SAPIENZA
UNIVERSITÀ DI ROMA

TODO Titolo

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

Alessio Bandiera
ID number 1985878

Advisor
Prof. Ivano Salvo

Academic Year 2023/2024

TODO Titolo
Bachelor’s Thesis. Sapienza University of Rome

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

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

TODO. test

Contents

Chapter 1

Introduction

MISSING INTRODUCTION ON THE WHOLE PAPER

introduction of the whole paper, putting the "todo" thing to remember later

1.1 Cancer

1.1.1 Carcinogenesis

Cancer is a medical condition characterized by uncontrolled cell proliferation, which allows cells to infiltrate into organs and tissues, thereby altering their functions and structure. This exponential growth is driven by mutations in cellular DNA, which encodes the instructions for cell development and multiplication, therefore errors in these instructions can lead to cancerous transformation. In most types of cancer, a single aberration is insufficient for cancer development; instead, multiple mutations are required. Some of these mutations are present since birth, while others occur throughout life due to chance or external factors. Additionally, for tumor proliferation to occur, mutations in genes that regulate cell growth are necessary [16]. Specifically, proto-oncogenes, which promote mitosis, and tumor suppressor genes, which inhibit cell growth, are involved in this process, known as *oncoevolution* [4].

expand on carcinogenesis

1.1.2 Current treatment

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

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

- **chemotherapy**, which employs cytotoxic drugs to block cellular division in both cancerous and healthy cells, but they can also induce side effects in rapidly renewing tissues.
- **hormone therapy**, which alters the balance of specific hormones, potentially leading to side effects such as joint pain or osteoporosis;
- **targeted therapy**, which involves drugs containing antibodies or inhibitory substances that specifically target cancer cells, promoting their destruction by the immune system; however, developing effective targeted therapies can be challenging due to the complexity of the target's structure or function; in addition, this approach may also induce unwanted side effects in various organs, and cancer cells may develop resistance if they find alternative ways to develop that do not rely on the therapy's target [14] .

check this out, and probably move this part in the next section

1.1.3 Target therapy

In particular, in recent years targeted therapy has been the focus of extensive research due to its potential to precisely affect only the desired target, thereby reducing the side effects that currently characterize most cancer treatments and potentially limiting damage to healthy cells [13].

expand target therapy on how it works; check README for link; consider making "Target therapy" section and adding subsections if needed

Chapter 2

Classifying mutations

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 [12]. These pathways have a direction in which the actions occur, with the terms *upstream* and *downstream* indicating the initial and final stages of these processes, respectively.

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

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

2.1 Mutations

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???

try to expand this section

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 [11].

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 [11]. 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 [11]; additionally, there is minimal overlap of mutated genes across sample pairs, even from the same patient [17], 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 [11]. 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 [15] — estimates suggest that the human genome contains more than 50,000 genes [7]. Hence, it is necessary to identify a property to leverage to conduct a 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 [11]. 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 [6, 3, 15]:

- one hypothesis is that mutually exclusive genes are functionally connected within a common pathway, acting on the same downstream effectors and creating functional redundancy; consequently, they would share the same selective advantage, meaning that the alteration of one mutually exclusive gene would be sufficient to disrupt their shared pathway, thereby removing the selective pressure to alter the others; this explanation, however, does not fully account for the phenomenon because the co-alteration of mutually exclusive genes should not result in negative effects on the cell.
- an alternative explanation is that the co-occurrence of mutually exclusive alterations is detrimental to cancer survival, leading to the elimination of cells that harbor such co-occurrences; moreover, some pairs of mutually exclusive genes could be *synthetic lethal*, meaning that while the alteration of one gene may be compatible with cell survival, the simultaneous aberration of both genes would be lethal to the cell .

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

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

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 [11].

2.3 Mutual exclusivity formalization

2.3.1 Hard and soft mutual exclusivity

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

2.3.2 Mutual exclusivity of a group

Searching for the most mutually exclusive gene group is equivalent to identifying a single driver pathway, for the aforementioned reasons. For a pair of genes, soft mutual exclusivity can be assessed using the Fisher's exact test . However, there is no agreed-upon method for analytically testing mutual exclusivity among more than two genes. One approach could involve checking whether each pair of genes within

studiati cos'è un minimo

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].

2.3.3 A deterministic equation

Vandin et al. [15], the authors of a landmark paper that developed two 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**: 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.

nei materiali supplementari mettono la dimostrazione che questo problema è NP-Hard, lo devo fare?

2.3.4 Extending the deterministic equation

considera di spostare questo alla fine

Leiserson et al. [11] refine the weight function of Vandin et al. [15], 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. $\sum_{\rho=1}^{t} W(M_\rho)$.

2.3.5 A statistical approach

Babur et al. [1] criticize the metric developed by Vandin et al. [15] 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.

QUESTO paper fa vedere in dettaglio come si fa, sono sicuro al 99% che si tratti della stessa cosa, lo inserisco?

specificare quali di preciso? non mi sembra rilevante

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 $i \in M$, mutations in $\Gamma(i)$ are independent of alterations in $\Gamma(M - \{i\})$. H_0 is then tested for each $i \in M$ 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. Using Dendrix's notation

i'm not sure i know what this means

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

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 . To account for it, first the null distribution of the initial p-values must be estimated for each gene, then it must be calculated the significance of the observed initial p-values for each member

add link?

non ho la minima idea di cosa voglia dire tutta questa frase; successivamente, qui viene spiegato in che modo stimano la "null distribution of the initial p-value", ma oltre che non riesco a capire che cosa facciano di preciso, per farlo riciclano l'algoritmo con il quale poi andranno a risolvere il problema generale, ma

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

2.3.6 A clustering approach

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. [9].

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:

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

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

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.*

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.

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

va bene se la metto su questo piano?

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

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

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 [6]. *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 [11].

correggi questa frase che non ha senso effettivamente

3.1 Dendrix

fare Dendrix

placeholder.

3.2 Multi-Dendrix

3.2.1 An alternative solution to Dendrix

Leiserson et al. [11], the authors of Multi-Dendrix (*de novo* [6]), try to solve the same problem posed by Vandin et al. [15], provided in Section 2.3.3.

qui manca una sezione iniziale in cui multi-dendrix descrive i motivi per cui l'approccio greedy di dendrix potrebbe non trovare soluzioni ottimali, e perché il loro approccio che trova le soluzioni tutte in un colpo solo è migliore

For these reasons, Leiserson et al. [11] formulate Dendrix's problem as an **ILP** (*Integer Linear Program*), which they refer to as $\text{Dendrix}_{ILP}(k)$. 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 3.1 (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 2.3.4, Leiserson et al. [11] proposes 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 2.3.3:

> **Multiple Maximum Weight Submatrices Problem**: Given an $m \times n$ mutation matrix A, and integer $t > 0$, find a collection $M = \{M_1, \ldots, M_t\}$ of $m \times k$ column submatrices that maximizes

k DOES NOT APPEAR ANYWHERE IN THE ILP???????

$$W'(M) := \sum_{\rho=1}^{t} W(M_\rho) \tag{3.6}$$

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 3.2 (Multi-Dendrix)**.** Multi-dendrix is defined by the following ILP:

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

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

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

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

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

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; when $k_{\min} < k_{\max}$ the ILP may choose gene sets with fewer than $k_{\max}$ genes if this maximizes the overall weight $W'(M)$ of the collection.

placeholder.

"tutti i casi precedenti avvengono davvero nella realtà" non l'ho ancora scritto perché va prima deciso cosa fare con iter-dendrix perché qui è menzionato e sarebbe da inserire nei bullet point

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. [17] involves a **GA** (*Genetic Algorithm*), which is versatile and flexible, and can be used to optimize a wide variety of scoring functions. 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 3.3 (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

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

Definition 3.4 (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 :

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

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

Definition 3.5 (Selection probability)**.** Given the rank r_i of an individual h_i based on its score, the **selection probability** is definied 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 3.5.

Definition 3.6 (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 3.7 (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.

Definition 3.8 (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 3.9 (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 most 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 3.10 (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 $\mathrm{pcc}(\cdot)$ 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 greedy approach

To identify the most mutually exclusive group, Babur et al. [1] employ a greedy algorithm called Mutex (*knowledge-based* [6]), 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?

placeholder.

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 parole, provo a farlo?

Definition 3.11 (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. [9] define three methods for assigning weights to the graph to perform the vertex clustering algorithm called C3 (*knowledge-based* [6]):

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 ME-CO

Definition 3.12 (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 NI-ME-CO

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. [9] 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) [8], 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 [10], 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; 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 $\mathscr{F}$. 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.

Definition 3.13 (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 3.14 (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 EX-ME-CO

todo

placeholder.

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

[3] 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.

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

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

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

[15] 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.

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

[17] 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.