

# A Comparative Analysis of Algorithms for Identifying Cancer Driver Pathways

Facoltà di Ingegneria dell'informazione, informatica e statistica  
Corso di Laurea in Informatica



**SAPIENZA**  
UNIVERSITÀ DI ROMA

Candidato: Alessio Bandiera

Relatore: Ivano Salvo

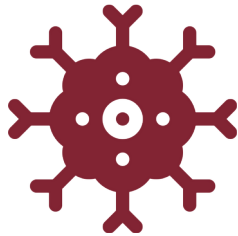
Anno Accademico: 2023/2024

# Il cancro

Il cancro è un gruppo di malattie caratterizzate dalla crescita incontrollata delle cellule.

# Il cancro

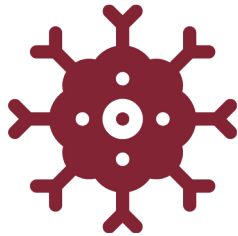
Il cancro è un gruppo di malattie caratterizzate dalla crescita incontrollata delle cellule.



Esistono oltre 100 tipi di cancro, e.g. carcinomi, sarcomi e leucemie.

# Il cancro

Il cancro è un gruppo di malattie caratterizzate dalla crescita incontrollata delle cellule.



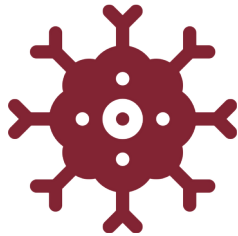
Esistono oltre 100 tipi di cancro, e.g. carcinomi, sarcomi e leucemie.



Ogni anno i decessi per il cancro sono nell'ordine dei milioni.

# Il cancro

Il cancro è un gruppo di malattie caratterizzate dalla crescita incontrollata delle cellule.



Esistono oltre 100 tipi di cancro, e.g. carcinomi, sarcomi e leucemie.



Ogni anno i decessi per il cancro sono nell'ordine dei milioni.



È importante trovare trattamenti efficaci contro questa malattia.

# Cure attuali

Le cure ed i trattamenti per il cancro attualmente disponibili sono:



- chirurgia
- radioterapia
- chemioterapia
- terapia ormonale

# Cure attuali

Le cure ed i trattamenti per il cancro attualmente disponibili sono:

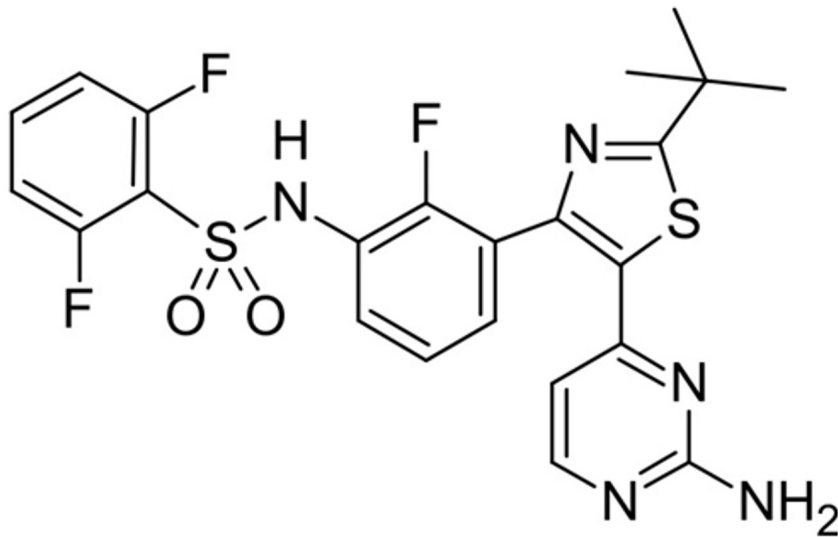


- chirurgia
- radioterapia
- chemioterapia
- terapia ormonale

**Problema.** Tutti i trattamenti attuali sono limitati, e possono portare a molteplici effetti collaterali.

# Terapia a bersaglio

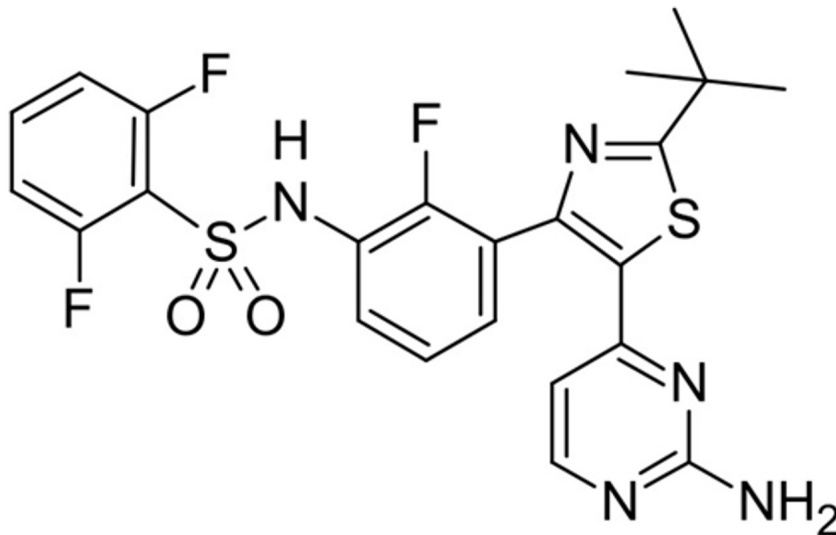
La **terapia a bersaglio** è un trattamento per il cancro che si concentra sulle proteine responsabili della crescita del tumore.





# Terapia a bersaglio

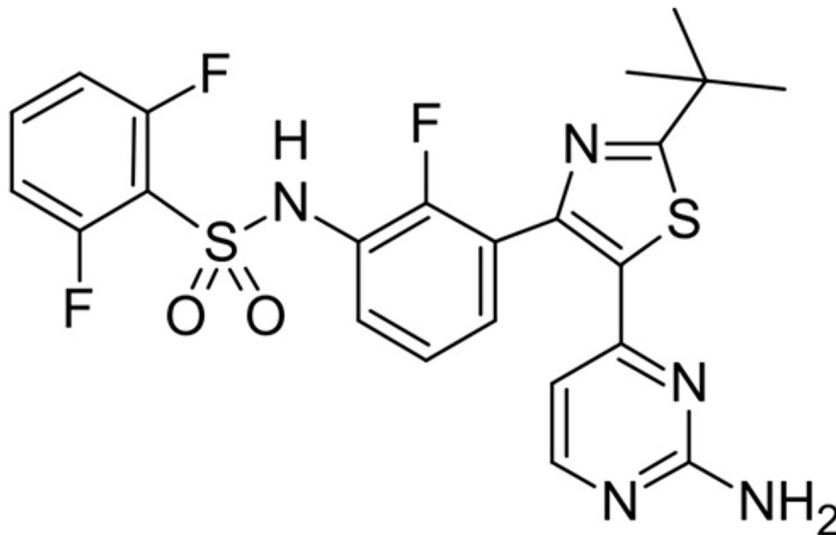
La **terapia a bersaglio** è un trattamento per il cancro che si concentra sulle proteine responsabili della crescita del tumore.



La terapia a bersaglio offre maggiore selettività e può aiutare a ridurre gli effetti collaterali.

# Terapia a bersaglio

La **terapia a bersaglio** è un trattamento per il cancro che si concentra sulle proteine responsabili della crescita del tumore.



La terapia a bersaglio offre maggiore selettività e può aiutare a ridurre gli effetti collaterali.



Cosa bersagliare?

# Il ruolo delle mutazioni nel cancro

Il cancro evolve attraverso un processo *multistep* in cui le cellule acquisiscono gradualmente caratteristiche maligne tramite **mutazioni**.

# Il ruolo delle mutazioni nel cancro

Il cancro evolve attraverso un processo *multistep* in cui le cellule acquisiscono gradualmente caratteristiche maligne tramite **mutazioni**.



Lo sviluppo del cancro è un **processo di mutazione** e selezione di cellule con capacità sempre maggiori di proliferare.

# Il ruolo delle mutazioni nel cancro

Il cancro evolve attraverso un processo *multistep* in cui le cellule acquisiscono gradualmente caratteristiche maligne tramite **mutazioni**.



Lo sviluppo del cancro è un **processo di mutazione** e selezione di cellule con capacità sempre maggiori di proliferare.



Le **mutazioni** ricoprono un ruolo fondamentale per lo sviluppo e la progressione del cancro.

# Tipi di mutazioni

**Definizione.** (Mutazione *passenger*) Una mutazione *passenger* è una mutazione che non conferisce vantaggio diretto al cancro.

# Tipi di mutazioni

**Definizione.** (Mutazione *passenger*) Una mutazione *passenger* è una mutazione che non conferisce vantaggio diretto al cancro.

**Definizione.** (Mutazione *driver*) Una mutazione *driver* è una mutazione che contribuisce direttamente alla crescita tumorale.

# Tipi di mutazioni

**Definizione.** (Mutazione *passenger*) Una mutazione *passenger* è una mutazione che non conferisce vantaggio diretto al cancro.

**Definizione.** (Mutazione *driver*) Una mutazione *driver* è una mutazione che contribuisce direttamente alla crescita tumorale.



Colpendo le mutazioni *driver* con terapie a bersaglio è possibile ridurre lo sviluppo del cancro.



# Tipi di mutazioni

**Definizione.** (Mutazione *passenger*) Una mutazione *passenger* è una mutazione che non conferisce vantaggio diretto al cancro.

**Definizione.** (Mutazione *driver*) Una mutazione *driver* è una mutazione che contribuisce direttamente alla crescita tumorale.



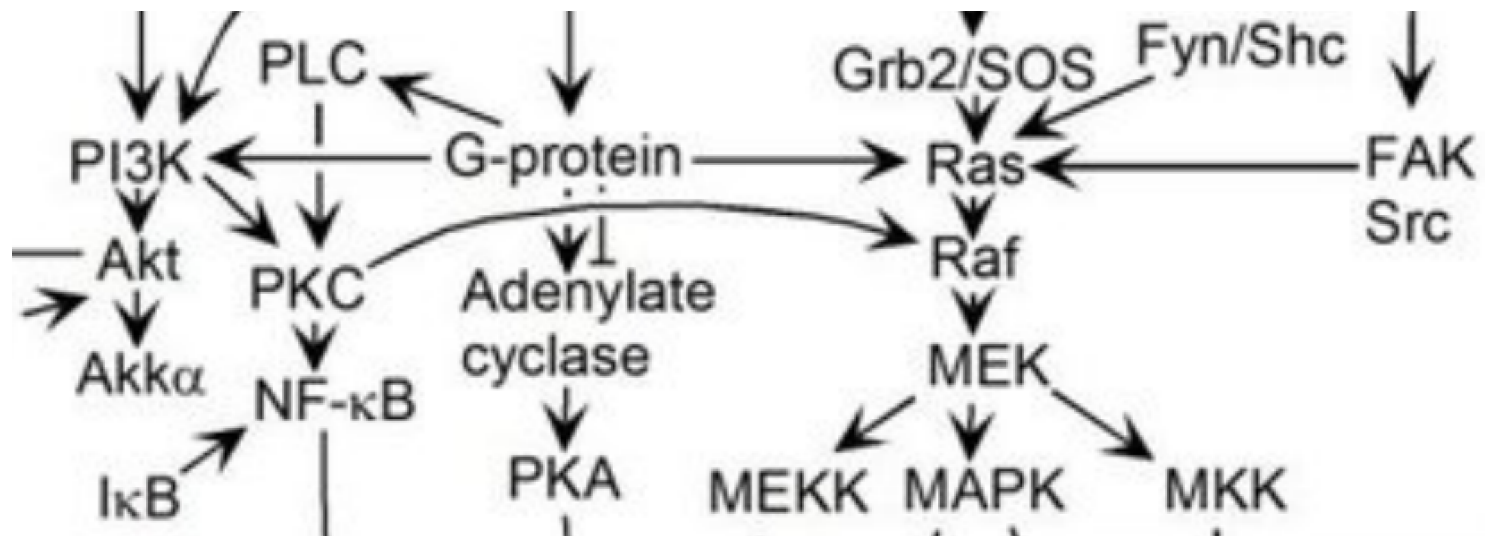
Colpendo le mutazioni *driver* con terapie a bersaglio è possibile ridurre lo sviluppo del cancro.



Classificare le mutazioni tra *driver* e *passenger* è essenziale.

# Pathway cellulari

**Definizione.** (Pathway) Un pathway cellulare è una catena di processi biochimici che avvengono all'interno di una cellula.



Siamo interessati ai geni che compongono i pathway.

## Cercare i pathway *driver*

I pathway sono importanti poiché nel loro contesto è possibile valutare la ricorrenza delle singole mutazioni.

## Cercare i *pathway driver*

I *pathway* sono importanti poiché nel loro contesto è possibile valutare la ricorrenza delle singole mutazioni.



Più mutazioni *driver* in geni diversi possono portare a simili effetti *downstream*, dunque il vantaggio selettivo è distribuito tra le frequenze delle varie alterazioni.

## Cercare i *pathway driver*

I *pathway* sono importanti poiché nel loro contesto è possibile valutare la ricorrenza delle singole mutazioni.



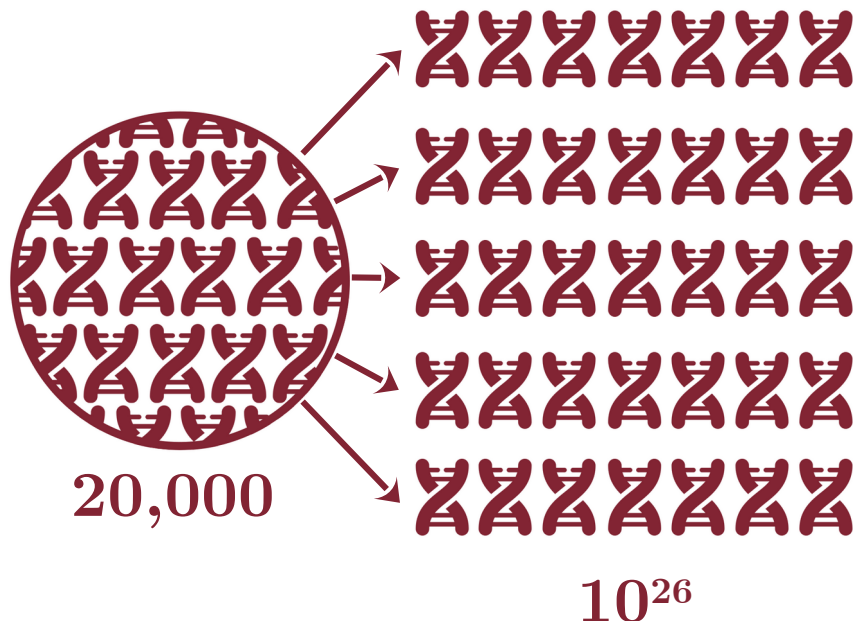
Più mutazioni *driver* in geni diversi possono portare a simili effetti *downstream*, dunque il vantaggio selettivo è distribuito tra le frequenze delle varie alterazioni.



Mutazioni diverse possono influenzare lo stesso *pathway* in vari campioni.

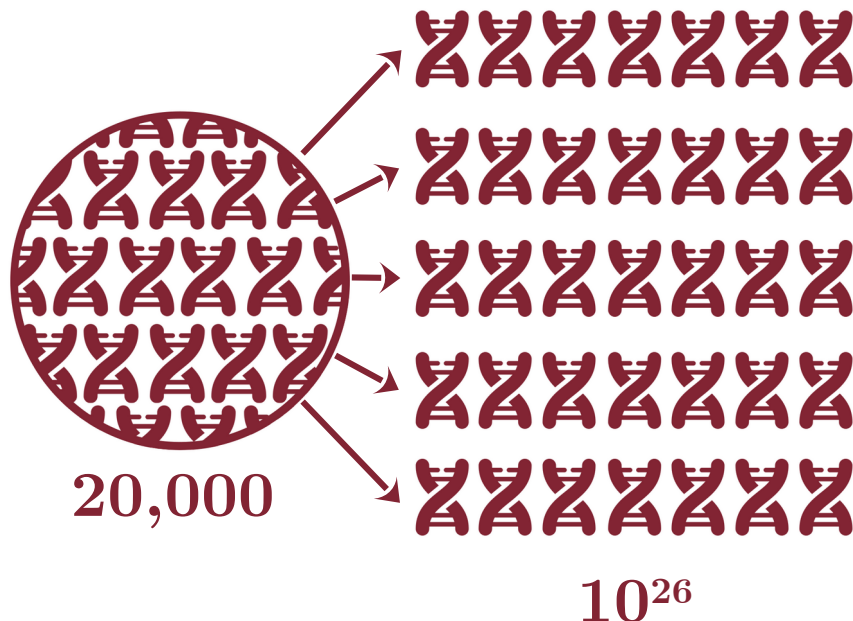
# Problemi nel cercare i pathway

**Problema.** Cercare pathway *driver* è complesso, per via dell'enorme numero di pathway possibili da verificare, e.g. ci sono più di  $10^{26}$  insiemi possibili di 7 geni.



## Problemi nel cercare i pathway

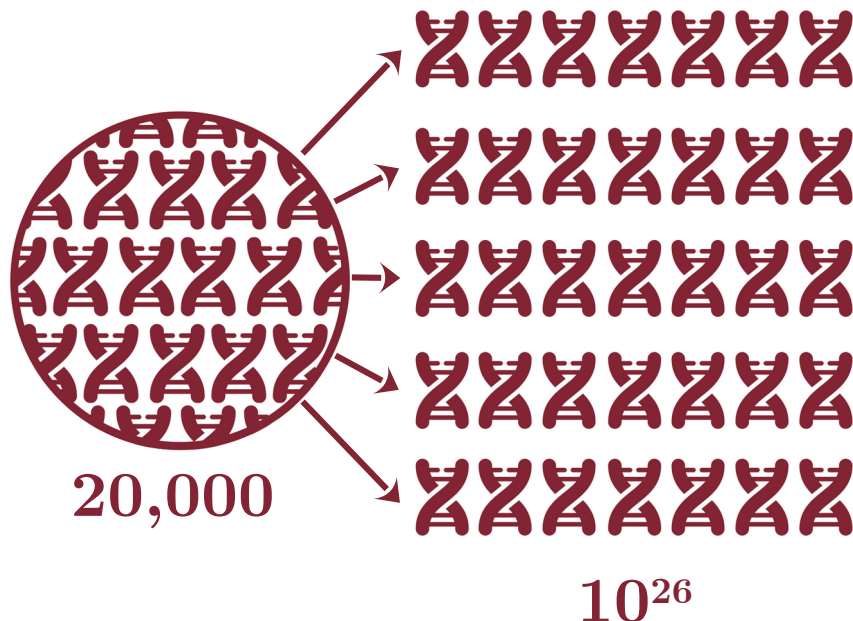
**Problema.** Cercare pathway *driver* è complesso, per via dell'enorme numero di pathway possibili da verificare, e.g. ci sono più di  $10^{26}$  insiemi possibili di 7 geni.



Non è possibile controllare ogni pathway.

# Problemi nel cercare i pathway

**Problema.** Cercare pathway *driver* è complesso, per via dell'enorme numero di pathway possibili da verificare, e.g. ci sono più di  $10^{26}$  insiemi possibili di 7 geni.



Non è possibile controllare ogni pathway.

Fortunatamente, statisticamente si sono osservate proprietà che permettono di ridurre il numero di pathway da controllare.



# Copertura

Pathway importanti per il cancro dovrebbero essere alterati in un numero significativo di pazienti.

# Copertura

Pathway importanti per il cancro dovrebbero essere alterati in un numero significativo di pazienti.



La maggior parte dei pazienti dovrebbe presentare mutazioni in geni di pathway importanti per il cancro.

# Copertura

Pathway importanti per il cancro dovrebbero essere alterati in un numero significativo di pazienti.



La maggior parte dei pazienti dovrebbe presentare mutazioni in geni di pathway importanti per il cancro.

**Assunzione.** (Copertura) I geni *driver* di pathway *driver* sono mutati nella maggior parte dei pazienti.

# Mutua esclusività

Empiricamente, si osserva che ogni paziente ha poche mutazioni *driver*, le quali però colpiscono molteplici pathway.

# Mutua esclusività

Empiricamente, si osserva che ogni paziente ha poche mutazioni *driver*, le quali però colpiscono molteplici pathway.



Ogni pathway avrà approssimativamente 1 mutazione *driver*.

# Mutua esclusività

Empiricamente, si osserva che ogni paziente ha poche mutazioni *driver*, le quali però colpiscono molteplici pathway.



Ogni pathway avrà approssimativamente 1 mutazione *driver*.

**Assunzione.** (Mutua esclusività) I geni *driver* all'interno dello stesso pathway sono approssimativamente mutuamente esclusivi.

# Pathway *driver*

**Assunzione.** (Copertura) I geni *driver* di pathway *driver* sono mutati nella maggior parte dei pazienti.

**Assunzione.** (Mutua esclusività) I geni *driver* all'interno dello stesso pathway sono approssimativamente mutuamente esclusivi.

# Pathway *driver*

**Assunzione.** (Copertura) I geni *driver* di pathway *driver* sono mutati nella maggior parte dei pazienti.

**Assunzione.** (Mutua esclusività) I geni *driver* all'interno dello stesso pathway sono approssimativamente mutuamente esclusivi.



**Definizione.** (Pathway *driver*) Un pathway *driver* è un pathway costituito da geni mutati in numerosi pazienti, e le cui mutazioni sono approssimativamente mutualmente esclusive all'interno del pathway.



## Matrice di Mutazione

**Definizione.** (Matrice di mutazione) Una matrice di mutazione è una matrice binaria che descrive le mutazioni dei pazienti.

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$p_1$	0	1	0	0	1
$p_2$	1	0	1	0	0
$p_3$	0	1	1	0	0
$p_4$	0	0	0	1	1

$$a_{i,j} = 1 \iff i \text{ ha il gene } j \text{ mutato}$$

## Copertura di un gene

**Definizione.** (Copertura di un gene) La copertura di un gene  $g$  è l'insieme dei pazienti che hanno  $g$  mutato.

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$p_1$	0	1	0	0	1
$p_2$	1	0	1	0	0
$p_3$	0	1	1	0	0
$p_4$	0	0	0	1	1

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

## Copertura di un gene

**Definizione.** (Copertura di un gene) La copertura di un gene  $g$  è l'insieme dei pazienti che hanno  $g$  mutato.

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$p_1$	0	1	0	0	1
$p_2$	1	0	1	0	0
$p_3$	0	1	1	0	0
$p_4$	0	0	0	1	1

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

## Copertura di un insieme di geni

**Definizione.** (Copertura di un insieme di geni) La copertura di un insieme di geni  $M$  è l'unione delle coperture dei geni di  $M$ .

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$p_1$	0	1	0	0	1
$p_2$	1	0	1	0	0
$p_3$	0	1	1	0	0
$p_4$	0	0	0	1	1

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

## Copertura di un insieme di geni

**Definizione.** (Copertura di un insieme di geni) La copertura di un insieme di geni  $M$  è l'unione delle coperture dei geni di  $M$ .

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$p_1$	0	1	0	0	1
$p_2$	1	0	1	0	0
$p_3$	0	1	1	0	0
$p_4$	0	0	0	1	1

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

## Mutua esclusività

**Definizione.** (Mutua esclusività)  $M$  è mutuamente esclusivo se non ci sono pazienti con più di una mutazione di geni di  $M$ .

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$p_1$	0	1	0	0	1
$p_2$	1	0	1	0	0
$p_3$	0	1	1	0	0
$p_4$	0	0	0	1	1

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

## Sovrapposizione di un insieme di geni

**Definizione.** (Sovrapposizione)  $\omega(M)$  rappresenta il numero di pazienti con più di un gene di  $M$  mutato.

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$p_1$	0	1	0	0	1
$p_2$	1	0	1	0	0
$p_3$	0	1	1	0	0
$p_4$	0	0	0	1	1

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

## Sovrapposizione di un insieme di geni

**Definizione.** (Sovrapposizione)  $\omega(M)$  rappresenta il numero di pazienti con più di un gene di  $M$  mutato.

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$p_1$	0	1	0	0	1
$p_2$	1	0	1	0	0
$p_3$	0	1	1	0	0
$p_4$	0	0	0	1	1

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



## Sovrapposizione di un insieme di geni

**Definizione.** (Sovrapposizione)  $\omega(M)$  rappresenta il numero di pazienti con più di un gene di  $M$  mutato.

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$p_1$	0	1	0	0	1
$p_2$	1	0	1	0	0
$p_3$	0	1	1	0	0
$p_4$	0	0	0	1	1

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

## Sovrapposizione di un insieme di geni

**Definizione.** (Sovrapposizione)  $\omega(M)$  rappresenta il numero di pazienti con più di un gene di  $M$  mutato.

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$p_1$	0	1	0	0	1
$p_2$	1	0	1	0	0
$p_3$	0	1	1	0	0
$p_4$	0	0	0	1	1

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

## Peso di un gruppo di geni

**Definizione.** (Peso) Il peso di un gruppo di geni è la differenza tra la sua copertura e la sua sovrapposizione.

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$p_1$	0	1	0	0	1
$p_2$	1	0	1	0	0
$p_3$	0	1	1	0	0
$p_4$	0	0	0	1	1

$$W(M) := |\Gamma(M)| - \omega(M)$$

## Peso di un gruppo di geni

**Definizione.** (Peso) Il peso di un gruppo di geni è la differenza tra la sua copertura e la sua sovrapposizione.

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$p_1$	0	1	0	0	1
$p_2$	1	0	1	0	0
$p_3$	0	1	1	0	0
$p_4$	0	0	0	1	1

$$W(M) := |\Gamma(M)| - \omega(M)$$

## Peso di un gruppo di geni

**Definizione.** (Peso) Il peso di un gruppo di geni è la differenza tra la sua copertura e la sua sovrapposizione.

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$p_1$	0	1	0	0	1
$p_2$	1	0	1	0	0
$p_3$	0	1	1	0	0
$p_4$	0	0	0	1	1

$$W(M) := |\Gamma(M)| - \omega(M)$$

## Peso di un gruppo di geni

**Definizione.** (Peso) Il peso di un gruppo di geni è la differenza tra la sua copertura e la sua sovrapposizione.

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$p_1$	0	1	0	0	1
$p_2$	1	0	1	0	0
$p_3$	0	1	1	0	0
$p_4$	0	0	0	1	1

$$W(M) := |\Gamma(M)| - \omega(M)$$

# Maximum Weight Submatrix Problem (MWSP)

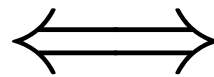
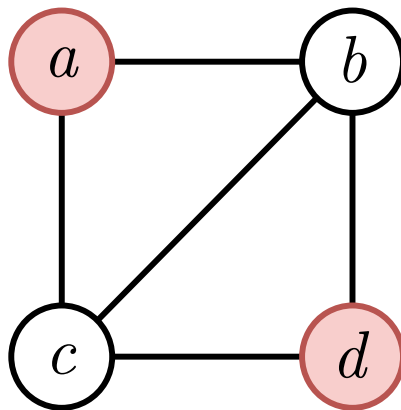
**Definizione.** (MWSP) Data una matrice di mutazione  $A$  di dimensioni  $m \times n$ , ed un intero  $k > 0$ , si trovi una sottomatrice  $m \times k$  di  $A$  tale da massimizzare  $W(M)$ .

	$g_1$	$g_2$	$g_3$	$g_4$	$g_5$
$p_1$	0	1	0	0	1
$p_2$	1	0	1	0	0
$p_3$	0	1	1	0	0
$p_4$	0	0	0	1	1

# Maximum Weight Submatrix Problem (MWSP)

**Definizione.** (MWSP) Data una matrice di mutazione  $A$  di dimensioni  $m \times n$ , ed un intero  $k > 0$ , si trovi una sottomatrice  $m \times k$  di  $A$  tale da massimizzare  $W(M)$ .

**Teorema.** (MWSP) L'MWSP è NP-completo.



	$a$	$b$	$c$	$d$
$s(a,b)$	1	1	0	0
$s(a,c)$	1	0	1	0
$s(b,c)$	0	1	1	0
$s(c,d)$	0	0	1	1
$s(b,d)$	0	1	0	1
$s_a^{(1)}$	1	0	0	0
$s_d^{(1)}$	0	0	0	1



## Un ILP per l'MWSP

**Definizione.** (Indicatrice di  $M$ )  $I_M(j)$  è la variabile indicatrice che descrive l'insieme di geni  $M$ .

$$I_M(j) = 1 \iff j \in M$$

## Un ILP per l'MWSP

**Definizione.** (Indicatrice di  $M$ )  $I_M(j)$  è la variabile indicatrice che descrive l'insieme di geni  $M$ .

$$I_M(j) = 1 \iff j \in M$$

**Definizione.** (Indicatrice di  $\Gamma(M)$ )  $C_i(M)$  è la variabile indicatrice che descrive quali pazienti copre  $M$ .

$$C_i(M) = 1 \iff \exists g \in M \mid i \in \Gamma(g)$$

## Un ILP per l'MWSP

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

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

$$\sum_{j=1}^n I_M(j) \cdot a_{i,j} \geq C_i(M), \quad 1 \leq i \leq m.$$

# Multiple Maximum Weight Submatrix Problem (MMWSP)

**Definizione.** (MMWSP) Data una matrice di mutazione  $A$  di dimensioni  $m \times n$ , ed un intero  $t > 0$ , si trovi la collezione  $M = \{M_1, \dots, M_t\}$  di sottomatrici colonna di  $A$  che massimizzi

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

	$g_1$	$g_2$	$g_3$	$g_4$
$p_1$	0	1	0	1
$p_2$	1	0	0	0
$p_3$	0	1	0	0
$p_4$	0	0	1	1

# Approcci statistici

**Problema.** La metrica  $W(M)$  assume che i pathway *driver* abbiano i geni esattamente mutuamente esclusivi, ma la mutua esclusività esatta nei dati reali si verifica raramente

# Approcci statistici

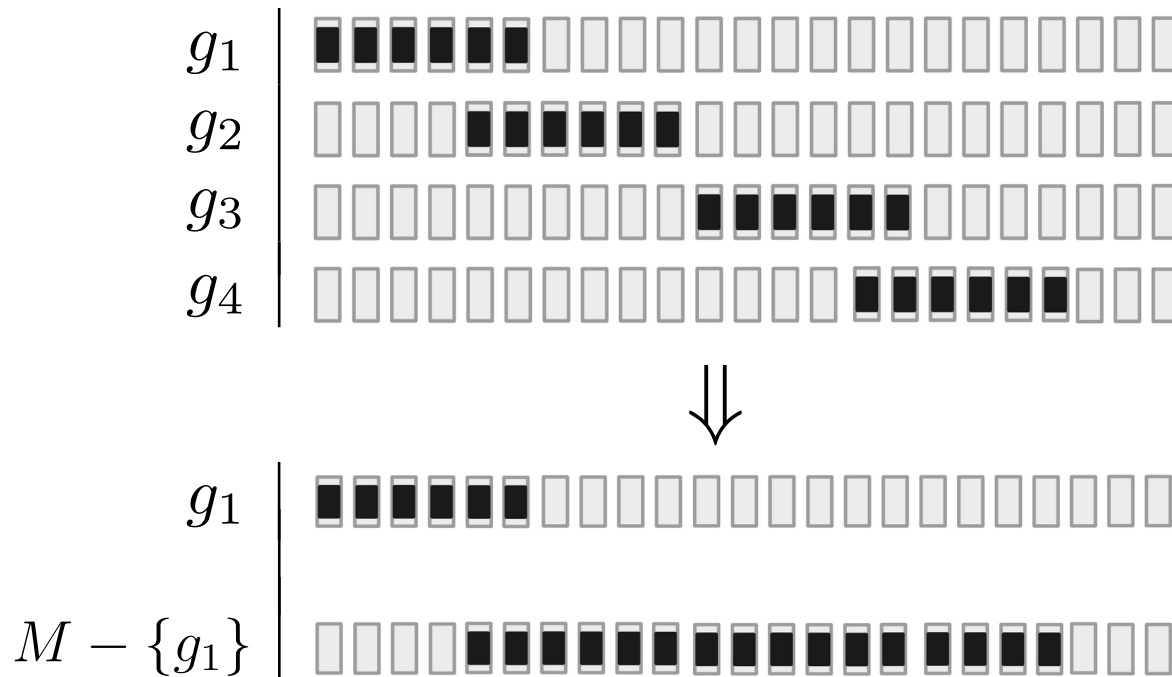
**Problema.** La metrica  $W(M)$  assume che i pathway *driver* abbiano i geni esattamente mutuamente esclusivi, ma la mutua esclusività esatta nei dati reali si verifica raramente.



Nonostante  $W(M)$  permetta di formulare facilmente problemi di ottimizzazione per trovare pathway *driver*, approcci statistici tendono a performare meglio su dati reali.

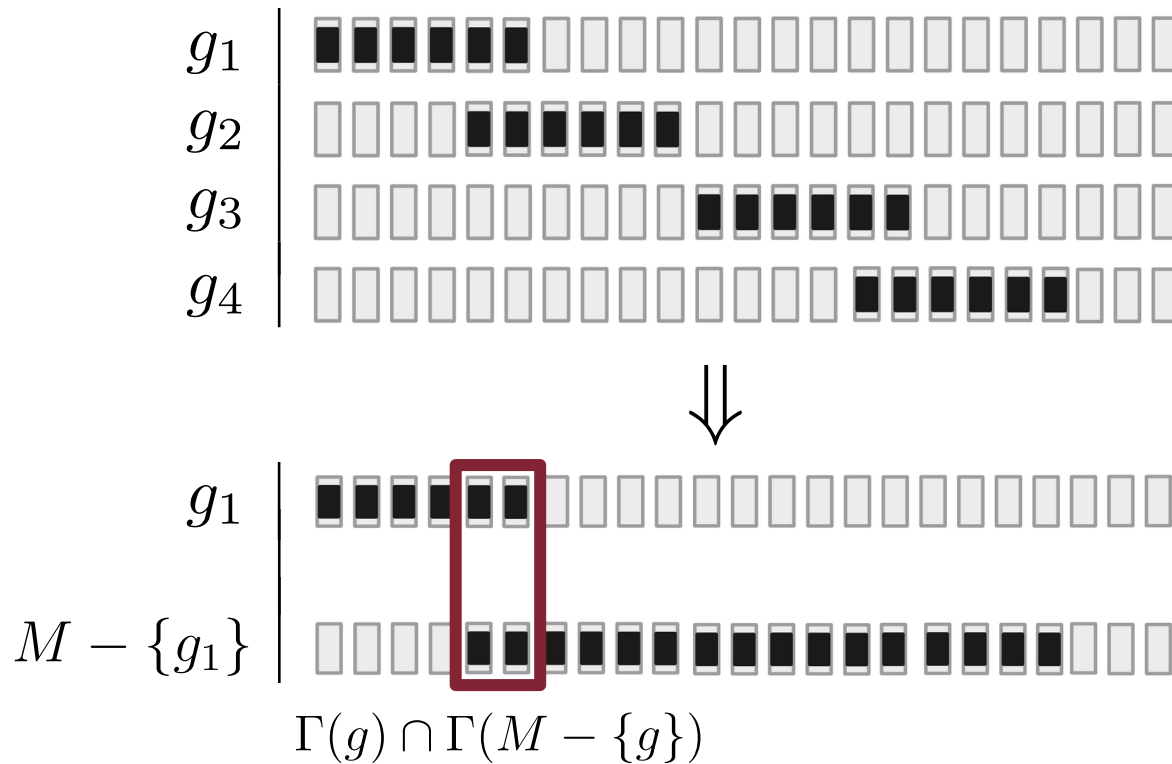
# Ipotesi nulla

**Definizione.** ( $H_0$ ) Dato un gruppo di geni  $M$ , un gene  $g$  di  $M$  è alterato indipendentemente dall'unione delle alterazioni dei geni in  $M - \{g\}$ .



# Ipotesi nulla

**Definizione.** ( $H_0$ ) Dato un gruppo di geni  $M$ , un gene  $g$  di  $M$  è alterato indipendentemente dall'unione delle alterazioni dei geni in  $M - \{g\}$ .





# Punteggio di un insieme di geni

**Definizione.** (Variabile aleatoria  $X$ )  $X$  è la variabile aleatoria che rappresenta il numero di pazienti aventi sia  $g$  che un qualsiasi altro gene in  $M - \{g\}$  mutato.

## Punteggio di un insieme di geni

**Definizione.** (Variabile aleatoria  $X$ )  $X$  è la variabile aleatoria che rappresenta il numero di pazienti aventi sia  $g$  che un qualsiasi altro gene in  $M - \{g\}$  mutato.

$$X \sim H(m, \Gamma(g), \Gamma(M - \{g\}))$$

## Punteggio di un insieme di geni

**Definizione.** (Variabile aleatoria  $X$ )  $X$  è la variabile aleatoria che rappresenta il numero di pazienti aventi sia  $g$  che un qualsiasi altro gene in  $M - \{g\}$  mutato.

$$X \sim H(m, \Gamma(g), \Gamma(M - \{g\}))$$

$$\Downarrow$$

$$p_g := P(X = \Gamma(g) \cap \Gamma(M - \{g\}))$$

## Punteggio di un insieme di geni

**Definizione.** (Variabile aleatoria  $X$ )  $X$  è la variabile aleatoria che rappresenta il numero di pazienti aventi sia  $g$  che un qualsiasi altro gene in  $M - \{g\}$  mutato.

$$X \sim H(m, \Gamma(g), \Gamma(M - \{g\}))$$

$$\Downarrow$$

$$p_g := P(X = \Gamma(g) \cap \Gamma(M - \{g\}))$$

$$\Downarrow$$

$$s_M := \max_{g \in M} p_g$$

# Algoritmo genetico

L'algoritmo genetico utilizza la stessa funzione di *fitness*  $W(M)$ .

$g_1$		0111001100
$g_2$		1011011000
$g_3$		0001110011
$g_4$		0101010110

⋮

$g_n$		0101010110
-------	--	------------

⋮

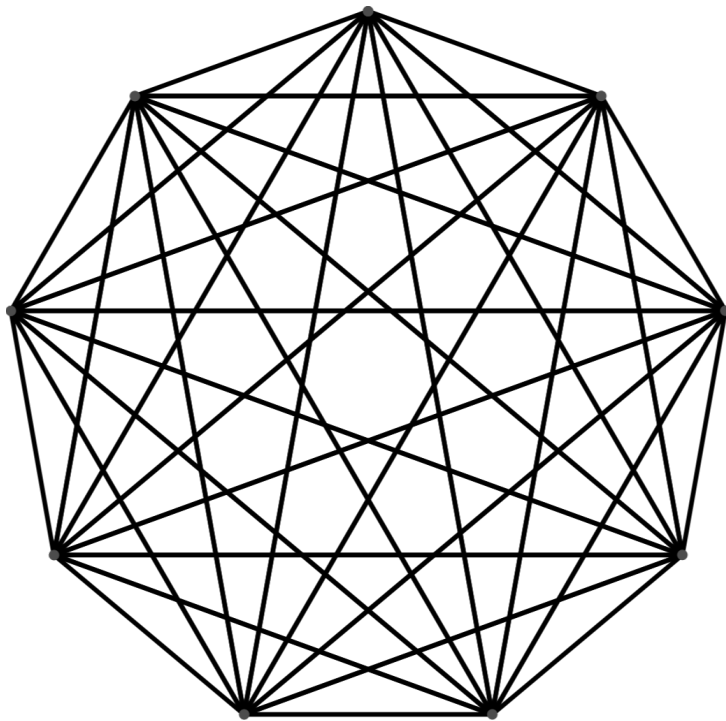
**Definizione.** (Membro) Un membro della popolazione è una stringa binaria che rappresenta un insieme di geni  $M$ .



**Definizione.** (Crossover) Un figlio eredita dai genitori i bit in comune, mentre gli altri sono casuali.

# Algoritmo di clustering

**Definizione.** (Grafo di geni) Un grafo di geni è un grafo completamente connesso in cui ogni arco ha assegnati due pesi.



**Definizione.** (Peso negativo) Il peso negativo di un arco  $(u,v)$  è il costo di posizionare  $u$  e  $v$  nello stesso cluster.

$$w_{uv}^- := w_{uv}^-(e)$$

**Definizione.** (Peso positivo) Il peso positivo di un arco  $(u,v)$  è il costo di posizionare  $u$  e  $v$  in cluster diversi.

$$w_{uv}^+ := w_1 w_{uv}^+(c) + w_2 w_{uv}^+(n) + w_3 w_{uv}^+(x)$$

## Lavori futuri

L'identificazione dei pathway *driver* offre prospettive promettenti per migliorare l'efficacia delle terapie a bersaglio, che potrebbero portare a terapie più personalizzate.

## Lavori futuri

L'identificazione dei pathway *driver* offre prospettive promettenti per migliorare l'efficacia delle terapie a bersaglio, che potrebbero portare a terapie più personalizzate.

Future ricerche potrebbero integrare tecnologie emergenti, come il *single-cell sequencing*.



## Lavori futuri

L'identificazione dei pathway *driver* offre prospettive promettenti per migliorare l'efficacia delle terapie a bersaglio, che potrebbero portare a terapie più personalizzate.

Future ricerche potrebbero integrare tecnologie emergenti, come il *single-cell sequencing*.

Sono necessari algoritmi che permettano di tenere in considerazione l'eterogeneità tumorale ed i meccanismi di resistenza adattativa alle terapie a bersaglio.

**Grazie per l'attenzione**

