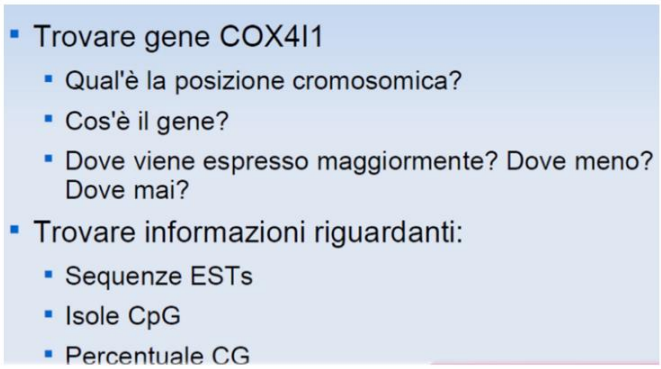
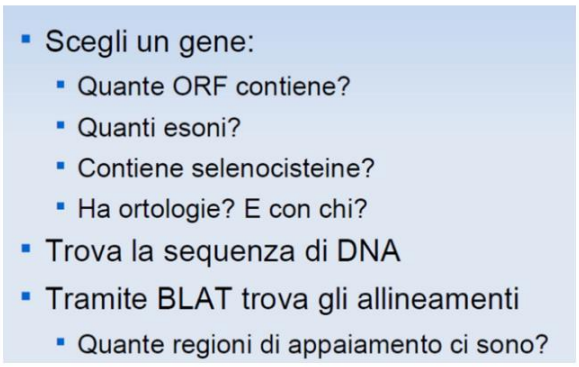
LEZIONE DI GENETICA N°16 18/05/2023
ANALISI IN SILICO

ESERCIZIO





Sito web utilizzato: https://genome.ucsc.edu

1. Si deve scegliere il genoma di riferimento (umano, topo,…..).
2. Si cerca il gene COX4I1.
3. Appaiano una serie di informazioni (che possono anche essere nascoste o mostrate in formato pack o full).

La barretta rossa indica dove è localizzato il gene sul cromosoma.
Le freccette sono usate per muoversi sul gene e possiamo zoomare su di esso (vedere informazioni su un errore in particolare, andando su "base" si possono vedere anche le basi che lo compongono).

COX4I1 un gene formato da 5 esoni.
Se voglio avere una informazione più specifica, clicco sul gene.
NM_001318794 è il codice di rifermento della sequenza nucleotidica. Se questo codice viene messo sul sito Nucleotide (https://www.ncbi.nlm.nih.gov/nucleotide/) , ci restituisce l'intera sequenza nucleotidica.
Le informazioni principali che troviamo su Genome Browser sono le seguenti:

- Posizione.
- Grandezza.
- Totale esoni
- Se possiede altri nomi (alias).
- Informazioni sul modello genico (es: se ha selenocisteine, se ha un codone di inizio…).
- Conformazione della proteina.
- Se ci sono ortologhi e paraloghi,
- Per alcune proteine è presente l'allineamento.
- In quali tessuti è maggiormente espresso il gene (in questo caso ventricolo sx).
- GC percent.
- Letteratura (in quali pubblicazioni è presente il gene).
- FASTA format.
- Gene Ontology (GO) Annotations (informazioni sulle funzioni molecolari, i processi biologici e le componenti cellulari.

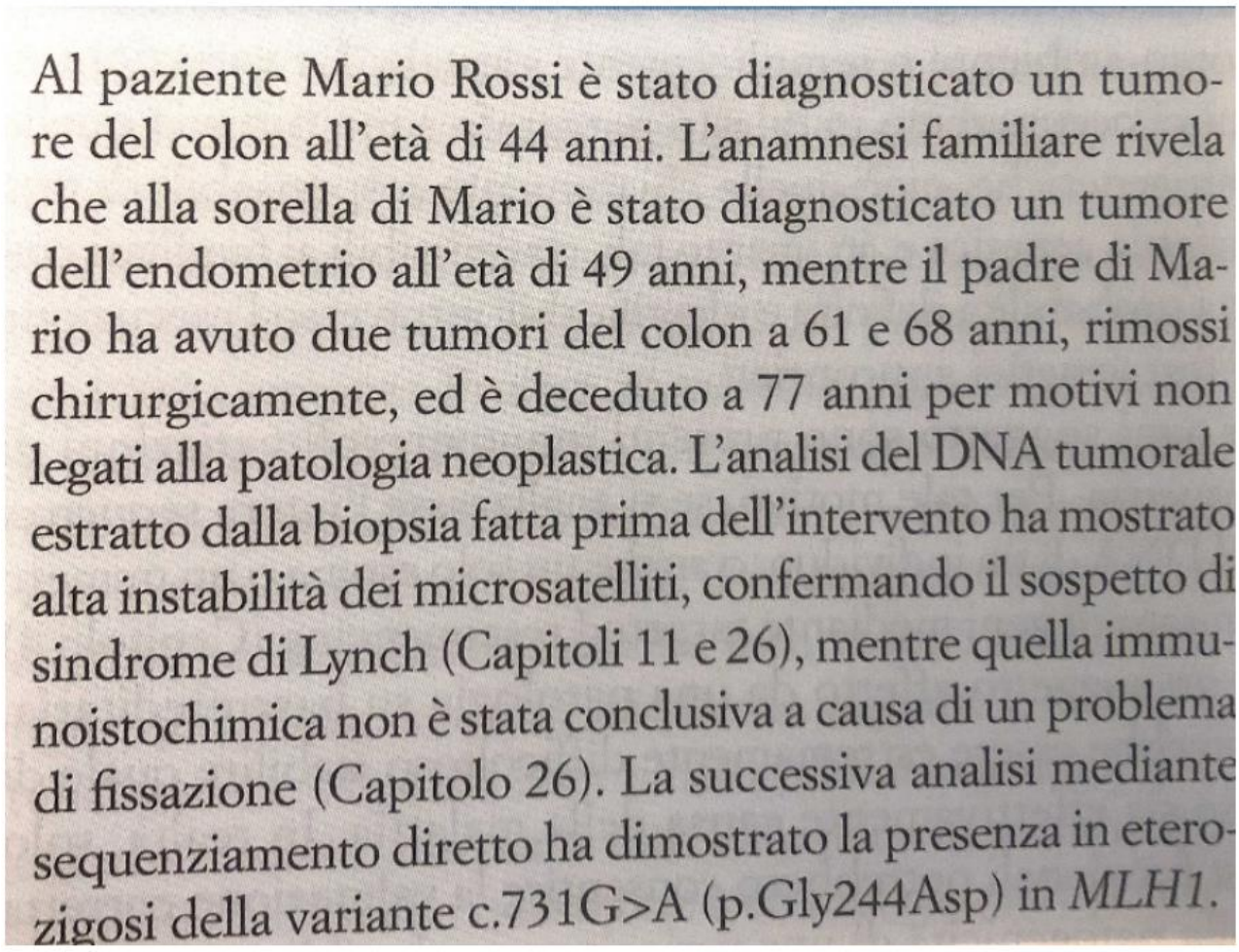
Molte di queste informazioni possiamo mostrarle nella schermata iniziale.
In ogni schermata vi sono delle cross references che ci conducono ad altri database.
Nella schermata iniziale possiamo visualizzare i fenotipi OMIM Gene
(https://omim.org/graph/radial/123864). Il punto interrogativo indica che non sono ancora state

scoperte.
È possibile visualizzare varianti e repeat. GnomAD è il più usato e le “barrette” sono polimorfismi). Cliccando su una variante, è possibile visualizzarne la posizione, la popolazione dove è presente e le pubblicazioni.

ANALISI IN SILICO

La locuzione è usata per indicare fenomeni biologici riprodotti in una simulazione matematica al computer, invece che in provetta o in un essere vivente. Infatti, il silicio, è la sostanza di cui sono fatti i componenti elettronici all'interno della quasi totalità dei computer, anche se il concetto di simulazione matematica non ha niente a che fare con il silicio. Al contrario, se il fenomeno biologico si riproduce in provetta, si dice "in vitro", mentre se si riproduce in un essere vivente si dice "In vivo".
La ricerca scientifica in silico è ovviamente il frutto della recente informatizzazione della ricerca. In biologia hanno assunto importanza fondamentale le basi di dati che contengono i dati di sequenziamento del DNA e i livelli di espressione genica di un gene.
Inoltre, sono stati creati numerosi programmi informatici per l'analisi e l'interpretazione di dati sperimentali.
La ricerca in silico consiste, dunque, nell'uso di tali programmi e raccolte di dati allo scopo di ricavare nuove informazioni dalla comparazione, interpretazione, analisi ed interconnessione dei dati.

ESERCIZIO


Al paziente Mario Rossi è stato diagnosticato un tumore del colon all’età di 44 anni. L’anamnesi familiare rivela che alla sorella di Mario è stato diagnosticato un tumore dell’endometrio all’età di 49 anni, mentre il padre di Mario ha avuto due tumori del colon a 61 e 68 anni, rimossi chirurgicamente, ed è deceduto a 77 anni per motivi non legati alla patologia neoplastica. L’analisi del DNA tumorale estratto dalla biopsia fatta prima dell’intervento ha mostrato alta instabilità dei microsatelliti, confermando il sospetto di sindrome di Lynch (Capitoli 11 e 26), mentre quella immunoistochimica non è stata conclusiva a causa di un problema di fissazione (Capitolo 26). La successiva analisi mediante sequenziamento diretto ha dimostrato la presenza in eterozigosi della variante c.731G>A (p.Gly244Asp) in *MLH1*.

731 G>A significa che in posizione 731 una guanina è sostituita da una adenina questo porta che nel codone 244 una Glicina è sostituita da Acido aspartico.

1. Si va su un genome browser (la prof usa https://www.ensembl.org/index.html);
2. Si cerca il gene (MLH1);
3. Clicco sul primo trascritto;
4. Vado su cDNA: ci dà più informazioni; ha 3 linee. I codoni codificanti sono in giallo. La prima linea è la sequenza genomica, la seconda la parte codificante e la terza la proteina;
5. Vado al codone 244 (trovo una G) o in posizione nucleotidica 731;
6. Posso consultare le varianti e i nomi delle varianti;
7. Verifico se è patogenica (in questo caso è patogenica e ha un significato incerto, quindi, è necessaria una analisi in silico).

Per l'analisi in silico, si usano diversi database (spesso specifici a seconda della patologia). In questo caso, essendo una malattia oncologica, il sito che si utilizza è: http://agvgd.hci.utah.edu/about.php.

Il sito ci mostra l'allineamento (se è uguale in tutte le specie, il cambiamento sarà deleterio).

Altri software disponibili per l'analisi in silico:

- PolyPhen (http://genetics.bwh.harvard.edu/pph2/).
- Sorting Intolerant From Tolerant (https://sift.bii.a-star.edu.sg/).
- Screening for Non-Acceptable Polymorphisms (https://rostlab.org/services/snap/).
- Protein Analysis Through Evolutionary Relationships(http://pantherdb.org/; http://www.mutationtaster.org)

La professoressa utilizza "SIFT".
SIFT utilizza l'omologia di sequenza tra geni e domini correlati tra le specie per prevedere l'impatto di tutti i 20 possibili amminoacidi in una data posizione, consentendo agli utenti di determinare quali nsSNP sarebbero di maggior interesse da studiare ordinando le varianti in base a questo punteggio di previsione.
Per usarlo, si inserisce la sequenza in formato FASTA e la sostituzione di nostro interesse (il formato FASTA sia del nucleotide che della proteina si trova su GenomeBrowser).

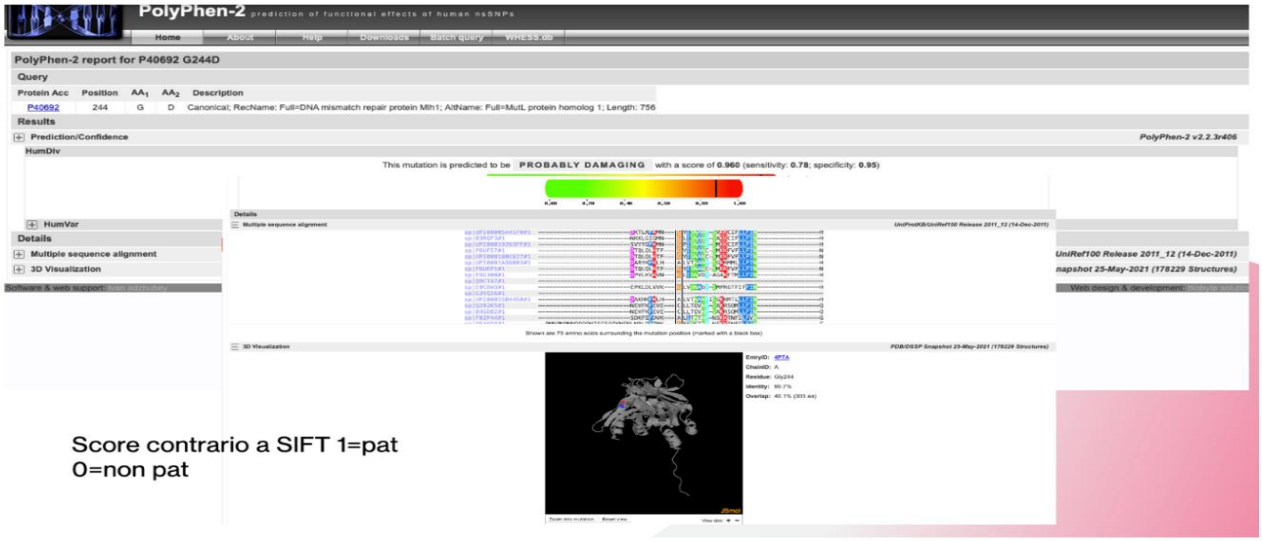
Ci restituisce una predizione ad ogni posizione del codone in cui a sx ci sono le posizioni non tollerate e a dx quelle tollerate. Se lo score è minore o uguale a 0,05 non è tollerato.

pos	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
242M	1.00	0.61	0.08	0.04	0.08	0.10	0.06	0.04	0.45	0.09	0.46	0.76	0.05	0.10	0.07	0.06	0.11	0.26	1.00	0.02	0.06
243N	1.00	0.13	0.02	0.58	0.21	0.02	0.22	0.07	0.02	0.42	0.04	0.02	1.00	0.07	0.12	0.09	0.21	0.12	0.04	0.01	0.04
244G	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
245Y	1.00	0.42	0.14	0.18	0.24	0.77	0.27	0.44	0.33	0.40	0.54	0.18	0.26	0.17	0.26	0.32	0.33	0.33	0.41	0.15	1.00
246I	1.00	0.03	0.01	0.01	0.01	0.26	0.01	0.00	1.00	0.01	0.20	0.04	0.01	0.01	0.01	0.01	0.01	0.03	0.62	0.01	0.02
247S	1.00	0.52	0.20	0.12	0.09	0.04	0.28	0.04	0.06	0.11	0.09	0.04	0.15	0.15	0.07	0.07	1.00	0.43	0.14	0.02	0.05
248N	1.00	0.05	0.01	0.11	0.04	0.01	0.08	0.02	0.01	0.04	0.01	0.00	1.00	0.03	0.03	0.02	0.34	0.07	0.01	0.00	0.01
249A	1.00	1.00	0.12	0.17	0.28	0.14	0.22	0.09	0.46	0.28	0.60	0.15	0.18	0.21	0.19	0.19	0.44	0.67	0.88	0.03	0.12
250N	1.00	0.02	0.01	0.31	0.04	0.00	0.07	0.02	0.00	0.03	0.01	0.00	1.00	0.02	0.02	0.02	0.06	0.03	0.01	0.00	0.01
251Y	1.00	0.02	0.01	0.00	0.01	0.78	0.01	0.01	0.03	0.00	0.07	0.02	0.00	0.01	0.00	0.00	0.01	0.01	0.03	0.06	1.00
252S	1.00	0.82	0.08	0.66	1.00	0.15	0.44	0.20	0.28	0.96	0.39	0.14	0.61	0.31	0.65	0.57	0.96	0.78	0.36	0.05	0.18
253V	1.00	0.87	0.10	0.62	0.82	0.14	0.48	0.17	0.27	0.67	0.34	0.12	0.55	0.32	0.46	0.40	1.00	0.95	0.43	0.04	0.16

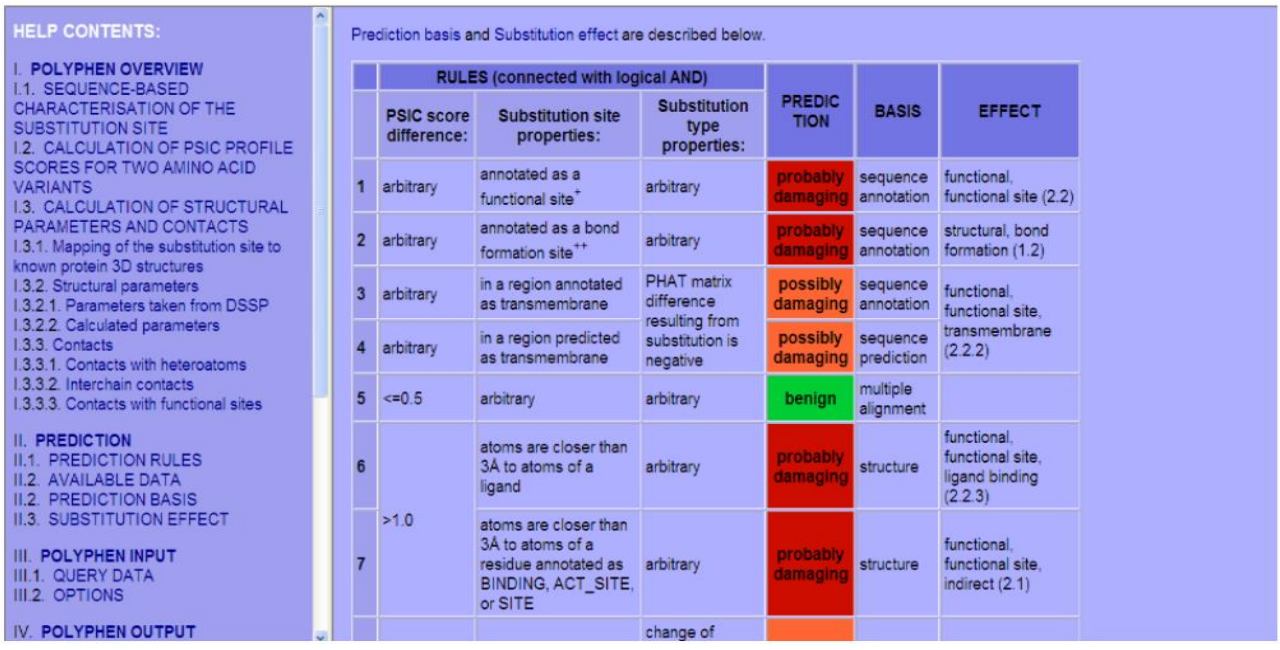
Per una maggiore sicurezza, si consultano anche altri siti di analisi in silico (per vedere se coincidono). Nella posizione 244 (di nostro interesse) come è possibile vedere nell'immagine, sono

tutte non tollerate tranne la proteina originaria (la G). La tolleranza si basa solo sull'amminoacido, a differenza di PolyPhen che analizza anche strutture e domini.

Anche l'algoritmo **PolyPhen**, come **SIFT,** adotta un approccio evolutivo nel distinguere gli nsSNP deleteri da quelli funzionalmente neutri. PolyPhen differisce da SIFT in quanto prevede quanto possa essere dannosa una particolare variante utilizzando un insieme di regole empiriche basate su informazioni di sequenza, filogenetiche e strutturali che caratterizzano una particolare variante.
Oltre a utilizzare gli allineamenti di sequenza, PolyPhen utilizza database di strutture proteiche, come **PDB** (Protein Data Bank) o **PQS** (Protein Quarternary Structure), **DSSP** (Dictionary of Secondary Structure in Proteins) e database di strutture tridimensionali per determinare se una variante può avere un effetto sulla struttura secondaria della proteina, sui contatti intercatena, sui siti funzionali e sui siti di legame.
Score > 1 = patologico.
Ci mostra una serie di allineamenti in cui G non è conservata e anche la posizione 3D della zona che è cambiata nella mutazione.



Position-Specific Independent Counts (PSIC) in PolyPhen riflette la probabilità che un aa occupi una specifica posizione in una sequenza proteica.



HELP CONTENTS:

I. POLYPHEN OVERVIEW
I.1. SEQUENCE-BASED CHARACTERISATION OF THE SUBSTITUTION SITE
I.2. CALCULATION OF PSIC PROFILE SCORES FOR TWO AMINO ACID VARIANTS
I.3. CALCULATION OF STRUCTURAL PARAMETERS AND CONTACTS
I.3.1. Mapping of the substitution site to known protein 3D structures
I.3.2. Structural parameters
I.3.2.1. Parameters taken from DSSP
I.3.2.2. Calculated parameters
I.3.3. Contacts
I.3.3.1. Contacts with heteroatoms
I.3.3.2. Interchain contacts
I.3.3.3. Contacts with functional sites

II. PREDICTION
II.1. PREDICTION RULES
II.2. AVAILABLE DATA
II.2. PREDICTION BASIS
II.3. SUBSTITUTION EFFECT

III. POLYPHEN INPUT
III.1. QUERY DATA
III.2. OPTIONS

IV. POLYPHEN OUTPUT

Prediction basis and Substitution effect are described below.

	RULES (connected with logical AND)			PREDICTION	BASIS	EFFECT
	PSIC score difference:	Substitution site properties:	Substitution type properties:			
1	arbitrary	annotated as a functional site+	arbitrary	probably damaging	sequence annotation	functional, functional site (2.2)
2	arbitrary	annotated as a bond formation site++	arbitrary	probably damaging	sequence annotation	structural, bond formation (1.2)
3	arbitrary	in a region annotated as transmembrane	PHAT matrix difference resulting from substitution is negative	possibly damaging	sequence annotation	functional, functional site, transmembrane (2.2.2)
4	arbitrary	in a region predicted as transmembrane		possibly damaging	sequence prediction	
5	<=0.5	arbitrary	arbitrary	benign	multiple alignment	
6	>1.0	atoms are closer than 3Å to atoms of a ligand	arbitrary	probably damaging	structure	functional, functional site, ligand binding (2.2.3)
7		atoms are closer than 3Å to atoms of a residue annotated as BINDING, ACT_SITE, or SITE	arbitrary	probably damaging	structure	functional, functional site, indirect (2.1)
			change of			

Mutation t@sting (https://www.mutationtaster.org) permette di analizzare il tipo di mutazione grazie all'inserimento di alcuni dati. Anche qui è possibile visualizzare il cDNA con le tre linee (sequenza genomica, parte codificante e proteina). Importante è scegliere il giusto genoma di riferimento dal quale sono partito (perché cambia il trascritto utilizzato).

Un database molto utilizzato è InSIGHT (https://www.insight-group.org/variants/databases) che contiene diversi varianti per diversi geni (ci segnala anche se l'effetto è patologico o incerto).
Altro database citato è IGSR (https://www.internationalgenome.org/) che è stato sviluppato grazie al progetto genoma umano.

Su Ensambl, se ricerchiamo la mutazione, sono già presenti i risultati delle analisi in silico dei principali siti web. Se sono contraddittori, è necessario verificare su altri software.

RIASSUNTO:

- La prima cosa da fare quando si ha una variante da un sequenziamento NGS o Sanger è consultare i database di popolazione per vedere se è stata ritrovata nel sequenziamento globale e a quale frequenza (se la frequenza è alta è un polimorfismo);
- Se non è presente, troviamo un database dedicato alla tipologia di malattia di nostro interesse;
- Se la ricerca non ci soddisfa, andiamo a condurre una analisi con i software in silico (almeno 3). Se i siti non sono concordi, bisogna fare delle analisi in vitro (prima bisogna capire dove è espresso per analizzare il campione corretto) e in vivo (in genetica è importante la segregazione della famiglia).