**Università degli Studi di Padova**
Corso di Laurea in Biologia Molecolare

Elaborato di Laurea

# Predizione *in silico* della predisposizione genetica alle dislipidemie

Tutor: **Prof. Silvio C. E. Tosatto**
Dipartimento di Biologia

Co-tutor: **Dott. Giovanni Minervini**
Dipartimento di Biologia

Laureanda: **Giada Trevisan**

Anno Accademico 2011/2012

## Abstract

Numerosi studi epidemiologici hanno dimostrato che elevati livelli di colesterolo rappresentano uno dei principali fattori di rischio per le malattie cardiovascolari. Nei paesi industrializzati, in particolar modo, si osserva un aumento di questo fenomeno, solo in parte spiegabile con il miglioramento delle condizioni alimentari. Le dislipidemie si presentano chiaramente come delle patologie multifattoriali, dove la componente genetica assume un ruolo molto importante. In questo elaborato sono state catalogate le mutazioni patologiche ad oggi conosciute per cinque dei principali geni coinvolti nella regolazione e nel metabolismo dei lipidi ematici (ApoB100, LDL-R, CETP, PCSK9, ABCA1). Le mutazioni sono state quindi organizzate in una vasta banca dati utilizzata come base di partenza per lo sviluppo di un algoritmo in grado di predire la predisposizione alla ipercolesterolemia. L'algoritmo è stato testato usando il genotipo di dieci volontari appartenenti al progetto Personal Genomic Project. I risultati ottenuti in questo lavoro hanno permesso di identificare correttamente la predisposizione alle dislipidemie per ognuno di essi. Le potenzialità dell'approccio usato nello svolgimento di questa tesi, ci permettono di immaginare un futuro prossimo in cui la diagnostica in silico permetterà di identificare correttamente le patologie ben prima che esse si presentino.

## 1. Introduzione:

Il miglioramento del tenore di vita degli ultimi decenni ha portato non solo ad un maggiore benessere, ma anche allo sviluppo di disordini legati al cambiamento dello stile di vita e della nutrizione. Alimentazione scorretta e ricca di grassi saturi, fumo e sedentarietà, ipertensione e obesità, hanno contribuito all'insorgenza di malattie cardiovascolari, così diffuse oggigiorno nei paesi industrializzati [1]. E' preoccupante notare che la principale causa di mortalità in questi paesi siano proprio le malattie cardiovascolari, causate da fattori, in parte direttamente correlati con l'alimentazione, quali alti valori ematici di colesterolo e trigliceridi. Numerosi studi clinici hanno confermato la relazione esistente tra la quantità di lipidi nel siero e il rischio di malattie cardiovascolari: in particolare è stato constatato che l'elevato livello di colesterolo LDL (lipoproteine a bassa densità) rappresenta uno dei principali fattori di rischio [1]. Secondo i dati pubblicati dal "Progetto Cuore" nel 2009, nella popolazione italiana tra i 35 e 74 anni, il 21% degli uomini e il 23% delle donne sono ipercolesterolemici. Il valore medio del colesterolo totale è circa di 205 mg/dl, mentre il valore per la HDL-colesterolemia è di 49 mg/dl negli uomini e 59 mg/dl nelle donne. I livelli raccomandati dal Ministero della Salute sono mostrati nella seguente tabella: [1]

| | Alto rischio (mg/dl) | Borderline (mg/dl) | Desiderabile (mg/dl) |
|---|---|---|---|
| Colesterolo Totale | Maggiore di 239 | 190-239 | Minore di 190 |
| LDL | Maggiore di 160 | 100-159 | Minore di 100 |
| HDL (uomo) | Minore di 35 | 35-39 | Maggiore di 39 |
| HDL (donna) | Minore di 40 | 40-45 | Maggiore di 45 |
| Trigliceridi | Maggiore di 200 | 150-200 | Minore di 150 |

**Tabella 1**. Valori ematici di riferimento raccomandati dal Ministero della Salute italiano (modificati da http://www.farmacoecura.it/alimentazione/ colesterolo-alto-hdl-ldl-valori-sintomi-cura-e-dieta/ il 17 nov 2011).

È auspicabile raggiungere e/o mantenere i valori desiderabili di colesterolemia e trigliceridemia per la prevenzione delle malattie cardiovascolari.
Uno dei disordini più diffuso e conosciuto è l'**ipercolesterolemia**, uno stato di alterazione patologica in cui i pazienti presentano un eccesso di colesterolo sierico, sia totale che legato alle lipoproteine plasmatiche. Questa condizione di alterazione, se non adeguatamente trattata, si concretizza in un rischio per la salute in quanto porta ad una maggiore probabilità di sviluppare **aterosclerosi**, una patologia che consiste nella formazione di placche di colesterolo depositanti sulla parete delle arterie [1]. L'occlusione delle arterie, causata da questi ateromi, oltre a determinare una evidente riduzione del lume arteriolare, con conseguenti ripercussioni sulla circolazione sanguigna periferica, può dare origine a trombi e causare infarti ed ictus. Oltre ad una ipercolesterolemia esogena, indotta da una

inadeguata alimentazione, esiste anche un'**ipercolesterolemia familiare**, un disordine genetico ad eredità autosomica dominante, causata da difetti funzionali o strutturali del recettore per le lipoproteine a bassa densità (LDL), che portano al non adeguato "uptake" (letteralmente *assorbimento*) di colesterolo nelle cellule. Il conseguente aumento dei livelli nel sangue può essere dell'ordine di 3-4 volte e anche di più. Numerosi studi epidemiologici hanno dimostrato che elevati livelli di colesterolo rappresentano uno dei principali fattori di rischio per le malattie cardiovascolari [2]. La diagnosi precoce, il controllo del peso, l'attività fisica, la riduzione dello stress, il cambio di stile di vita, un'alimentazione povera di grassi saturi e le eventuali terapie farmacologiche possono riportare i livelli di colesterolo entro la norma [1]. Dobbiamo però tener conto che solo il 20% di colesterolo viene assunto attraverso l'alimentazione, mentre il restante 80% è prodotto per via endogena dal fegato. Pertanto il contributo della predisposizione genetica può essere anche molto rilevante. Per questo motivo lo studio dei fattori non modificabili, come le mutazioni all'interno dei geni coinvolti nel metabolismo del colesterolo, può essere una strategia per attuare una terapia specifica per il trattamento. In questo elaborato prenderemo in considerazione cinque geni coinvolti nel metabolismo del colesterolo. Le dislipidemie primarie, condizioni cliniche su base genetica in cui sono presenti elevate concentrazioni di lipidi nel sangue, possano essere causate da mutazioni in decine di altri geni del pathway del colesterolo.

### 1.1 Gene ApoB100

Un gene strettamente correlato alle lipoproteine a bassa densità è il gene per l'ApoB100, la principale apolipoproteina sulla superficie delle LDL, che permette il loro assemblaggio a livello del fegato e la cessione del loro contenuto alle cellule. La proteina, infatti, funziona da ligando per il recettore LDL-R e permette l'endocitosi delle LDL [3]. Il fatto che sia presente soltanto una molecola di ApoB100 per ogni particella lipoproteica, porta alla conseguenza che il numero di lipoproteine prodotte dal fegato dipenda linearmente con la quantità di proteina sintetizzata. In questo modo alti livelli di ApoB100 portano ad elevati valori di LDL, correlati positivamente al rischio cardiovascolare e alla formazione di placche aterosclerotiche. **L'ipercolesterolemia familiare (FH)**, non è quindi causata solo da mutazioni difettive nel gene per LDL-R ma può essere determinata anche da variazioni nella sequenza genica per ApoB100. La sovraespressione della proteina conduce ad una maggiore biosintesi di particelle a bassa densità, contribuendo all'aumento dei livelli ematici di colesterolo LDL [2]. Anche difetti di interazione tra LDL-R e ApoB100, che diminuiscono l'affinità di legame, sono responsabili della non adeguata endocitosi delle lipoproteine [4]. Questa forma di ipercolecolesterolemia, chiamata **ipercolesterolemia familiare di tipo B**, risulta meno severa rispetto a quella causata da difetti del recettore, e ha un minore effetto sul rischio cardiovascolare.

La patologia complementare è l'**ipobetalipoproteinemia (FHBL)**, un disordine genetico dominante causato da mutazioni non senso nel gene per questa proteina [4]. Pazienti presentanti questa condizione mostrano livelli molto bassi di colesterolo LDL e lieve malassorbimento dei lipidi, ma la peculiarità è la presenza nel siero di prodotti proteici di ApoB molto più corti, corrispondenti a forme troncate di ApoB100 [4]. Queste specie proteiche, prodotte da varianti alleliche, sono associate con la ridotta concentrazione di ApoB100 wt a catena integra disponibile per la biogenesi delle LDL.

### 1.2 Gene LDL-R

La principale patologia associata a questo gene è l'**ipercolesterolemia familiare (FH)**: le cause di questo disordine, come già riportato, coinvolgono mutazioni a carico del gene codificante per il recettore delle LDL (LDL-R), o altresì, a carico del gene per l'apolipoproteina ApoB100 [3]. Le mutazioni con perdita di funzione per LDL-R, infatti, non permettono l'adeguato "uptake" delle lipoproteine a bassa densità, con conseguente incremento nei livelli plasmatici. Questa condizione porta ad aumento del rischio di mortalità cardiovascolare e di aterosclerosi precoce. La funzione della proteina è di agire da recettore per le lipoproteine LDL, permettendone l'endocitosi da parte delle cellule dei tessuti extraepatici. Il legame tra l'apolipoproteina B100 e LDL-R promuove l'internalizzazione del complesso recettore-ligando attraverso una vescicola rivestita di clatrina, una proteina di rivestimento che si associa alla membrana vescicolare e porta alla formazione della cosiddetta "fossetta rivestita". La formazione di un endosoma e la conseguente fusione con un lisosoma permette di separare i recettori, che verranno riciclati, di degradare l'apoB100 e idrolizzare gli esteri del colesterolo rilasciando acidi grassi e colesterolo. [2]. Quest'ultimo verrà usato dalle cellule per rimodellare le membrane o per la sintesi di ormoni steroidei [5]. L'accumulo di quantità eccessive di colesterolo dall'endocitosi delle LDL inibisce la sintesi endogena nel fegato [2]. La tappa limitante della sintesi, in cui avviene la regolazione, è la prima reazione della via metabolica, e coinvolge la riduzione di HMG-CoA (un intermedio a sei atomi di carbonio formato dalla condensazione di tre AcetilCoA), a mevalonato, il precursore del colesterolo. Molecole strutturalmente simili all'enzima chiave (HMGCoA reduttasi) di questo passaggio enzimatico, come le **statine**, possono agire da inibitori competitivi, diminuendone l'attività e portando così ad abbassare il livello di colesterolo circolante e di conseguenza il rischio cardiovascolare. Attualmente le statine costituiscono la terapia farmacologica di elezione per il trattamento dell'ipercolesterolemia, in quanto generalmente ben tollerate, e comportano una riduzione sostanziale del rischio di eventi cardiovascolari [2].

### 1.3 Gene CETP

Altre mutazioni collegate alla precoce comparsa dell'aterosclerosi coinvolgono la proteina di trasferimento degli esteri del colesterolo (CETP). Questa proteina plasmatica ha il compito di facilitare lo scambio degli esteri del colesterolo e dei trigliceridi tra le diverse lipoproteine. Permette, infatti, il trasferimento dei TAG (triacilgliceroli) dalle VLDL e LDL alle HDL, e viceversa per gli esteri di colesterolo [5]. Mutazioni che portano alla sovraespressione o all'aumento dell'attività del trasportatore sono responsabili dell'incremento di rischio cardiovascolare e aterosclerosi in quanto riducono i livelli di HDL e aumentano i livelli di LDL e VLDL. Al contrario, polimorfismi che portano ad abbassare l'attività o la concentrazione nel siero sono responsabili di aumentare notevolmente la concentrazione di HDL **(iperalfalipoproteinemia)**, diminuire i livelli di LDL e contribuire ad un effetto antiaterogenico [5]. In particolare, il **polimorfismo I450V** in omozigosi (genotipo VV) è stato associato ad un'eccezionale longevità e al mantenimento delle funzioni cerebrali [6]. In questi pazienti si riscontrano lipoproteine HDL e LDL di dimensioni maggiori e bassi livelli di CETP. La spiegazione di come questa variante possa proteggere dalla demenza senile, causata da ridotto apporto di sangue alle arterie cerebrali, si basa sul fatto che lipoproteine di diametro maggiore sono benefiche in quanto hanno una minor probabilità di rimanere intrappolate nei vasi sanguigni, dove possono ostruire le arterie [6]. I numerosi studi sulle varianti alleliche hanno dimostrato che la riduzione dell'attività di questo enzima risulta in un aumento dei livelli di HDL e in una diminuzione dei valori di LDL. Pertanto sono stati sviluppati degli inibitori farmacologici specifici il cui uso porta ad incrementare il colesterolo HDL, a diminuire quello non HDL e ad inibire la progressione dell'aterosclerosi[5].

### 1.4 Gene PCSK9

L'omeostasi del colesterolo e l'internalizzazione delle LDL è regolata anche dalla proteina PCSK9 (pro proteina convertasi subtilisin/kexin di tipo 9), la cui funzione è di indurre la degradazione del recettore LDL-R sulla superficie delle cellule [4]. La proteina viene secreta nel plasma dal fegato e si lega al dominio EGF-A del recettore per le LDL e ne induce la degradazione attraverso un processo intracellulare non ancora ben noto in cui si pensa che sia coinvolto il sistema di ubiquitinazione [4]. Una minore disponibilità di recettore risulta in una diminuzione del metabolismo delle lipoproteine a bassa densità con conseguente incremento dei livelli plasmatici di colesterolo-LDL, che possono peggiorare un tratto ipercolesterolemico. Mutazioni "gain of function ", che aumentano l'attività proteasica o l'affinità per il ligando, prevengono l'assorbimento di colesterolo dalle cellule e possono essere responsabili della terza forma di **ipercolesterolemia familiare (FH3)**, condizione rara autosomica dominante caratterizzata da elevati valori sierici di colesterolo LDL [5]. L'inibizione della funzionalità di questo enzima, pertanto, può essere ritenuta un mezzo per abbassare i livelli di

colesterolo. Individui con severe mutazioni “loss of function”, che aboliscono la secrezione di questa proteina, non mostrano alcun fenotipo svantaggioso. È possibile pensare che terapie basate sull’uso di RNA antisenso possano bloccare l’espressione di questo gene [4]. Infatti, variazioni nella sequenza genica che danneggiano la normale funzione della proteina (es. mutazioni non senso) sono associati con la diminuzione dei livelli di colesterolo LDL **(ipocolesterolemia)** a causa dell’aumento della disponibilità di recettori sulla superficie delle cellule [4].

### 1.5 Gene ABCA1 (ATP Binding Cassette A1)

La concentrazione totale di colesterolo nel siero è determinata anche dalle lipoproteine ad alta densità (HDL) le quali hanno l’importante funzione di rimuovere il colesterolo in eccesso dai tessuti periferici, e riportarlo al fegato, dove verrà eliminato o riciclato. Queste lipoproteine si formano per aggregazione dei vari componenti: colesterolo, trigliceridi, apolipoproteine e fosfolipidi. In particolare il fegato secerne l’ApoA1 a cui si aggregano colesterolo e TAG dai tessuti circostanti, grazie al trasportatore ABCA1, e altre apolipoproteine dalle altre particelle, formando così una HDL nascente. La proteina ABCA1, localizzata sulle membrane cellulari di quasi tutti i tessuti, funziona come una pompa di efflusso di colesterolo nel pathway di biogenesi delle HDL e di rimozione lipidica[7]. Essendo la concentrazione di HDL inversamente proporzionale al rischio di malattie cardiovascolari, è auspicabile una concentrazione sierica elevata. Le principali patologie determinate da alterazioni dell’attività di ABCA1 sono la malattia di Tangier e l’ipoalfalipoproteinemia familiare. Sebbene questi due disordini abbiano dei tratti in comune poichè forme alleliche dello stesso gene, sono da mantenere distinti in quanto hanno opposta modalità di ereditarietà e differenti caratteristiche biochimiche e cliniche [7]. La **malattia di Tangier**, infatti, consiste in un raro disordine recessivo caratterizzato da livelli estremamente ridotti di HDL sieriche. Questa condizione risulta nell’accumulo di esteri di colesterolo nei tessuti e nella conseguente predisposizione a premature malattie cardiovascolari. La forma più comune di insufficienza di lipoproteine ad alta densità è **l’ipoalfalipoproteinemia familiare**, ad eredità dominante, che differisce dalla malattia di Tangier per l’assenza dei suoi sintomi caratteristici [7]. Nonostante siano numerosi i geni associati al pathway del colesterolo che possono influire sulle dislipidemie, la limitata durata del tirocinio ha permesso lo studio solo di un ristretto numero di geni. Tra questi, ApoB e LDL-R in quanto direttamente coinvolti e maggiormente studiati; PCSK9, CETP e ABCA1 come geni largamente riportati nelle pubblicazioni scientifiche in stretta associazione alle dislipidemie. In una futura ricerca è auspicabile allargare lo studio coinvolgendo tutti i geni implicati nel metabolismo del colesterolo per poter aumentare le proprietà predittive del metodo. Lo scopo della tesi è stato quello di elaborare un algoritmo che, basato su un database costruito sulla correlazione mutazione-fenotipo per i suddetti geni, sia in grado di predire dal genoma di un paziente il grado di predisposizione a malattie dislipidemiche. La possibilità al

giorno d'oggi di far sequenziare il proprio genoma può costituire un enorme vantaggio per l'identificazione delle cause genetiche di malattie a cui si è predisposti. Estendendo questo tipo di studio non solo ai geni coinvolti nei disordini dislipidemici, ma all'intero genoma, si potrebbe rivoluzionare la diagnostica medica identificando rapidamente attraverso l'uso di software, gli alleli che causano una determinata malattia. Questa metodica pertanto può costituire un potente mezzo da accompagnare in un prossimo futuro ai test diagnostici di routine.

## 2. Metodi

La prima procedura eseguita è stata la costruzione di un database, ad uso interno, di tutti le mutazioni e degli SNP (Polimorfismi a Singolo Nucleotide) individuati nell'uomo, per ognuno dei cinque geni trattati. Di particolare importanza è stata la ricerca, ove possibile, della frequenza della singola mutazione nelle diverse popolazioni, dell'eterozigosità (proporzione degli individui eterozigoti) e dell'esatto fenotipo ad essa associato. Per ogni singolo polimorfismo sono state inoltre collezionate informazioni riguardo: la localizzazione genica (eventualmente cromosomica o del messaggero), il tipo di mutazione nucleotidica e proteica (esempio: delezione di basi, mutazione non senso, mutazione riguardante i siti di splicing), quale esone o quale dominio ne è stato affetto, la patologia o il fenotipo correlato e il numero di identificazione univoco secondo dbSNP. La raccolta dei dati ha impiegato la consultazione dei principali database biologici internazionali quali : dbSNP come banca dati dei polimorfismi conosciuti (http://www.ncbi.nlm.nih.gov/projects/SNP/), Pubmed come archivio delle pubblicazioni scientifiche (http://www.ncbi.nlm.nih.gov/pubmed/), Uniprot come risorsa per le informazioni proteiche (http://www.uniprot.org/), SNPeffect per un ulteriore integrazione dei dati (http://snpeffect.switchlab.org/), PheGenI (http://www.ncbi.nlm.nih.gov/gap/PheGenI), OMIM (http://www.ncbi.nlm.nih.gov/omim) e Variation Viewer per ulteriore completezza delle informazioni (http://www.ncbi.nlm.nih.gov/sites/varvu). PheGenI è uno strumento per la correlazione fenotipo-genotipo che integra informazioni dagli studi GWAS (genome-wide association study) e dai dati presenti in NCBI, Gene, dbGaP, OMIM, GTEx e dbSNP. Gli utenti possono condurre la ricerca di un gene in base alla localizzazione cromosomica, al nome del gene o al fenotipo e scaricare i risultati completi di SNPs e associazioni con altri geni o altri fenotipi. Attualmente i termini di ricerca sono basati sul vocabolario MeSH (per l'indicizzazione degli articoli della letteratura scientifica). OMIM è il database delle malattie genetiche umane ad eredità mendeliana e contiene un sommario di informazioni riguardanti tutti i disordini genetici conosciuti associati a più di 12.000 geni. Come PheGenI anch'esso si focalizza sulla relazione tra geni e fenotipo, ed è validato dalle pubblicazioni scientifiche raccolte in Pubmed. È aggiornato giornalmente ed è liberamente consultabile. Per

OMIM e PhegenI la ricerca è stata condotta sia in base al nome del gene che ai fenotipi/patologie ad essi associate (esempio: “ABCA1” e “Tangier disease”). Le numerose informazioni estratte da ogni database sono state raffinate attraverso l’eliminazione delle voci ridondanti e l’integrazione con il maggior numero di dati disponibili di ogni singola mutazione. Data di aggiornamento dei database: marzo-aprile 2012, periodo in cui sono state condotte le ricerche.

La fase successiva è stata l’elaborazione dei dati di sequenziamento a nostra disposizione, ricavati dai genomi completi di 10 individui e pubblicati nell’ambito del Personal Genome Project (PGP) http://www.personalgenomes.org/. Lo strumento informatico che è stato utilizzato a questo scopo è il software “ANNOVAR” [8]. Questo programma è stato concepito come un mezzo per la manipolazione dei dati ottenuti dai sequenziatori di nuova generazione. È un software molto efficace per l’annotazione funzionale delle varianti geniche e l’utilizzo delle informazioni ricavate dai diversi genomi. Data una lista di varianti con il tipo di mutazione e la sua posizione nel genoma, esso è in grado di eseguire:

1. **Gene-based annotation:** identificazione degli SNPs che possono causare cambiamenti a livello proteico e quali aminoacidi ne saranno affetti. Gli utenti possono scegliere quale sistema di definizione usare tra RefSeq genes, UCSC genes, ENSEMBL genes, GENCODE genes, 1000 Genomes e altri.
2. **Region-based annotation:** identificazione delle varianti in specifiche regioni genomiche: domini conservati nelle varie specie, potenziali siti di legame dei fattori di trascrizione, regioni duplicate, siti di ipersensibilità alla Dnasi I e molto altro.
3. **Filter-based annotation:** identificazione delle varianti riportate in dbSNP, identificazione delle varianti più comuni (MAF > 1%) dai dati di 1000 Genomes, identificazione degli SNPs non sinonimo, predizione di patogenicità di SIFT.
4. **Altre funzionalità:** una qualsiasi combinazione delle funzionalità appena menzionate permette estrema flessibilità per ogni ambito di ricerca. Inoltre, l’utente può definire manualmente nuovi filtri personalizzati, utili ad evidenziare mutazioni di interesse.

Dai dati di sequenziamento di genomi o esomi, ANNOVAR genera un file Excel-compatibile con l’annotazione genica, cambiamenti aminoacidici, punteggio di SIFT (che quantifica la patogenicità di una variante), numero identificativo di dbSNP, frequenza allelica di 1000 Genomes e molte altre informazioni. L’efficienza di questo software permette di eseguire la “Gene-based annotation” in circa 4 minuti (computer desktop 3GHz Intel Xeon CPU, 8 Gb memory) e la procedura di riduzione delle varianti più comuni (modulo 3) in 15 minuti [8].

ANNOVAR è stato usato, nel nostro studio, per l'integrazione delle informazioni geniche ottenute dai dati di sequenziamento dei dieci genomi a nostra disposizione. Per la specifica analisi da noi condotta è stato utilizzato il modulo 1, in quanto le altre funzioni non erano adatte al nostro scopo. Il documento prodotto da ANNOVAR consisteva in un file arricchito con le informazioni geniche, per ogni genoma a nostra disposizione. Ad esempio in quale esone è presente quel polimorfismo, l'eventuale cambiamento nella sequenza proteica, la frequenza dell'allele e il numero identificativo secondo dbSNP. Il risultato di ANNOVAR è stato processato attraverso un software sviluppato all'interno del nostro laboratorio. I dati del sequenziamento sono stati elaborati in modo da essere rappresentati con una struttura "ad albero": l'intero genoma è stato dapprima suddiviso nei 23 cromosomi, ogni cromosoma è stato diviso nei vari geni, ogni gene è stato ripartito in regioni quali esoni e introni. Ogni variazione genica diversa dalla sequenza del genoma di riferimento è stata quindi collocata in una specifica posizione. Il passo successivo è stata l'elaborazione, con la stessa procedura, del database inizialmente costruito ("genoma artificiale malato"): ogni mutazione collezionata è stata rappresentata a livello di cromosoma, di gene, di esone o introne, ed esatta posizione, strutturando un albero per ogni mutazione. Il confronto per uguaglianza tra gli alberi del genoma e gli alberi del database ha permesso la ricerca delle sole mutazioni significative del genoma del paziente.

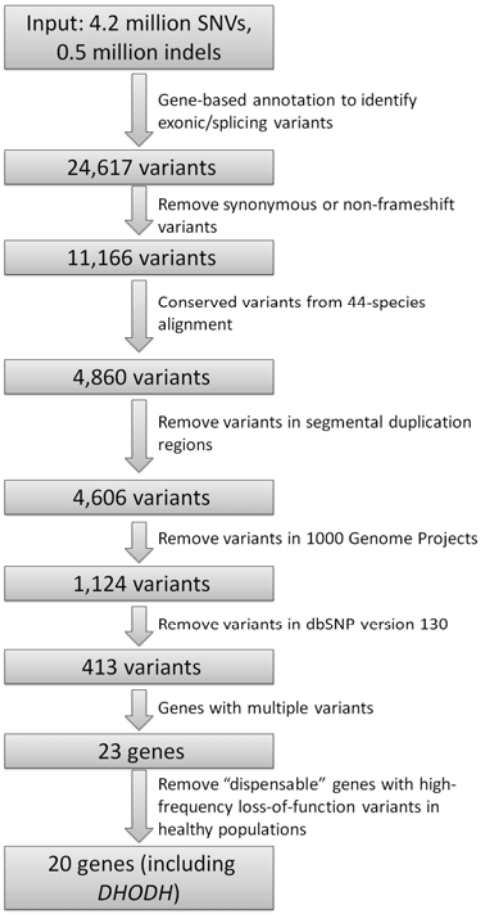


**Figura 1.** Esempio di flusso di filtraggio delle varianti effettuato dal software ANNOVAR

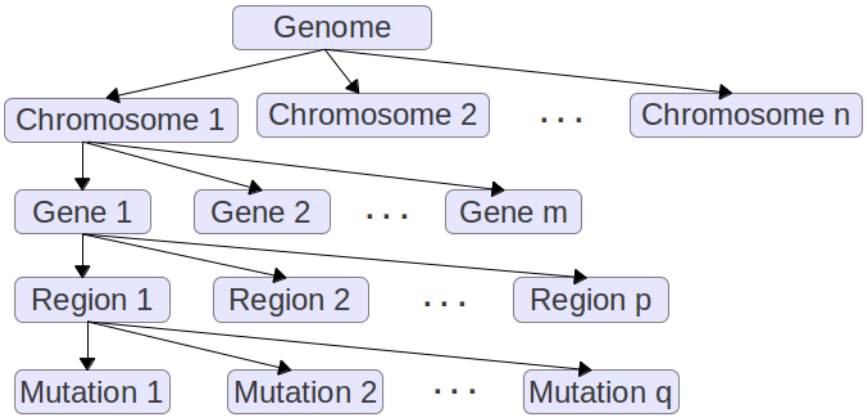


**Figura 2.** Esemplificazione dell'organizzazione dei dati nella memoria del calcolatore

La rappresentazione " ad albero" dei dati ha consentito di ottimizzare l'algoritmo di ricerca ed ha contribuito all'alta velocità ed efficienza del sistema, nonostante la lunga procedura iniziale per la sua costruzione. Infatti, un modello formato da una lunga lista di mutazioni sarebbe stato molto più semplice da costruire ma avrebbe richiesto altrettanti numerosi confronti e sarebbe stato perciò molto più lento. Più formalmente, la complessità della ricerca in una generica lista è O(*n*), dove *n* è il numero di varianti. Tale funzione, diventa O(*log n*) in una struttura albero. In altre parole, con questa particolare struttura dati si ha una relazione logaritmica tra il numero di mutazioni ed il tempo di calcolo. Per esempio, considerando i tipici 4 milioni di SNP in una persona:

- La ricerca su una generica lista richiederebbe 4 milioni di confronti nel caso pessimo
- La ricerca nella nostra struttura dati richiederebbe al più *log(4.000.000)* di confronti, cioè circa 7. Questo ha importanti ricadute nel tempo di esecuzione, che si riduce nella pratica di 5 ordini di grandezza, e rende lo strumento efficace nell'ambito dell'analisi di grandi moli di dati.

Dal confronto abbiamo ricavato il numero di mutazioni patogenetiche presenti in entrambi i dataset, e per ognuna è stato associato un peso a seconda della gravità della mutazione. È stato assegnato un maggior peso alle variazioni geniche correlate con l'aumento di colesterolo, alle mutazioni non senso, alle mutazioni frameshift, alle delezioni ed inserzioni estese, alle mutazioni con diretta associazione alle malattie cardiovascolari. Un peso minore è stato assegnato alle mutazioni sinonimo, alle inserzioni/delezioni non frameshift; mentre i polimorfismi senza correlazione di fenotipo sono stati eliminati in quanto di peso nullo. Le simulazioni sono state seguite su: HP G62, 450 SL, Core i3 2.26GHz, 6Gb RAM, con sistema operativo Linux Ubuntu versione 11.10, compilatore GCC versione 4.6.1. Il software di analisi è stato scritto in linguaggio C++.

## 3. Risultati

Nello svolgimento del presente lavoro di tesi sono state raccolte circa 50.000 annotazioni di mutazioni conosciute per i 5 geni oggetto di studio. In particolare, queste mutazioni sono state oggetto di una attenta revisione manuale al fine di scremare le numerose ridondanze presenti. I dati di partenza infatti, sono una collezione di dati preesistenti generati con le finalità più disparate e quindi non omogenei tra loro. Il primo risultato di questo lavoro è stata quindi una banca dati composta da 300 annotazioni uniche. Per ogni gene sono stati collezionati dati riguardanti la tipologia di mutazione, la variazione nucleotidica osservata, l'eventuale mutazione amminoacidica corrispondente e la frequenza associata. Sono state inoltre riportate le informazioni di localizzazione cromosomica nonché il codice identificativo dbSNP associato. Per ognuno dei 5 geni analizzati sono state create delle sotto tabelle nel quale è stato riportato il fenotipo associato. Per tutti i fenotipi catalogati si è correlato il grado di patogenicità; le mutazioni quindi, sono state catalogate come: neutre, predisponenti alla patologia e causanti la patologia. La seconda parte di questo lavoro, di tipo più strettamente informatico ed applicativo, ha visto la scrittura di un algoritmo per la determinazione e successiva assegnazione di un eventuale fenotipo patologico, partendo dal solo genotipo del paziente. Per l'esecuzione dell'algoritmo si è usata la piattaforma ANNOVAR. Maggiori dettagli sono presenti nella sezione 2 di questo elaborato. L'algoritmo è catalogabile come un "sistema esperto" ovvero un algoritmo che partendo da conoscenze pregresse e regole di associazione, simula la valutazione di un esperto umano. Entrando nel dettaglio, come secondo risultato di questa tesi, sono state create delle regole di associazione tra le mutazioni collezionate ed i fenotipi osservati. L'algoritmo quindi è stato istruito e testato su un campione di 10 genotipi pubblicamente disponibili e facenti parte del progetto internazionale "Personal Genome Project" http://www.personalgenomes.org/; in breve, questo progetto di ricerca mette a disposizione i genotipi volontari, non necessariamente affetti da patologie, per i quali è stato sequenziato l'intero genoma. I risultati ottenuti mostrano come l'algoritmo da noi sviluppato sia in grado di assegnare correttamente un profilo di rischio ipercolesterolemico partendo dal solo genotipo. Per i 10 genotipi disponibili sono state cercate eventuali mutazioni compatibili con quelle presenti nel nostro database. Alcune mutazioni sono risultate essere più comuni di altre. Una tale condizione ha suggerito che queste mutazioni fungano da fattore di rischio e che da sole non siano necessariamente predisponenti la patologia. Andando ad analizzare i singoli geni caso per caso, sono stati ottenuti i risultati mostrati in Tabella 2.

| | Genotipo | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Gene** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| ABCA1 | Sano | G656A | Sano | Sano | G656A | G656A | G656A | G656A | G656A | Sano |
| APOB | Sano | Sano | Sano | Sano | Sano | Sano | Sano | Sano | Sano | Sano |
| CETP | Sano | G1264A | G1264A | G1264A | Sano | G1264A | G1264A | G1264A | G1264A | G1264A |
| LDLR | Sano | Sano | Sano | Sano | C1236T | C1236T | Sano | Sano | Sano | Sano |
| PCSK9 | Sano | Sano | Sano | Sano | Sano | Sano | Sano | Sano | Sano | Sano |

**Tabella 2.** Nella tabella sono riportate le mutazioni predisponenti l'ipercolesterolemia, rinvenute nei 10 genotipi del progetto PGP attraverso l'utilizzo dell'algoritmo sviluppato in questo lavoro.

Le mutazioni a carico del gene ABCA1 risultano presenti nei pazienti 2, 5, 7, 8, 9. Nessuno tra i pazienti esaminati ha mutazioni per il gene APOB compatibili con quelle presenti nella nostra banca dati. In alcuni soggetti è però presente una mutazione rispetto al genoma di riferimento. Tale risultato indica quindi che sono presenti variazioni genotipiche del gene APOB ancora non catalogate. Una possibile interpretazione prevede che tali mutazioni siano silenti o che comunque non alterino la funzionalità della proteina codificata. Mutazioni a carico del gene CETP sono state evidenziate per quasi tutti i pazienti esclusi il numero 1 e 4. Similmente a quanto detto per il gene APOB, probabilmente si tratta di mutazioni molto frequenti che da sole non determinano l'insorgenza della patologia. Risultati più interessanti sono stati ottenuti per il gene LDLR. I pazienti numero 5 e 6 sono risultati portatori della mutazione nucleotidica C1236T. A livello amminoacidico la risultante mutazione Pro412Pro sembrerebbe indicare una mutazione silente. Dati bibliografici, invece, indicano questa mutazione come un possibile fattore di rischio. Al momento attuale non si è ancora correttamente interpretato il fenomeno [9]. Una possibile spiegazione prevede un'alterazione in qualche meccanismo di regolazione genica introdotto dalla mutazione. Nessuna mutazione significativa a carico del gene PCSK9 è stata evidenziata nei pazienti analizzati in questo studio. Per il paziente numero 5 sono presenti in bibliografia informazioni sullo stato di salute. In particolare il paziente è un soggetto che presenta ipercolesterolemia. La nostra analisi riconosce correttamente questo paziente come soggetto a rischio.

## 4. Discussione

Le displipidemie e le conseguenti patologie associate stanno aumentando proporzionalmente al benessere economico. La diagnosi clinica riveste quindi un ruolo fondamentale nel riconoscere i fattori di rischio come l'ipercolesterolemia. L'ipercolesterolemia, letteralmente, è una condizione in cui si osserva un eccesso di colesterolo nel sangue, risultato di una compartecipazione di fattori abiotici, quali un'alimentazione sbilanciata, e fattori biotici quali una predisposizione genetica. Gli attuali mezzi diagnostici possono riconoscere l'ipercolesterolemia

solo quando questa è già presente. Non hanno quindi caratteristiche desiderabili quali capacità predittive e non sono in grado di discriminare un paziente sano da un paziente potenzialmente a rischio. In questo lavoro di tesi, partendo da dati bibliografici riportanti le mutazioni genetiche per 5 dei geni coinvolti nella sindrome ipercolesterolemica, si è creato un metodo di predizione *in silico,* in grado di riconoscere la predisposizione genetica alla malattia. A questo fine, è stata organizzata una banca dati contenente un esteso numero di mutazioni dei geni ABCA1, APOB, CETP, LDLR e PCSK9; geni questi, direttamente associati alle displipidemie. E' stato quindi scritto un algoritmo e istruito con regole di associazione genotipo-fenotipo emerse durante lo svolgimento di questa tesi. L'algoritmo ottenuto, testato su 10 casi reali ha fornito interessanti risultati. In primo luogo, si è osservato che mutazioni su un singolo gene tra quelli presi in esame, non sono direttamente associate a fenotipi a rischio, ma che l'ipercolesterolemia è presente quando più geni sono contemporaneamente mutati. La componente alimentare ha comunque un suo peso. Il paziente numero 1, pur non presentando nessuna delle mutazioni oggetto di studio è risultato essere in trattamento con farmaci per il controllo del colesterolo. Analizzando però nel dettaglio il soggetto, si scopre che lo stesso dichiara un peso di corporeo che abbondantemente supera i 100 Kg ed un indice di massa corporea ben al di sopra dei valori consigliati. Il risultati ottenuti quindi, in linea con quanto già presente in bibliografia, confermano la multifattorialità della patologia dislipidemica. Un ulteriore risultato interessante ci è stato dato dalla presenza, in alcuni dei soggetti usati per testare l'algoritmo, di mutazioni inattese. In particolare, per il gene PCSK9 è emersa una mutazione presente in più pazienti, ma non elencata nella nostra banca dati. Al fine di spiegare questa anomalia è stata condotta una ricerca bibliografica da cui è emerso che al momento di scrittura di questo elaborato, tale mutazione non risulta essere annotata tra i fattori di rischio. Al fine di poter valutare l'impatto di questa mutazione in assenza di dati bibliografici, in futuro pensiamo di introdurre dei sistemi automatici di predizione degli effetti degli SNP. Ad esempio, PolyPhen (genetics.bwh.harvard.edu/pph2/) è uno degli strumenti più noti in letteratura in questo ambito. Inoltre, possono essere integrati anche sistemi per la valutazione del cambio della stabilità strutturale delle proteine come FoldX [10].

In generale quindi, l'approccio *in silico* mostra delle buone capacità predittive. Da un altro punto di vista, si potrebbe argomentare che il limitato numero di pazienti sui quali è stata condotta l'indagine, non sia sufficiente per confermare i risultati ottenuti. La ridotta disponibilità di interi genomi pubblicamente accessibili è senza dubbio un fattore limitante, ma gli incoraggianti risultati ottenuti in questo lavoro lasciano ben sperare su di un possibile utilizzo di algoritmi come quello sviluppato in questa tesi, come supporto alla decisione umana nell'ambito diagnostico e nella medicina preventiva.

## 5. Referenze


1. Ministero della Salute **"Linee guida per la prevenzione dell'aterosclerosi"** (2004) www.salute.gov.it/imgs/C_17_pubblicazioni_1097_allegato.pdf - www.cuore.iss.it/fattori/colesterolemia.asp

2. Joseph L. Goldstein and Michael S. Brown (2009) **"History of Discovery: The LDL Receptor"** Arterioscler Thromb Vasc Biol. 29(4): 431–438. doi: 10.1161/ATVBAHA.108.179564

3. Tremblay AJ, Lamarche B, Ruel IL, Hogue JC, Bergeron J, Gagné C, Couture P. (2004) **"Increased production of VLDL apoB-100 in subjects with familial hypercholesterolemia carrying the same null LDL receptor gene mutation."** J Lipid Res. 45(5):866-72.

4. Angelo B. Cefalù- Salvatore Amato - Emanuela Fertitta - Francesca Fayer – Vincenza Valenti - Maria C. Gueli - Ugo Di Blasi - Michele Pagano – Isabella Nardi - Gaspare Cusumano – Paolo Gulotta - Alessandro Raffa – Tiziana Doveri - Maurizio R. Averna (2007) **"Il gene PCSK9: un nuovo gene implicato nel controllo della colesterolemia"**. Acta Medica Mediterranea, 23: 11

5. Oliveira HC, de Faria EC (2011) **"Cholesteryl ester transfer protein: the controversial relation to atherosclerosis and emerging new biological roles."** IUBMB Life. 63(4):248-57. doi: 10.1002/iub.448.

6. Barzilai N, Atzmon G, Schechter C, Schaefer EJ, Cupples AL, Lipton R, Cheng S, Shuldiner AR. (2003) **"Unique lipoprotein phenotype and genotype associated with exceptional longevity."** JAMA. 15;290(15):2030-40

7. Stefková J, Poledne R, Hubácek JA. (2004) **"ATP-binding cassette (ABC) transporters in human metabolism and diseases. "** Physiol Res.;53(3):235-43.

8. Kai Wang, Mingyao Li and Hakon Hakonarson. (2010) **"ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data"** Oxford Journals, Life Sciences, Nucleic Acids Research, Volume 38, Issue 16, Pp. e164

9. Yandell M, Huff C, Hu H, Singleton M, Moore B, Xing J, Jorde LB, Reese MG. (2011) "**A probabilistic disease-gene finder for personal genomes."** Genome Res. 21(9):1529-42.

10. Schymkowitz J. W., Rousseau F., Martins I. C., Ferkinghoff-Borg J., Stricher F., Serrano L. (2005) **"Prediction of water and metal binding sites and their affinities by using the Fold-X force field."** Proc Natl Acad Sci USA, vol 102, p 10147-52.

**INDICE**