Prefazione

La seguente trattazione presenta un progetto dedito allo studio dell'efficienza computazionale di cluster a basso consumo energetico adoperati per l'analisi biofisica. In particolare, la ricerca mira a dimostrare come l'utilizzo di macchine a minor dispendio energetico possano essere pi convenienti e potenzialemente pi potenti rispetto alle odierne macchine sfruttate nel ramo della ricerca biomedica e sanitaria.
Il processo attraverso il quale stato possibile strutturare una tale ricerca avvenuto concentrando gli interessi verso uno tra i sistemi pi moderni di ricerca delle mutazioni genetiche causanti varie tipologie di tumori: il sistema GATK-LOD$_{\text{n}}$. L'implementazione di una componente di questo metodo in un innovativa versione della nota utility Make, e quindi la creazione di un nuovo eseguibile, ha permesso una gestione pi libera dei singoli passaggi del suddetto sistema.
Questo nuovo strumento, denominato Snakemake, in grado di organizzare i diversi compiti del metodo biofisico su diverse macchine conservando il corpo unico del programma. Proprio questa caratteristica stata sfruttata per verificare la fruibilit di insiemi di computer che collaborano come un solo apparato: i cluster.
La formazione di gruppi di computer nasce dalla necessit dei moderni organi di ricerca di potenziare le capacit computazionali, a cui deve essere anche alternato un contenimento dei costi sia economici che energetici. In questi tempi infatti, gli studi nel campo biomedico prevedono la collaborazione di professionisti in statistica dotati di computer ad alta potenza, che hanno il compito di fornire i risultati proficuamente e rapidamente. Lo sviluppo e il progresso nei vari settori ha necessitato, e necessita tuttora, di un contemporaneo aumento della potenza di computazione, il quale ha alcuni risvolti problematici. La crescita delle prestazioni dei computer ha come conseguenza di base un'inevitabile innalzamento dei costi di tali servizi e quindi una minor accessibilit alla maggioranza dei gruppi di ricerca. Questo per, non l'unico effetto negativo poich all'aumento della potenza segue un aumento del consumo energico, che non s un fattore trascurabile.
Queste ragioni hanno portato, negli ultimi anni, alcuni studiosi ad interessarsi a metodi alternativi per la computazione e in questa ricerca approfondito il tema riguardante le simulazioni su cluster a basso consumo energetico. L'idea di fondo il poter garantire ai ricercatori un risultato in qualit di tempi almeno pari a quello ottenuto con la metodologia tradizionale. Accompagnato per, dal vantaggio di consumare minor energia e spendere una quantit di fondi inferiore.
Nel primo capitolo saranno introdotti e approfonditi gli elementi cardine del

progetto a partire dall'esposizione del metodo GATK-LOD$_n$ sia nel funzionamento che nei risultati. Successivamente sar evidenziata la componente del metodo che implementata nell'ambiente di sviluppo del tool Snakemake e, quindi, saranno approfondite le capacit di questo strumento. Infine, sar specificato il significato di low power e saranno mostrate le macchine adoperate nell'analisi, per poi concludere con un accenno al funzionamento dei cluster.

Nel secondo capitolo sar spiegato dettagliatamente il funzionamento del programma con alcuni approfondimenti sui parametri utilizzati e verranno evidenziati i passaggi per un corretto uso del sistema, dall'installazione ai dati ottenuti. In pi sar specificato quale tipo di analisi stata compiuta su tali valori a disposizione.

Nel terzo capitolo verrano mostrati i risultati finali pi rilevanti con l'aggiunta di tabelle e grafici utili ad impreziosire le analisi e ad accompagnare l'esposizione. Inoltre sar presente anche un breve accenno ai dati non considerati proficui e a coloro che sono stati trascurati.

Per terminare sar presente un paragrafo destinato alle conclusioni, alle considerazioni finali e agli sviluppi futuri del progetto, tutto basato sugli esiti pi interessanti tratti dalle indagini svolte.

Capitolo 1

Introduzione

Questo capitolo introduce le componenti principali del progetto ed ha il compito di spiegare in maniera approfondita le operazioni svolte da esse. Per consentire una piena comprensione del sistema, tali componenti sono state raggruppate nei tre campi su cui stato condotto il lavoro.
Il primo di questi il campo bioinformatico e, quindi, un approfondimento del metodo GATK-LOD$_n$ riguardante la ricerca sull'origine dei tumori attraverso nozioni di bioinformatica. Inoltre sar sottolineata l'importanza della fisica nell'ambiente di lavoro e nelle procedure degli algoritmi.
Il secondo il campo di sviluppo informatico, ovvero la spiegazione del programma Snakemake scelto, le cui funzionalit concendono ampie possibilit sulla gestione delle risorse offerte dalle macchine.
Il terzo il campo dei dispositivi low power, accompagnati dalla descrizione di coloro che sono stati adoperati per la computazione.
La combinazione tra tali differenti ambiti ha quindi fornito i sufficienti strumenti per proseguire con la costruzione e il funzionamento del programma; i quali saranno affrontati nel prossimo capitolo.

1.1 La ricerca delle mutazioni genetiche

Questa sezione si occuper di dare un'idea generale su un ramo degli studi riguardanti le mutazioni genetiche che sviluppano e alimentano i tumori.
Negli ultimi anni l'indagine sulle forme cancerogene basata sulle variazioni che avvengono nel codice genetico ha suscitato sempre pi interesse e ci ha portato allo sviluppo di un discreto numero di programmmi, software e metodi atti all'analisi del DNA. Ognuna di queste applicazioni ha come fine la determinazioni di quelle mutazioni nei geni che comportano l'insorgere delle

malattie tumorali ora conosciute. La conoscenza di una tale correlazione di vitale importanza per la pianificazione di un piano di cura adeguato e, in altri casi, per un intervento anticipato in grado di prevenire il presentarsi della malattia.
Un altro aspetto delicato la risposta del cancro alle cure a cui sottoposto il paziente e che, in taluni casi, si concretizza in una forma di resistenza a questi interventi, come la resistenza alla chemioterapia. La gestione di una tale conseguenza pu essere aiutata dalla comprensione dell'origine genetica del tumore ed possibile quindi un agevolazione sulla scelta dei percorsi di guarigione pi opportuni.
Seppur avendo lo stesso obiettivo, gli innumerevoli algoritmi utilizzati spesso si ritrovano in contrasto tra di loro e queste discrepanze sono rilevate quando si procede con il confronto dei risultati finali, che spesso non concordano o sono proprio differenti. Questi sistemi lavorano i dati sperimentali grezzi attraverso vari processi che, generalmente, sfruttano svariati algoritmi di statistica le cui radici provengono dai campi di matematica e fisica applicata. La netta differenza di prestazioni, insieme alla diverse metodologie operate, lascia ai gruppi di ricerca un arduo compito nella scelta del metodo pi ideoneo da seguire.
In questo progetto il metodo considerato prende il nome di GATK-LOD$_n$ ed ideato dalla combinazione di due tra i tools pi comuni nel panorama bioinformatico: GATK e MuTect. Prima di esporre questo algoritmo, necessario definire il meccanismo generale con cui analizzato il materiale genetico nei tempi moderni e i motivi per cui esso il pi conveniente.

1.1.1 L'analisi del DNA

1.1.2 Il metodo GATK-LOD$_n$

L'ideazione di questo algoritmo dovuta ad un gruppo di ricercatori del Dipartimento di Fisica e Astronomia dell'Universit di Bologna nell'ambito dello studio sulla scoperta di polimorfismi somatici del singolo nucleotide nel sequenziamento dell'esoma. Precisamente, questo genere di polimorfismo denonimato con la sigla SNP, single nucleotide polymorphism, e indica quelle variazioni nei singoli nucleotidi che si verificano con frequenza significativa in una specifica posizione del genoma. In particolare, l'esoma comprende quelle regioni del genoma in cui sono codificate le istruzioni per l'RNA e per la sintesi delle proteine. Le mutazioni somatiche inoltre, hanno un ruolo chiave nella progressione della malattia e nella resistenza alla chemioterapia.
L'interesse nel proporre questo metodo nasce dal desiderio di poter predisporre di uno strumento che non si aggiunga al gruppo di software gi esistenti

bens che ottimizzi e potenzi alcuni tra questi. Perci, il team di studiosi ha considerato due applicazioni standard, GATK e MuTect, e li ha composti in modo da migliorare i prodotti finali di entrambi. Infatti alla completa esecuzione di GATK stato applicato una componente di MuTect, ovvero un classificatore Bayesiano conosciuto come LOD_n il cui scopo verificare un ulteriore volta i risultati ottenuti. I passaggi previsti dall'algoritmo sono vari e di seguito saranno esposti coloro ritenuti pi rilevanti per una spiegazione sufficientemente piena del medesimo.
Dopo aver raccolto i campioni normali e quelli per alcune specie di tumori attraverso specifiche metodologie sperimentali, essi sono stati sottoposti ad un controllo di qualit tale da rimuovere le letture considerate a bassa confidenza, cio non soddisfacenti una soglia minima predefinita. A questo punto, sono state applicati in successione i tools BWA-MEM e Picard, dove il primo allinea le reads e il secondo le ordina, indicizza e in pi ne marca i duplicati. Una volta completati sugli allineamenti una fase di riallineamento locale e di ricalibrazione sulla qualit delle letture, possibili grazie all'utilizzo di alcuni strumenti del Genome Analysis Toolkit, infine sono stati eseguiti GATK e MuTect per la ricerca delle varianti sui singoli nucleotidi(SNV).
La differenza procedurale tra queste due applicazioni che, mentre MuTect ritrova le mutazioni contemporaneamente tra i campioni normali e tumorali; GATK le chiama indipendentemente. Un'osservazione da sottolineare che i due metodi scovano pure diverse varianti che non sono condivise da entrambi, indicando la natura incompleta dei sistemi utilizzati. Un ulteriore distinzione tra GATK e MuTect data dal tipo di risultati raccolti poich, se il primo dotato di una maggiore sensibilit alle mutazioni, dalle 3 alle 20 volte superiore al secondo; quest'ultimo possiede una maggiore specificit degli SNVs.
Avendo GATK un elevato numero di falsi positivi nella chiamata alle varianti, stato aggiunto in fondo all'algoritmo il classificatore Bayesiano di MuTect per cercare di ridurre questo errore. Il compito del LOD_n consiste nel calcolare il rapporto tra due eventi probabilistici. Il primo che la mutazioni nel campione normale dovuta a rumori di fondi e quindi in realt non esiste. La seconda invece, considera il caso in cui la mutazione esiste nel campione normale ed dovuta ad una variante germinativa eterozigote. A questo punto, se il rapporto tra le probabilit (il Log Odds, da cui LOD) eccede un valore di soglia fissato, il classificatore definisce la variante come somatica.
Il gruppo di studiosi ha confrontato i prodotti finali dei tre algoritmi e ha ricavato che l'uso di GATK-LOD_n riduce notevolmente il numero di chiamate degli SNVs di GATK, mantenendo una sensibilit nettamente superiore a MuTect. Questa riduzione dovuta all'eliminazione di un sostanziale numero di falsi positivi, ovviamente dipendenti dalla tipologia di tumore. Perci, il vero successo il raggiungimento dello scopo prefissato, che consisteva nel dotarsi

di uno strumento pi performante di GATK e MuTect. Il miglioramento non si limita solo alla filtrazione dei falsi positivi ma anche al mantenimento della sensibilit di GATK e all'aumento della specificit. Infatti, GATK-LOD_n ha presentato frequenze di validit superiori a GATK e un miglior PPV, Positive Predictive Value, di GATK su differenti range di VAF, Variant Allelic Frequency. Le metodologie adoperate per confrontare la sensibilit e la specificit non sono state ritenuti utili allo scopo della presentazione e per questo non verranno trattate.
A posteriori delle verifiche sperimentali, GATK-LOD_n si rivelato uno strumento utile ad allargare le capacit di GATK e a scovare varianti non trovate da MuTect senza dover rinunciare alla specificit e sensibilit.
Visti i risultati positivi ricavati dall'algoritmo, il team di ricercatori fiducioso che un metodo di questo tipo possa aiutare a definire con maggior dettaglio le mutazioni somatiche di genomi cancerogeni, favorendo le valutazioni mediche e gli approcci sui percorsi di cura.
In modo da fornire una conoscenza pi approndita delle operazioni che avvengono nel sequenziamento, saranno esposti brevemente i concetti fondamentali alla base delle applicazioni utilizzate nel metodo GATK-LOD_n: GATK e MuTect.

GATK

MuTect