraints helps to avoid the limitations of many clustering algorithms that often produce non-informative "giant clusters" or singleton clusters .

non menzionano nessuna fonte dalla quale tirano fuori questa informazione, ritrovatela da solo e aggiungi qualcosa

di nuovo, non ci sono citazioni o esempi menzionati, ma vorrei mettecene

Definition 3.20 (C3's ILP)**.** The **C3 algorithm** can be defined by the following ILP:

$$\text{minimize} \sum_{e \in E(G)} (w_e^+ x_e + w_e^-(1 - x_e)), \tag{3.24}$$

$$\text{subject to } x_{uv} \leq x_{uz} + x_{zv}, \ u, v, z \in V(G) \text{ distinct}, \tag{3.25}$$

$$\sum_{\substack{v \in V(G) \\ u \neq v}} (1 - x_{uv}) \leq k, \ u \in V(G), \tag{3.26}$$

$$x_e \in \{0, 1\}, \ e \in E(G). \tag{3.27}$$

In this formulation, x_e allow to describe any clustering of the vertices of G, since $x_e \in \{0, 1\}$; also, note that Equation 3.24 aligns with the definition provided in Section 2.2.5, in fact $x_{uv} = 1$ implies that u and v should belong to different clusters, while $x_{uv} = 0$ implies that the two vertices should be placed into the same cluster. Furthermore, Equation 3.26 states that for a fixed vertex $u \in V(G)$, the number of variables x_{uv} equal to 0 — for any $v \in V(G)$ such that $u \neq v$, meaning that u and v are adjacent and in the same cluster — must not exceed k. Lastly, Equation 3.25 is the triangle inequality, which ensures that if u and z are placed in the same cluster, and z and v are also placed in the same cluster, then u and v will be placed in the same clustered. This means that belonging to the same cluster is a transitive property , since

$$\begin{cases} x_{uz} = 0 \\ x_{zv} = 0 \\ x_{uv} \leq x_{uz} + x_{zv} \end{cases} \implies x_{uv} = 0$$

3.5.7   The rounding procedure

Since solving ILPs is NP-Hard, Hou et al. [13] relax the problem by changing Equation 3.27 to an interval constraint

$$0 \leq x_e \leq 1$$

leading to an LP program, the solution of which may be fractional. Hence, to obtain a valid clustering, the fractional solutions have to be rounded. Therefore, instead of solving the LP, Hou et al. [13] remove Equation 3.26 from the linear program, and employ the following rounding procedure.

Algorithm 3.3 *Rounding procedure*: given a solution $\{x_e\}_{e \in E(G)}$ of Definition 3.20 (without the size constraint Equation 3.26), a rational value α, and an integer k, the algorithm rounds the solution to integer values.

1: **function** ROUNDINGPROCEDURE(G, $\{x_e\}_{e \in E(G)}$, α, k)
2:     $\mathcal{C} := \varnothing$ ▷ the output set of clusters
3:     $S := V(G)$
4:     **while** $S \neq \varnothing$ **do**
5:         Choose an arbitrary $u \in S$ ▷ this is the *pivot vertex*
6:         $T := \{w \in S - \{u\} \mid x_{uw} \leq \alpha\}$ ▷ u's neighbors under α's threshold
7:         **if** $\sum_{w \in T} x_{uw} \geq \frac{\alpha}{2}|T|$ **then**
8:             $\mathcal{C} = \mathcal{C} \cup \{u\}$ ▷ add a singleton cluster $\{u\}$
9:             $S = S - \{u\}$
10:        **else if** $|T| \leq k$ **then**
11:            $\mathcal{C} = \mathcal{C} \cup (\{u\} \cup T)$ ▷ add the cluster $(\{u\} \cup T)$
12:            $S = S - (\{u\} \cup T)$
13:        **else**
14:            Partition T into $\{T_0', T_1, \ldots, T_p\}$, such that:

                  • $|T_0'| = k$

                  • $|T_i| = k + 1$ for each $0 < i < p$

                  • $|T_p| \leq k + 1$

15:            $T_0 := T_0' \cup \{u\}$ ▷ T_0 has $k + 1$ elements
16:            **for** $i \in [0, p]$ **do**
17:                $\mathcal{C} = \mathcal{C} \cup T_i$ ▷ add each partition as a cluster
18:            **end for**
19:            $S = S - (\{u\} \cup T)$
20:        **end if**
21:     **end while**
22:     **return** $\mathcal{C}$
23: **end function**

---

placeholder. ——————————————————————

placeholder. ——————————————————————

Definition 3.21 (C3)**.** The **C3 algorithm** is defined as follows: first, the next ILP is solved

$$\text{minimize} \sum_{e \in E(G)} (w_e^+ x_e + w_e^-(1 - x_e)), \tag{3.28}$$

$$\text{subject to } x_{uv} \leq x_{uz} + x_{zv}, \ u, v, z \in V(G) \text{ distinct}, \tag{3.29}$$

$$0 \leq x_e \leq 1, \ e \in E(G). \tag{3.30}$$

and then the rounding procedure defined in Algorithm 3.3 is applied.

volevo dare una spiegazione a parole dell'algoritmo ma dopo averlo scritto mi sono reso conto che è abbastanza autoesplicativo, l'unica cosa che è un po vaga è la condizione della riga 7 che non comprendo del tutto

nel materiale supplementare trattano di un valore ottimale per α, lo vedo?

Chapter 4

Discussion

placeholder.

4.1 Dendrix

4.1.1 The deterministic formalization

Vandin et al. [23] provided one of the first mathematical formalizations of the phenomena of mutual exclusivity and coverage, in the context of gene mutations. Specifically, the definitions introduced offer a very intuitive approach to formalizing these biological concepts:

- the coverage of a gene is defined as the set of patients exhibiting a mutation of the gene, equivalent to the number of 1s in its column of the mutation matrix;

- a set of genes is defined to be mutually exclusive if no patient has more than one mutated gene in the set, i.e. no row of the set's associated matrix has more than 1 one;

- the coverage of a gene set is the set of patients with at least one mutation in the set;

- the coverage overlap of a gene set is the count of patients who possess more than one mutation within the gene set;

- the weight of a gene set is calculated as the difference between the coverage of the gene set and its coverage overlap.

Consequently, a higher weight for a gene set indicates both greater coverage and mutual exclusivity among its genes. The weight formula suggests that the optimal gene set, i.e. the one that maximazes its weight, is the one where the associated matrix has a high number of rows with at least one 1, and a minimal number of rows with more than one 1.

In my view, this metric stands out as the most elegant among those discussed in this work: it not only provides a clear and intuitive measure, but also offers a simple and straightforward formula. While this formula may seem to oversimplify the challenge of identifying driver pathways — given that mutual exclusivity alone does not cover all aspects of pathway analysis, and exact mutual exclusivity is rarely observed in real data — it remains a highly regarded deterministic formalization, and numerous studies (some of which are discussed in this work) agree that this metric represents the most refined approach to date.

4.1.2   Additional considerations

I want to commend Vandin et al. [23] for their precise and methodical explanation of their methodology. With only a few minor, negligible details to consider, their work is exceptionally clear regarding their objectives, the methods employed to achieve them, and their actual outcomes. Additionally, their mathematical analysis is thorough and well-supported: in the supplemental material, they provide extensive proofs, including the NP-hardness of the MWSP (which is described in Theorem 2.1), the correctness of their greedy algorithm, and the rapid mixing property of their MCMC approach. I greatly appreciate the clarity of their presentation, which significantly facilitated my understanding of their study. Indeed, this level of clarity is notably superior compared to other papers I have analyzed, which lack such clear explanations, as will be discussed in the subsequent sections.

As a final note, I believe that conducting a comparative analysis between the MCMC approach and a random search method would be interesting. For instance, consider the following algorithm:

1. *initialization*: given the set of all genes $\mathcal{G}$, choose an arbitrary subset $M_0 \subseteq \mathcal{G}$ ok k genes;

2. *iteration*: for $t = 1, 2, \ldots$ derive M_{t+1} from M_t as follows:

   (a) define $W \subseteq \mathcal{G}$ and $V \subseteq M_t$ randomly;
   (b) choose $(\hat{w}, \hat{v}) \in \underset{(w,v) \in W \times V}{\arg \max} \, W\left((M_t - \{v\}) \cup \{w\}\right)$;
   (c) set $M_{t+1} := (M_t - \{\hat{v}\}) \cup \{\hat{w}\}$.

At each step, this algorithm selects a predetermined amount of *random adjustments*, choosing the one that maximizes the weight as the base set for the next iteration. It would be interesting to evaluate how this approach performs on real data, and whether the MCMC algorithm outperforms it, particularly when applied to data under the GIM model.

4.2   Multi-Dendrix

4.2.1   An ILP for the MWSP

Acknowledgements

TODO

Bibliography

[1] Özgün Babur 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] Tiziano Bernasocchi et al. "Dual functions of SPOP and ERG dictate androgen therapy responses in prostate cancer". In: *Nature Communications* 12.1 (Feb. 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-020-20820-x. URL: http://dx.doi.org/10.1038/s41467-020-20820-x.

[3] *Cancer of any site - Cancer Stat Facts*. URL: https://seer.cancer.gov/statfacts/html/all.html.

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

[5] Jaroslaw Cisowski et al. "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.

[6] "Comprehensive genomic characterization defines human glioblastoma genes and core pathways". In: *Nature* 455.7216 (Sept. 2008), 1061–1068. ISSN: 1476-4687. DOI: 10.1038/nature07385. URL: http://dx.doi.org/10.1038/nature07385.

[7] Geoffrey M Cooper. *The development and causes of cancer*. 2000. URL: https://www.ncbi.nlm.nih.gov/books/NBK9963/.

[8] Yulan 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.

[9] Máire A. Duggan et al. "The Surveillance, Epidemiology, and End Results (SEER) Program and Pathology: Toward Strengthening the Critical Relationship". In: *American Journal of Surgical Pathology* 40.12 (Dec. 2016), e94–e102. ISSN: 0147-5185. DOI: 10.1097/pas.0000000000000749. URL: http://dx.doi.org/10.1097/PAS.0000000000000749.

[10] P. EHRLICH. "Experimental Researches on Specific Therapy". In: *The Collected Papers of Paul Ehrlich*. Elsevier, 1960, 106–117. ISBN: 9780080090566. DOI: 10.1016/b978-0-08-009056-6.50015-4. URL: http://dx.doi.org/10.1016/b978-0-08-009056-6.50015-4.

[11] P. Andrew Futreal et al. "A census of human cancer genes". In: *Nature Reviews Cancer* 4.3 (Mar. 2004), 177–183. ISSN: 1474-1768. DOI: 10.1038/nrc1299. URL: http://dx.doi.org/10.1038/nrc1299.

[12] Dorit Hochbaum. *Approximation algorithms for NP-hard problems ed. by Dorit S. Hochbaum*. PWS Publ, 1996.

[13] Jack P. Hou et al. "A new correlation clustering method for cancer mutation analysis". In: *Bioinformatics* 32.24 (Aug. 2016), 3717–3728. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw546. URL: http://dx.doi.org/10.1093/bioinformatics/btw546.

[14] "International network of cancer genome projects". In: *Nature* 464.7291 (Apr. 2010), 993–998. ISSN: 1476-4687. DOI: 10.1038/nature08987. URL: http://dx.doi.org/10.1038/nature08987.

[15] M. Kanehisa. "KEGG: Kyoto Encyclopedia of Genes and Genomes". In: *Nucleic Acids Research* 28.1 (Jan. 2000), 27–30. ISSN: 1362-4962. DOI: 10.1093/nar/28.1.27. URL: http://dx.doi.org/10.1093/nar/28.1.27.

[16] Sushant Kumar et al. "Passenger Mutations in More Than 2, 500 Cancer Genomes: Overall Molecular Functional Impact and Consequences". In: *Cell* 180.5 (Mar. 2020), 915–927.e16. ISSN: 0092-8674. DOI: 10.1016/j.cell.2020.01.032. URL: http://dx.doi.org/10.1016/j.cell.2020.01.032.

[17] Mark D. M. Leiserson 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.

[18] *Side effects of cancer treatment*. URL: https://www.cancer.gov/about-cancer/treatment/side-effects.

[19] Rebecca L. Siegel et al. "Cancer statistics, 2024". In: *CA: A Cancer Journal for Clinicians* 74.1 (Jan. 2024), 12–49. ISSN: 1542-4863. DOI: 10.3322/caac.21820. URL: http://dx.doi.org/10.3322/caac.21820.

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

[21] *The genetics of cancer*. Aug. 2024. URL: https://www.cancer.gov/about-cancer/causes-prevention/genetics.

[22] Roman K Thomas et al. "High-throughput oncogene mutation profiling in human cancer". In: *Nature Genetics* 39.3 (Feb. 2007), 347–351. ISSN: 1546-1718. DOI: 10.1038/ng1975. URL: http://dx.doi.org/10.1038/ng1975.

[23] Fabio Vandin et al. "De novo discovery of mutated driver pathways in cancer". In: *Genome Research* 22.2 (June 2011), 375–385. ISSN: 1088-9051. DOI: 10.1101/gr.120477.111. URL: http://dx.doi.org/10.1101/gr.120477.111.

[24] Bert Vogelstein et al. "Cancer Genome Landscapes". In: *Science* 339.6127 (Mar. 2013), 1546–1558. ISSN: 1095-9203. DOI: 10.1126/science.1235122. URL: http://dx.doi.org/10.1126/science.1235122.

[25] Michael R. Waarts et al. "Targeting mutations in cancer". In: *Journal of Clinical Investigation* 132.8 (Apr. 2022). ISSN: 1558-8238. DOI: 10.1172/jci154943. URL: http://dx.doi.org/10.1172/JCI154943.

[26] Nhs Website. *Signs and symptoms.* June 2024. URL: https://www.nhs.uk/conditions/cancer/symptoms/.

[27] *What is cancer?* Oct. 2021. URL: https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

[28] Christian Widakowich et al. "Review: Side Effects of Approved Molecular Targeted Therapies in Solid Cancers". In: *The Oncologist* 12.12 (Dec. 2007), 1443–1455. ISSN: 1549-490X. DOI: 10.1634/theoncologist.12-12-1443. URL: http://dx.doi.org/10.1634/theoncologist.12-12-1443.

[29] Junfei Zhao 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.