Relazione sul paper di BML

Introduzione

I dati ottenuti dalle cellule tumorali mettono in risalto le varie anormalità genetiche che si sono accumulate durante la progressione della malattia dal tessuto sano. Queste situazioni si generano progressivamente, con una continua acquisizione guidata, tra le altre cose, anche dal benessere della sottopopolazione così creata, che fa sì che solo una piccola parte di questi eventi riesca a venire fissata nel genotipo. Un fenomoeno frequente in questo contesto è chiamato **epistasis**, consistente nell'interazione fra geni distinti, che può contribuire all'evoluzione molecolare limitando e organizzando l'acquisizione di mutazioni.

La **funzione di benessere** (fitness function) detta anche **panorama** (landscape) è lo spazio di tutti i genotipi dipendenti sia dall'importanza che dal comprtamento delle interazioni epistatiche fra i geni. I genotipi tumorali sono quindi il risultato delle interazioni fra diversi percorsi evolutivi che si distibuiscono su un panorama complesso e i pattern che ricorrono fra le mutazioni somatiche contengono informazioni su entrambi questi elementi.

Le metodologie attuali (prima di BML, 2014) che cercano di estrapolare queste informazioni sono molto complesse computazionalmente e o non permettono di prendere in considerazione tutti i dati disponibili e sono costretti a operare su un sottoinsieme degli scenari possibili, o sono utilizzabili esclusivamente su piccoli gruppi di geni.

Approccio

Il metodo presentato si chiama **BML** (Bayesan Mutation Landscape) e ha la caratteristiche di prendere in considerazione anche quei genotipi ancestrali sconosciuti che precedono la condizione attuale del tumore prima di andare ad inferire i **percorsi di progressione evolutiva** (EPP in inglese). Gli stadi ancestrali, la cui non considerazione è un problema intrinseco a questo tipo di analisi, vengono stimati da BML direttamente dai dati. Inoltre, in questo metodo vengono inserite delle ottimizzazioni algoritmiche che permettono di calcolare gli EPP anche dei più grandi dataset disponibili considerando tutti i dati (nel 2014).

BML è basato su un modello **probabilistico** dove ogni percorso evolutivo dal genotipo sano a ognuno di quelli tumorali ha probabilità non nulla. Indichiamo con $P(g)$ la probabilità che un genotipo g sia presente nell'evoluzione del tumore (probabilità evolutiva) e che raggiunga la fissazione. Per ogni genotipo questo valore è uguale alla somma di tutti i percorsi evolutivi passanti per lui e che terminano con un genotipo

tumorale (Dunque $P(s) = 1$ dove s il genotipo sano). Si osserva inoltre che $P(g)$ è approssimabile con il valore $n(g)/N$ dove con $n(g)$ indichiamo il numero di campioni che hanno g come stato ancestrale o corrente e con N il numero complessivo dei campioni. In questo modo $P(g)$ è influenzato dal panorama di benessere che si sviluppa attraverso gli stati ancestrali di g attraversati durante l'evoluzione somatica. Questa metodologia inoltre **assume** che **le mutazioni si verifichino una per volta** e quindi non è adatta alle mutazioni di tipo strutturale come le CNV, mentre è adatta alle SNV, piccole inserzioni e piccole delezioni. Inoltre viene **ignorato l'effetto delle mutazioni già presenti nella linea germinale**.

> Nota: forse si potrebbe lavorare per adeguarlo ai vari tipi di mutazione

Questo approccio cerca di stimare la probabilità evolutiva utilizzando un modello basato sulle **reti bayesiane**, che possono essere rappresentate con dei grafi orientati e aciclici (DAG). BML stima la probabilità dei percorsi evolutivi P anche per gli stati ancestrali basansdosi su probabili percorsi evolutivi illustrati con degli alberi biforcati. I nodi interni di questi alberi rappresentano i genotipi ancestrali e assieme a quelli tumorali vengono usati per il learning della rete bayesiana. Dato che i percorsi evolutivi non sono noti è necessario operarare sugli alberi e in questo caso è stato scelto il metodo chiamato **nearest nighbour interchange** (NNI) che permette così il computo di un ottimo locale a partire da un albero casuale. La rete bayesiana successivamente stimerà P (tranne che per un fattore di normalizzazione, che successivamente sarà impostato proprio per avere $P(s) = 1$ con s il genotipo normale) e questa stima verrà usata per calcolare i percorsi di progressione evolutiva più probabili e successivamente visualizzarli. Inoltre sarà possibile osservare la distinzione fra **interazioni epistatiche positive dirette e indirette** (le negative non sono considerate) dove quelle dirette saranno indicate da archi nella rete.

Learning della rete bayesiana e ricostruzione degli EPP

I Dati

I dataset su cui BML è stato testato sono diversi, disponibili pubblicamente e toccano diversi tipi di cancro, come quello del colon retto, il glioblastoma, quello ai polmoni, alle ovaie e al seno. I dati sono stati preprocessati prima dell'analisi per fare in modo di rimuovere samples anormali, selezionare solo i geni più mutati (tramite una soglia) e considerare solo le mutazioni compatibili con questo metodo. Da queste informazioni viene poi estratta una matrice mutazioni/campioni che ha in ogni posizione un 1 se il gene nel dato sample è mutato e 0 altrimenti. BML è inoltre compatibile con programmi esterni di selezione delle mutazioni per poter operare sotto criteri differenti. In assenza di informazioni funzionali, l'output di questo programma va letto in soli termini probabilistici, senza assegnare ruoli specifici ai percorsi evolutivi mostrati.

Il modello

Una **rete bayesiana** $B(G, \Theta)$ è definita da G un grafo orientato e aciclico (DAG) i cui vertici (V) sono variabili booleane aleatorie e $\Theta = \{\theta_C | C \in V\}$ un insieme di parametri che rappresentano le probabilità condizionali $Pr(C = c|\Pi_C = \pi) \equiv \theta_C(c|\pi)$, ove c è lo stato di una variabile e π il relativo stato dei suoi genitori (fra tutti gli stati dei genitori dei vari nodi, Π_C) in G. Si consideri quindi D la matrice di input descritta nella sezione "I dati" e sia V (della rete) l'insieme dei geni e S l'inseme dei campioni. La matrice può essere utilizzata per acquisire un numero di dati sufficiente al learning della rete effettuando conteggi del tipo $n_{c,\pi}$ dove sono contati il numero di campioni in cui $C = c$ e $\Pi = \pi$.

Per selezionare le strutture adatte alla rete viene usato il **BIC** (Bayesan Information Criterion):

$$logPr(D|B) = \sum_{C \in V} Fam(C, \Pi_C)$$

dove $Fam(C, \Pi_C)$ è il punteggio BIC per la famiglia $\{C, \Pi_C\}$ (geni e stato dei relativi genitori) che è dato da:

$$Fam(C, \Pi_c) = max_{(\theta_c)}(\sum_{\pi \in \Pi_c}(\sum_{c \in C}(n_{c,\pi} log[\theta_C(c|\pi)] - log(n)/2)))$$

Dal quale, tramite massimizzazione, è possibile, con un numero sufficiente di campioni da una rete bayesiana reale, ricostruirla.

Apprendere le distribuzioni di probabilità

Come già detto, per calcolare P è necessario considerare un dataset contenente sia i genotipi tumorali che quelli ancestrali. Per riuscire ad ottenere questi ultimi vengono costruiti degli alberi biforcati con il genotipo normale alla radice, i genotipi tumorali come foglie e i nodi interni con un genitore e esattamente due figli, che vengono poi usati per il learning della rete. Si consideri ogni nodo come una sequenza di valori booleani (1 mutato, 0 non mutato) che, essendo associati ordinatamente a ogni gene, rappresentano il genotipo di una data sub-popolazione tumorale o ancestrale; la radice è quindi di soli zeri.

Semplificazioni e assunzioni:

1. Le mutazioni sono **IRREVERSIBILI** ossia se vi è una mutazione in un gene di un nodo, allora questa comparirà in tutti i suoi discendenti
2. La probabilità degli stati mutati deve essere più piccola di quella dello stato normale (deriva dal modello utilizzato, è implementata facendo modo che il numero di samples che presentano una mutazione su qualche gene non superi la metà del numero di samples analizzati)
3. Si fa in modo che, dato un genotipo, la probabilità di accumulare una mutazione in un gene non aumenti perdendo una mutazione in un altro gene. Ossia che

$$Pr(C = 1|\Pi_C = \pi_a) \geq Pr(C = 1|\Pi_C = \pi_b) \forall (\pi_b | \pi_b \subset \pi_a)$$

con $\pi_b \subset \pi_a$ che va inteso indicando che tutte le mutazioni presenti in π_b sono presenti anche in π_a

Obiettivo

A questo punto il problema del learning della struttra è dato dal trovare T_* albero e B_* rete bayesiana che massimizzano la BIC score:

$$(T_*, B_*) = arg \max_{T,B} log(Pr(D|B))$$

con $D = T \cup O$ dove O sono i genotipi della matrice passata in input.

Algoritmo di learning della struttura della rete bayesiana

Dato un albero realizzato come sopra si ricerca la struttura ottimale della rete (DAG) tramite una metodologia detta **ordering-based search** (OBS) che inizializza un ordinamento delle variabili e obbliga ognuna di queste a scegliere i propri genitori esclusivamente dai precedenti nell'ordine. L'algoritmo successivamente ricerca nello spazio di tutti gli ordinamenti invertendo l'ordine di ogni coppia adiacente (per la ricerca nello spazio degli alberi viene usato un set di operazioni chiamato NNI).

> aggiungi schemino algoritmo

Algoritmo di learning dei parametri della rete bayesiana

Per apprendere i parametri della rete bayesiana viene usato un Dirichlet empirico a priori. Per ogni gene C , i parametri sono scelti come:

$$\theta(C = c|\Pi_C = \pi) = \frac{(n_{c\pi} + \alpha_c)}{(n_{0\pi} + n_{1pi} + 1)}$$

dove l'iper parametro α_c indica la frazione di campioni di D che hanno $C = c$ e $n_{c\pi}$ è il numero di campioni dove $C = c$ e $\Pi_C = \pi$

Ricostruzione dei EPP più probabili

Gli EEP sono ricostruiti basandosi sul fatto che gli antenati più probabili per un dato genotipo sono quelli con la più alta probabilità evolutiva. L'algoritmo procede considerando tutti gli stati con k mutazioni e ricostuisce la loro storia evolutiva collegando ogni genotipo al suo più probabile stato ancestrale secondo P. L'input è costituito da tre parametri ossia P, k e c dove $0 < c < 1$ rappresenta un parametro per una soglia che fa in modo di permettere la visualizzazione dei soli nodi (con x mutazioni) con probabilità superiore a $c * m_x$ dove m_x rappresenta la probabilità massima ottenuta considerando genotipi con x

mutazioni ($x \leq k$). L'algoritmo itera partendo da $x = k$ e termina quando $x = 0$ decrementando x di 1 ad ogni passo. (si noti che k è fissato a 3 al momento)

Fitness epistasis e la sequenza delle mutazioni

Le interazioni epistatiche contribuiscono sia alla distribuzione delle mutazioni nello spazio dei genotipi e sia alla probabilità evolutiva. **BML non segue alcun modello specifico** per le dinamiche evolutive. In ogni caso si sfruttano queste due definizioni nell'analisi:

1. *fitness epistasis positiva*: intesa come la maggior probabilità che due mutazioni A e B occorrano assieme anziche separatamente ($P(AB) \leq P(A)P(B)$).
2. *sign epistasis*: si ha che $P(A) >> P(B)$ ma, se A occorre con frequenza sufficiente e l'interazione epistatica è sufficientemente forte si ha che $P(A) \leq P(AB) \leq P(B)$ (ossia una mutazione che da sola non ha successo, lo ha se si trova assieme ad un'altra mutazione positiva)

In BML si sfrutta la stima di P per osservare quando il genotipo doppiamente mutante ha una probabilità evolutiva fra quella dei singoli mutanti, dato che questa condizione permette di dare un ordinamento univoco alla relazione di epistasis fra i geni considerati.

(es: si osserva che TP53 precede TTN)

Procedure di bootstrap

Il bootstrap parametrico richiede di effettuare il learning del modello e di simularlo al fine di generare nuovi dataset per il learning permettendo di stimare accuratezza e robustezza dell'algoritmo di learning. I campioni di bootstrap sono ottenuti simulando la rete bayesiana calcolata precedentemente e selezionandone solamente quelli che mostrano almeno una mutazione e dove tutte le mutazioni presenti sono quelle rilevate in almeno uno dei genotipi osservati.