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 TODO

placeholder.

introduzione sul cancro

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. Nonetheless, 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 [6].

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 [6].

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 [3]:

- *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 [17, 19].

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 [20].

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 [19].

The concept of targeted therapy originates from the german Nobel Prize Paul Ehrlich's idea of a "*magic bullet*" [8], 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 [20].

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 [17].

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 [17].

- *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 [17].

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 [17].

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 [16], and some patients may be highly sensitive to these drugs and may develop specific and severe toxicities [20].

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 [17].

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 [19]:

- 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 [19].

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 [19]. 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

introduction to the chapter

2.1 Mutations and pathways

2.1.1 Passenger and driver mutations

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

DO I ADD THIS AS A CITATION???

expand this section

2.1.2 Importance of pathways

placeholder

questa sezione fa riscritta perché la roba che sta scritta qui non mi piace

Cell signaling is the process by which cells interact with each other, themselves, or their environment. It concerns the transduction of signals, which can be chemical, or can involve various types such as pressure, temperature, or light signals [5]. **Pathways** are sequences of molecular interactions within a cell that lead to a change in the cell or the production of a specific product [15]. These pathways have a direction in which the actions occur, with the terms *upstream* and *downstream* indicating the initial and final stages of these processes, respectively.

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

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

2.2 Classifying mutations

2.2.1 Frequency

To classify mutations into the two categories described, assessing their biological function is essential, though this remains a challenging task. Numerous methods exist to predict the functional impact of mutations based on *a priori* knowledge. However, these approaches often fail to integrate information effectively across various mutation types and are limited by their reliance on known proteins, rendering them less effective for less-studied ones [14].

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 [14]. Indeed, key driver events, such as TP53 loss-of-function mutations, can be identified by their significantly high frequency of occurrence across a set of tumors [1]. However, in many cases, since driver mutations are predominantly located in genes that are part of cell signaling pathways, different patients may harbor mutations in different pathway loci. Indeed, driver mutations can vary extensively between patient samples, even within the same cancer type [14]; additionally, there is minimal overlap of mutated genes across sample pairs, even from the same patient [21], reducing the statistical power of frequency analyses.

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

Therefore, studies should be conducted at the pathway level, as it is well established that different mutations can affect the same pathway across multiple samples [14]. However, since each pathway involves multiple genes, numerous possible combinations of driver mutations could impact a crucial cancer pathway, making it computationally unfeasible to test every possible gene permutation [18] — estimates suggest that the human genome contains more than 50,000 genes [9]. Hence, it is necessary to identify a property to leverage to search efficiently.

2.2.2 Mutual exclusivity and coverage

Most techniques developed in recent years for recognizing driver mutations leverage a statistical property observed in cancer patient data: each patient typically has a relatively small number of mutations that affect multiple pathways, thus each pathway will contain *1 driver mutation on average* per sample. This concept of mutual exclusivity among driver mutations within the same pathway, as statistically observed in patient samples, is then axiomatized and employed by research algorithms designed to identify driver mutations [14]. Additionally, mutual exclusivity *does not affect different pathways*; it is a phenomenon that occurs exclusively within a single pathway. While the precise explanation for this occurrence is not yet fully understood, several hypotheses appear plausible [7, 4, 18]:

- 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 (WT) 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].

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

sai che cosa aggiungere, aggiungilo appena puoi

Thus, *a driver pathway consists of genes that are mutated in numerous patients, with mutations being approximately mutually exclusive*. It is also observed that pathways exhibiting these characteristics are generally shorter and comprised of fewer genes on average [14].

2.3 Mutual exclusivity formalization

2.3.1 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 formulations, based on different assumptions, which will be discussed in later sections.

2.3.2 A deterministic formalization of mutual exclusivity

Vandin et al. [18], the authors of a landmark paper that developed an algorithm called Dendrix, provided the following mathematical formalization for the properties of mutual exclusivity and coverage for a set of genes.

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. 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

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

Under the previous definitions of mutual exclusivity, M is **mutually exclusive** if no patient has more than one mutated gene, formally

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

Definition 2.3 (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 in M is mutated, and it is defined as follows

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

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

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

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

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

To evaluate these two attributes, a measure that quantifies the trade-off between coverage and mutual exclusivity is introduced.

Definition 2.4 (Coverage overlap)**.** Given a set M of genes, the **coverage overlap** is defined as follows:

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

Note that the sum in Definition 2.4 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.* Note that $\omega(M) \geq 0$, with equality holding only if the sum equals the coverage of M, which means that no patient has more than one mutated gene of M.

Definition 2.5 (Mutually exclusive set)**.** A gene set M is considered to be **mutually exclusive** if $\omega(M) = 0$.

Definition 2.6 (Weight of gene set)**.** Given a set of genes M, to take into account both coverage and coverage overlap, the following measure is introduced:

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

Note that $W(M) = \Gamma(M)$ when M is mutually exclusive.

In order to find an optimal gene set, 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), and it can be proven that it is NP-Hard.

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

Proof. The proof is by reduction from the Independent Set Problem (ISP), which is known to be NP-Hard [11]. 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 $\mathcal{S}_v := \left\{ s_v^{(1)}, \dots, s_v^{(\Delta - \deg(v))} \right\}$ 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	$\dots$	v_n
e_1	$\ddots$		
$\vdots$		$\ddots$	
e_m			$\ddots$
$s_{v_1}^{(1)}$	$\ddots$		
$\vdots$		$\vdots$	
$s_{v_1}^{(\Delta - \deg(v_1))}$			
$\vdots$		$\vdots$	
$s_{v_n}^{(1)}$			
$\vdots$		$\vdots$	
$s_{v_n}^{(\Delta - \deg(v_n))}$			$\ddots$

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:

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 2.6, it follows that the maximum value $W(M)$ can reach is
$$\begin{aligned} W(M) &= 2\,|\Gamma(M)| - \sum_{i=1}^{k} |\Gamma(v_i)| \\ &= 2k\Delta - k\Delta \\ &= k\Delta \end{aligned}$$
- 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. □

2.3.3 Extending the deterministic equation

considera di spostare questo alla fine

Leiserson et al. [14] refine the weight function of Vandin et al. [18], aiming to extend the metric to assess mutual exclusivity across multiple driver pathways. In particular, while identifying individual driver pathways is crucial, most cancer patients are likely to have driver mutations across multiple pathways.

To effectively identify multiple driver 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 to find a collection $M = \{M_1, \ldots, M_t\}$ of gene sets which maximizes the sum of individual weights, i.e. $\sum_{\rho=1}^{t} W(M_\rho)$.

2.3.4 A statistical approach

Babur et al. [1] criticize the metric developed by Vandin et al. [18] 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 $\Gamma(M - \{j\})$. 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, Babur et al. [1] employ the following permutation-based correction:

- first,

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

2.3.5 A clustering approach

change this title soo similar with the previous one

Another notable approach utilized in several papers involves constructing gene graphs and identifying clusters based on specific criteria; this method is demonstrated by Hou et al. [12].

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*;
- a **negative weight** w^-_{uv}, which represents *the cost of placing* u *and* v *in the same cluster*;

i.e. by making w^+_{uv} large, placing u and v in different clusters is discouraged, and viceversa; the same concept applies for w^-_{uv}. Indeed, as weights representations suggest, *genes in the same cluster are likely to be mutually exclusive.*

Weights are calculated using four types of datasets: gene mutation data, copy number variation (CNV), network information, and gene expression data. To appropriately combine the sources from which the information was obtained, linear combinations were utilized to account for the reliability of the sources from which the data was drawn.

Additionally, as each type of data contributes differently to the driver discovery process, linear combinations are used, based on the importance or accuracy of each, specifically:

- 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*.

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. Therefore, if $m_{i,j} = 0$, no mutation of gene j is recorded in sample i, otherwise gene j is *deemed mutated.*

menziono i valori che hanno usato loro? menziono le conseguenze che valori alti/bassi di queste soglie hanno, secondo loro?

Definition 2.7 (Coverage of a vertex)**.** Given a vertex $u \in V(G)$, i.e. a gene, the **coverage of** u

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

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.

Definition 2.8 (Mutual exclusivity component)**.** The **mutual exclusivity component** between two genes $u, v \in V(G)$ is defined as follows:

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

where a is a user-defined scaling parameter.

This ratio is often referred to as **IoM** (*Intersection over Minimum*), 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

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

Definition 2.9 (Negative weights)**. Negative weights** only depend on the mutual exclusivity component, i.e.

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

By contrast, positive weights can depend on multiple factors which will be presented in the following sections . Focusing on **coverage**, if two genes u and v increase the coverage of the set significantly, $w^{+}_{uv}(\mathrm{c})$ should be large such that they are encouraged to be placed in the same cluster. Let

va bene se la metto su questo piano?

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

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

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

Note that, similar to Equation 2.3

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

The linear combinations that define w^{+}_{uv} will be discussed afterward.

Chapter 3

Finding driver mutations

Although the true explanation for mutual exclusivity remains unknown, and its therapeutic potential is still uncertain, this phenomenon is frequently observed in data and may lead to discoveries in cancer treatment. Existing approaches can be categorized into two types: ***de novo*** approaches, which identify mutually exclusive patterns using only genomic data from patients, and ***knowledge-based*** methods, which integrate the analysis with external *a priori* information [7]. *De novo* approaches might lack sufficient information as they do not utilize existing databases . Conversely, given that our understanding of gene and protein interactions in humans is still incomplete and many pathway databases fail to accurately represent the specific pathways and interactions present in cancer cells, *knowledge-based* approaches may be limited by their dependence on existing data sources. Consequently, *de novo* methods might yield new but potentially less accurate results, while *knowledge-based* approaches may limit the discovery of novel biological insights [14].

correggi questa frase che non ha senso effettivamente

3.1 Dendrix

3.1.1 A greedy approach

To find a solution to the problem described in Section 2.3.2, Vandin et al. [18] developed the following greedy algorithm called Dendrix (*de novo* [7]).

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 $W(M)$.

```
function GREEDYDENDRIX($\mathcal{G}$, $k$)
    $M := \{g_1, g_2\}$ such that $M$ maximizes $W(M)$        ▷ pick the best gene pair
    for $i \in [3, k]$ do
        $\hat{g} := \arg\max_{g \in \mathcal{G}} W(M \cup \{g\})$        ▷ TODO A PARIMERITO???
        $M = M \cup \{\hat{g}\}$
    end for
    return $M$
end function
```

The time complexity of the algorithm is $O\left(n^2 + kn\right) = O\left(n^2\right)$.

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 $\left|\hat{M}\right| = k$; 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;

WHAT ARE THESE??

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

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

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;
- 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 $W(\hat{M})$.

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. [18] prove that Algorithm 3.1 can correctly identify $\hat{M}$ with high probability when the mutation data come from the GIM generative model.

lo dimostrano nel materiale supplementare, lo vedo/metto?

Moreover, note that under the GIM, genes in $\hat{M}$ may have observed mutation frequencies similar to those of genes not in $\hat{M}$, making it impossible to distinguish between them based solely on mutation frequency, regardless of the number of patients. In practical applications, the utility of this greedy algorithm depends on the availability of mutation data from a sufficient number of patients.

3.1.2 Using MCMC

placeholder

la faccio? probabilmente si ma dovrei prima vedere cosa le MCMC siano

3.2 Multi-Dendrix

3.2.1 An alternative solution to Dendrix

Leiserson et al. [14], the authors of Multi-Dendrix (*de novo* [7]), try to solve the same problem posed by Vandin et al. [18], described in Section 2.3.2, by formulating the problem as an *Integer Linear Program* (ILP), which they refer to as Dendrix$_{ILP}(k)$.

Consider a gene set M defined by a set of indicator variables, one for each gene $j \in M$, 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

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

Definition 3.2 (Dendrix$_{ILP}(k)$)**.** 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 the sum in Equation 3.3 is the second version of the definition provided in Definition 2.6.

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 $j \in M$. □

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.

placeholder.

non mi ricordo cosa volevo scrivere qua

3.2.2 The ILP

As outlined in Section 2.3.3, Leiserson et al. [14] propose that the most effective approach to conducting the research is to identify a collection of gene sets that maximize the sum of their individual weights. To achieve this result, they solve the following problem, which is an extension of Section 2.3.2:

> **Multiple Maximum Weight Submatrices Problem**: Given an $m \times n$ mutation matrix A, and integer $t > 0$, and two integers $k_{\min}, k_{\max} \geq 0$, find a collection $M = \{M_1, \dots, M_t\}$ of column submatrices that maximizes
>
> $$W'(M) := \sum_{\rho=1}^{t} W(M_\rho) \tag{3.6}$$
>
> 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 this problem is NP-Hard, as for the case $t = 1$. Furthermore, collections M with a large value of $W'(M)$ are also likely to have higher coverage $\Gamma(M_\rho)$ for each individual gene set ρ. As a result, optimal solutions tend to produce collections where many patients have mutations in more than one gene set, or they may be pairs or larger groups of co-occurring mutations, a phenomenon observed in cancer.

qua inseriscono una citazione, potrebbe valere la pena di inserirla e/o indagare?

Definition 3.3 (Multi-Dendrix)**.** Multi-Dendrix is defined by the following ILP:

$$\text{maximize} \sum_{\rho=1}^{t} \sum_{i=1}^{m} \left(2 \cdot C_i(M_\rho) - \sum_{j=1}^{n} I_{M_\rho}(j) \cdot a_{i,j} \right), \tag{3.7}$$

$$\text{subject to} \sum_{j=1}^{n} I_{M_\rho}(j) \cdot a_{i,j} \geq C_i(M_\rho), \tag{3.8}$$

$$k_{\min} \leq \sum_{j=1}^{n} I_{M_\rho}(j) \leq k_{\max}, \tag{3.9}$$

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

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

Note that:

- Equation 3.7 and Equation 3.8 expand Equation 3.3 and Equation 3.4 respectively;
- Equation 3.9 allows each gene group to have a size between k_{min} and k_{max};
- Equation 3.10 states that each gene can appear in at most one set within the collection.

Leiserson et al. [14] 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 $W'(M)$.

```
function iterDendrix($\mathcal{G}$, $k$, $t$)
    $M := \varnothing$
    for $i \in [1, t]$ do
        $M_i :=$ greedyDendrix($\mathcal{G}, k$)        ▷ procedure defined in Algorithm 3.1
        $M = M \cup \{M_i\}$
        $\mathcal{G} = \mathcal{G} - M_i$
    end for
    return $M$
end function
```

Denoting with M and I the collections of gene sets obtained from Multi-Dendrix and Iter-Dendrix respectively, Leiserson et al. [14] 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 $k_{\text{min}} < k_{\text{max}}$ Multi-Dendrix may choose gene sets with fewer than k_{max} genes if doing so maximizes the overall weight $W'(M)$.

Leiserson et al. [14] state that all of these scenarios occur when analyzing real mutation data.

Lastly, Multi-Dendrix can be extended to allow gene sets to overlap, since the genes in the intersection may be involved in multiple biological processes. Hence, Equation 3.10 is replaced with the following equation:

$$\sum_{\rho=1}^{t} I_{M_\rho}(j) \leq \Delta, \quad 1 \leq j \leq n \tag{3.11}$$

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

$$\sum_{j=1}^{n} \sum_{\rho=1}^{t} \sum_{\substack{\rho'=1 \\ \rho \neq \rho'}}^{t} I_{M_\rho}(j) \cdot I_{M_{\rho'}}(j) \leq \tau, \quad 1 \leq \rho \leq t \tag{3.12}$$

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

3.3 MDPFinder

3.3.1 The genetic algorithm

placeholder.

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

The approach used by Zhao et al. [21] 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.

Definition 3.4 (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 3.5 (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 3.6 (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 3.6.

Definition 3.7 (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 3.8 (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.

fixa questa definizione

Definition 3.9 (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 3.10 (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 3.11 (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**:

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?

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

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

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 $R(E_M)$

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* [7]), which is applied to a directed graph constructed from databases containing information about biological pathways .

menziono i database che hanno usato?

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 .

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

placeholder.

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 parole, provo a farlo?

Definition 3.12 (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.

QUESTA DEFINIZIONE NON VA QUA

3.5 C3

3.5.1 Multiple versions

Hou et al. [12] define three methods for assigning weights to the graph to perform the vertex clustering algorithm called C3 (*knowledge-based* [7]):

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 Definition 2.9, 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 3.13 (ME-CO)**.** In the **ME-CO** version of the algorithm, the following definitions apply:

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

$$\forall u, v \in V(G) \quad w^+_{uv} := w^+_{uv}(\mathrm{c}) \tag{3.14}$$

The definitions for $w^-_{uv}(\mathrm{e})$ and $w^+_{uv}(\mathrm{c})$ are provided in Definition 2.8 and Definition 2.10 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. [12] 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) [10], 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.

non ho inserito l'immagine ma volendo la inserisco, la metto?

placeholder.

per la foto eventualmente

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 [13], 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

$$f(u,v) := \frac{|\mathscr{N}'(u) \cap \mathscr{N}'(v)|}{|\mathscr{N}'(u) \cup \mathscr{N}'(v)|} \tag{3.15}$$

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

$$\mathscr{F} := \{f(u,v) \mid u,v \in V(G')\} \tag{3.16}$$

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

Definition 3.14 (Network information component)**.** The **network information**

component is defined as follows:

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

Definition 3.15 (NI-ME-CO)**.** The **NI-ME-CO** version of the algorithm is defined by the following equations:

$$\forall u, v \in V(G) \quad w^-_{uv} := w^-_{uv}(\mathrm{e}) \tag{3.17}$$

$$\forall u, v \in V(G) \quad w^+_{uv} := w_1 w^+_{uv}(\mathrm{c}) + w_2 w^+_{uv}(\mathrm{n}) \tag{3.18}$$

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

parlare del rescaling?

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 $\mathbf{z}(\mathrm{u})$ be the vector of the time-evolving expression values of u . Thus, let

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

$$g(u,v) := \frac{|\langle \mathbf{z}(\mathrm{u}), \mathbf{z}(\mathrm{v}) \rangle|}{||\mathbf{z}(\mathrm{u})||\,||\mathbf{z}(\mathrm{v})||} \tag{3.19}$$

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 $\mathbf{z}(\mathrm{u})$ and $\mathbf{z}(\mathrm{v})$ — 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

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

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

Moreover, let

$$\mathscr{G} := \{g(u,v) \mid u, v \in V(G)\} \tag{3.20}$$

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

Definition 3.16 (Expression data component)**.** The **expression data component** is defined as follows:

$$w^+_{uv}(\mathrm{x}) := \begin{cases} 1 & g(u,v) > T''(J'') \\ \dfrac{g(u,v)}{T''(J'')} & g(u,v) \leq T''(J'') \end{cases}$$

Definition 3.17 (EX-ME-CO)**.** The **EX-ME-CO** version of the algorithm is defined by the following equations:

$$\forall u, v \in V(G) \quad w^{-}_{uv} := w^{-}_{uv}(\mathrm{e}) \tag{3.21}$$

$$\forall w^{+}_{uv} := w_1 w^{+}_{uv}(\mathrm{c}) + w_2 w^{+}_{uv}(\mathrm{x}) \tag{3.22}$$

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

placeholder.

parlare del rescaling?

3.5.5 Other versions

Hou et al. [12] 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 3.18 (NI-EX-ME-CO)**.** The **NI-EX-ME-CO** version of the algorithm is defined by the following equations:

$$\forall u, v \in V(G) \quad w^{-}_{uv} := w^{-}_{uv}(\mathrm{e}) \tag{3.23}$$

$$\forall w^{+}_{uv} := w_1 w^{+}_{uv}(\mathrm{c}) + w_2 w^{+}_{uv}(\mathrm{n}) + w_3 w^{+}_{uv}(\mathrm{x}) \tag{3.24}$$

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. [12] 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 3.19 (C3's ILP)**.** The **C3 algorithm** can be defined by the following ILP:

$$\text{minimize} \sum_{e \in E(G)} (w_e^+ x_e + w_e^- (1 - x_e)), \tag{3.25}$$

$$\text{subject to } x_{uv} \leq x_{uz} + x_{zv}, \;\; u, v, z \in V(G) \text{ distinct}, \tag{3.26}$$

$$\sum_{\substack{v \in V(G) \\ u \neq v}} (1 - x_{uv}) \leq k, \;\; u \in V(G), \tag{3.27}$$

$$x_e \in \{0, 1\}, \;\; e \in E(G). \tag{3.28}$$

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 3.25 aligns with the definition provided in Section 2.3.5, 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 3.27 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 3.26 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

dopo aver scritto questo mi sono fatto una domanda alla quale non trovo risposta nel paper: i cluster si possono intersecare? io non direi altrimenti questa definizione sottoforma di ILP non credo potrebbe avere senso, giusto?

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

3.5.7 The rounding procedure

Since solving ILPs is NP-Hard, Hou et al. [12] relax the problem by changing Equation 3.28 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. [12] remove Equation 3.27 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 3.19 (without the size constraint Equation 3.27), a rational value α, and an integer k, the algorithm rounds the solution to integer values.

```
 1: function RoundingProcedure($G$, $\{x_e\}_{e \in E(G)}$, $\alpha$, $k$)
 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 $\alpha$'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 $\{T'_0, T_1, \dots, T_p\}$, 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.

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?

Definition 3.20 (C3)**.** The **C3 algorithm** is defined as follows: first, the next ILP is solved

$$\text{minimize} \sum_{e \in E(G)} (w_e^+ x_e + w_e^- (1 - x_e)), \tag{3.29}$$

$$\text{subject to } x_{uv} \leq x_{uz} + x_{zv}, \;\; u, v, z \in V(G) \text{ distinct}, \tag{3.30}$$

$$0 \leq x_e \leq 1, \;\; e \in E(G). \tag{3.31}$$

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

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

[4] 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.

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

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

[7] 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.

[8] 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.

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

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

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.