Bioinformatica

UniShare

Davide Cozzi
@dlcgold

Indice

Capitolo 1

Introduzione

Questi appunti sono presi a lezione. Per quanto sia stata fatta una revisione è altamente probabile (praticamente certo) che possano contenere errori, sia di stampa che di vero e proprio contenuto. Per eventuali proposte di correzione effettuare una pull request. Link: `https://github.com/dlcgold/Appunti`.

Capitolo 2

Introduzione alla bioinformatica

La genomica ha dimostrato negli ultimi anni ha dimostrato una capacità incredibile di produrre dati e questo ha portato alla nascita del bioinformatico, che diventa un esperto della gestione di questi dati sia dal punto di vista algoritmico che dal punto vista sistemistico.
A partire dal 2000/2001 ma sopratutto poco prima del 2010 si ha una crescita dei **dati genomici** non indifferente. I dati genomici sono quelli provenienti dal sequenziamento del DNA. Negli ultimi anni questa crescita ha superato la curva della **legge di Moore** quindi la crescita in termini di hardware (che si stima migliorare ogni 18 mesi) non riesce più a soddisfare la stima di richiesta di hardware necessario per il sequenziamento. Questa stima di sequenziamento è basata su Illumina, che produce le più diffuse macchine fisiche per il sequenziamento. Case farmaceutiche e laboratori che studiano il sequenziamento hanno almeno una macchina Illumina. La quantità di dati ha raggiunto i livelli dei petabyte e quindi ci si aspetta (e in parte già è così) che l'hardware non sia più in grado di elaborare tali dati.
La bioinformatica riceve quindi questa tipologia di dati. La bioinformatica è cruciale nell'ambito della ricerca in biologia molecolare (riguardante prettamente DNA), dove sempre più si ha necessità dell'appoggio dell'informatica, avendo a che fare con dati, nel dettaglio grandi dati.
Un altro aspetto è quello legato alle nanotecnologie e alla così detta **DNA-based computation**. Un esempio è legato al fatto che ormai si è in grado di manipolare il DNA al punto di essere in grado di assemblarlo in laboratorio, tramite un meccanismo a *tiling (tasselli)*, dove il tiling tendenzialmente è una figura regolare (triangolare, rettangolare, esagonale, etc...) con cui si compone del materiale biologico. Si riescono a fare letteralmente figure con il DNA (anche stelle, smile etc...) ma, sopratutto di questi tempi, vaccini,

che sono appunto manipolazione genetica di DNA o RNA. **Questa parte non è trattata nel corso**.

2.1 Breve introduzione biologica

Nel corso tratteremo prevalentemente sequenze di DNA. All'interno della cellula si hanno i **cromosomi** e un **genoma** altro non è che la collezione di cromosomi all'interno di un individuo. Il singolo cromosoma è rappresentato da filamenti di DNA "attorcigliati". Il cromosoma sostanzialmente è formato dalla coppia di due filamenti che si uniscono in una parte centrale detta **centromero**. I cromosomi, dal punto di vista informatico, sono vere e proprie sequenze (con i 4 nucleotidi, adenina, citosina, guanina e timina, ricordando la complementarietà delle basi A-T C-G), anche se si hanno varie regole per gestire questa "semplificazione". Un altro aspetto è il passaggio dal DNA alle **proteine**, anche se nel corso non verrà trattata la **proteomica**, ovvero lo studio delle proteine in se. In merito al passaggio da DNA a proteine si ha che il DNA contiene i **geni** da cui poi derivano le proteine. Un gene può portare a più di una proteina e questo si è scoperto grazie al sequenziamento. Allo stato attuale per "leggere" il DNA di un individuo dobbiamo passare per macchine di sequenziamento che però non possono leggerlo interamente ma, prendendo il DNA da una provetta (anche a partire da una singola cellula nel **sequenziamento single-cell**), si ha in output un file con dei frammenti del DNA originale, replicati in coppie, dette **read**. Tramite vari algoritmi siamo poi in grado di arrivare a capire e studiare il DNA per poi arrivare, si spera, ad uno dei principali fini della bioinformatica, quello di curare la vita, tramite terapie mediche (si parla di **medicina translazionale**, ovvero non curo un paziente tramite protocolli generali ma sulla base del DNA del paziente, che viene studiato ai fini di stabilire la migliore terapia, che diventa personalizzata per l'individuo). Le scoperte biologiche più attuali sono ottenute praticamente sempre grazie all'intervento anche dell'informatica e della bioinformatica.
Un esempio di uso delle sequenze è confrontare regioni genomiche di varie specie per valutare eventuali somiglianze. Un primo modo è diretto, un secondo è confrontare dopo l'allineamento, con l'inserimento di gap (studieremo la cosa nel dettaglio).
Il bioinformatico fornisce al biologo/biotecnologo la strumentazione necessaria per fare le varie analisi.

2.2 Progetto Genoma Umano

Un elemento chiave nella bioinformatica è il **Human Genome Project (*progetto genoma umano*)**, progetto partito prima del 2000 (la prima base è del 1990) con vari obiettivi:

- identificare tutti i circa 30.000 geni nel DNA umano
- determinare le sequenze dei 3 miliardi di coppie di basi chimiche che compongono il DNA umano
- memorizzare queste informazioni in banche dati/db
- migliorare gli strumenti per l'analisi dei dati

La bioinformatica è andata avanti quasi sempre con progetti globali e il Progetto Genoma Umano è stato il primo di questi progetti, diciamo che lì nacque la bioinformatica. Si hanno vari *milesstones*:

- *1990:* progetto avviato come sforzo congiunto del U.S. Department of Energy e del National Institutes of Health (NIH)
- *Giugno 2000:* completamento di una bozza di lavoro dell'intero genoma umano
- *Febbraio 2001:* vengono pubblicate le analisi della bozza di lavoro
- *Aprile 2003:* Il sequenziamento del Progetto Genoma Umano è completato e il progetto è dichiarato finito due anni prima del previsto

Quest'anno, nel 2020, è stato lanciato un progetto ulteriore in quanto ora si è anche in grado di sequenziare il DNA nei pressi dei **telomeri**, ovvero le terminazioni dei cromosomi, che sono le regioni più difficili da ricostruire tramite il sequenziamento. Per farlo si hanno algoritmi e software davvero molto sofisticati.
Vediamo qualche numero:

- il genoma umano contiene 3 miliardi (3×10^9) di basi nucleotidiche chimiche che sono 4:
  - adenina (A)
  - citosina (C)
  - guanina (G)

  - timina (T)

- il gene mediamente è composto da 3000 basi, ma le dimensioni variano molto, con il più grande gene umano noto che è la Distrofina con 2.4 milioni di basi
- il numero totale di geni è stimato a circa 30000, molto inferiore alle stime precedenti da 80000 a 140000 (in quanto prima c'era il dogma che un gene codificasse una sola proteina, e si avevano circa 140000 proteine, che si conoscevano anche solo per le analisi del sangue)
- quasi tutte (99.9%) le basi nucleotidiche sono esattamente le stesse in tutte le persone. Basta lo 0.01% di differenze tra basi per "fare la differenza", anche differenziando predisposizioni geniche per una certa malattia
- le funzioni sono sconosciute per oltre il 50% del gene scoperto

Vediamo anche qualche numero (in stima) in merito agli organismi più studiati dai bioinformatici (spesso organismi con poche basi), più l'attualissimo *sars-cov-2*:

organismo	numero basi	numero di geni
uomo (Homo sapiens)	3 miliardi	30000
ratto di laboratorio (M. musculus)	2.6 miliardi	30000
arabetta comune (A. thaliana)	100 milioni	25000
nematoda (C. elegans)	97 milioni	19000
mosca della frutta (D. melanogaster)	137 milioni	13000
lievito (S. cerevisiae)	12.1 milioni	6000
batterio (E.coli)	4.6 milioni	3200
Human immunodeficiency virus (HIV)	9700	9
sars-cov-2	$\sim$27 milioni	$\sim$15

2.3 Variazioni

Una volta conosciuta la sequenza dell'uomo si è cercato di studiare quello 0.01% di differenze tra vari esseri umani. Queste differenze sono dette **SNPs (*single nucleotide polymorphism*)** (*detti a voce "snips"*) che rappresentano la variabilità nella popolazione umana. Sono le differenze a livello di singolo nucleotide. Subito dopo il Progetto Genoma Umano è partito, sempre

tramite il National Institutes of Health (NIH), un progetto che confrontasse popolazione africana, asiatica e statunitense per calcolare queste differenze, individuate tramite tool informatici, tramite il cosiddetto **assemblaggio di aplotipi**, che è prettamente un problema informatico, *NP-complete*, la cui soluzione più recente è data da un **algoritmo parametrico**. Dagli aplotipi vengono estratti gli SNPs e questo sarà visto tra qualche lezione. Gli SNPs sono serviti a determinare differenze tra le varie popolazioni campione in merito, ad esempio alla predisposizione alla Talassemia nelle popolazioni mediterranee. Questi studi servono appunto capire le predisposizioni delle varie popolazioni. Se una popolazione ha, nella maggior parte dei casi, una certa base in una certa posizione allora si ha uno SNPs. Il famoso 0.01% forma questi SNPs, il 99.9% della popolazione porta il cosiddetto **allele di maggioranza** mentre lo 0.01% l'**allele di minoranza**.
Uno studio ha dimostrato che, in Italia, solo i Sardi hanno un profilo genetico ben definito, tutti gli altri sono dei "mix genetici" e questo si è scoperto studiando gli SNPs.
Dal Progetto genoma Umano si è poi passati a confrontare il genoma di piccolissimi campioni, ad esempio 1000 individui, con il 1000 Genomes Project, un altro progetto con sforzi internazionali, fatto per mappare le variazioni su una popolazione di 1000 individui. Si segnala che per sequenziare un individuo ci sono voluti 10 anni nel primo caso ma poi ci è voluto molto meno. Ora un singolo individuo si sequenzia in qualche ora, a costi molto ridotti.
Dal DNA si sono anche ricavati i flussi migratori avvenuti nel corso della storia.

2.4 Pangenoma

Si vedrà, durante il corso, che dire **il genoma è una singola sequenza**, è ormai sostanzialmente errato. Avendo sequenziato milioni di individui si parla di **pangenoma** e le analisi devono ormai essere fatte non su un singolo genoma di riferimento ma si usa quello abbinato a tutta la serie di 0.01% di SNPs individuati finora. Nel dettaglio un pangenoma è una collezione di genomi multipli che sono correlati tra loro (variando solo in pochi punti). Si ha il pangenoma dell'uomo, di un batterio etc...
Dal punto di vista informatico diciamo comunque che il DNA è una sequenza sotto l'assunzione della **complementarietà delle basi**:

- adenina e timina sono complementari
- citosina e guanina sono complementari

e questo mi permette di poter studiare solo uno dei due filamenti del DNA.

Esempio 1. *Sia data la sequenza:*

$$S = acctacga$$

la complementare è:

$$S' = tggatgct$$

Se prendo la sequenza (o meglio una porzione di essa) di S_1 di un individuo h_1 e la sequenza S_2 di un individuo h_2 avrò un alta somiglianza con eventualmente uno o più SNPs.

La posizione dello SNP è detto **locus**. Uno SNP si ha quando nel 99.9% dei casi tutti gli individui hanno una certa base in una data posizione, avendo l'*allele di maggioranza*, mentre lo 0.01% degli individui ne ha una diversa, avendo l'*allele di minoranza* (e lo rilevo confrontando una popolazione).

Esempio 2. *Si hanno:*

$$S_1 = acc\mathbf{t}acga$$

$$S_2 = acc\mathbf{g}acga$$

ho uno SNP nel locus 4. Ipotizzando che il 99.9% degli individui siano come l'individuo con la sequenza s_1 ho che la base t è un allele di maggioranza mentre la base g è un allele di minoranza.

L'uomo si dice essere **biallelico** in quanto le "opzioni" per una certa posizione sono solo due. Alcuni cambiamenti possono anche essere del tipo *inserzione/delezione* (anche per sequenze di più basi contigue), parlando di **variazioni strutturali** (che sono comunque più complesse e meno tipiche). Per rappresentare il fatto che si hanno più sequenze con queste variazioni, soprattutto se sono inserimenti e delezioni, ma considerando che il 99.9% delle basi è uguale (cercando quindi una rappresentazioni che ottimizzi questa cosa), rappresentando quindi un pangenoma, dal punto di vista computazionale è un **grafo**. Ogni sequenza identica collassa in un solo nodo, avendo poi singoli nodi per le variazioni.

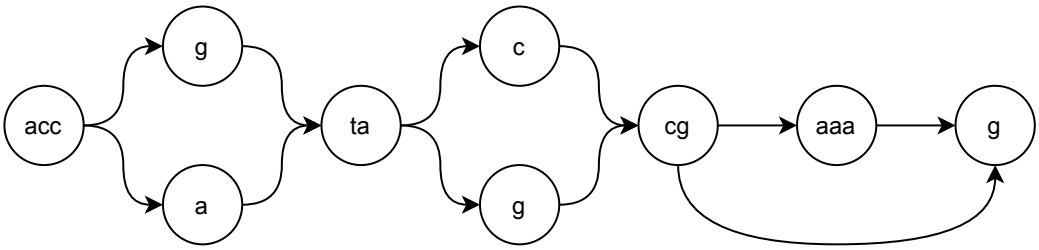
Esempio 3. *Ipotizzo di avere (con $-$ per indicare delezioni):*

$$S_1 = acc\mathbf{g}ta\mathbf{c}cg\mathbf{aaa}g$$

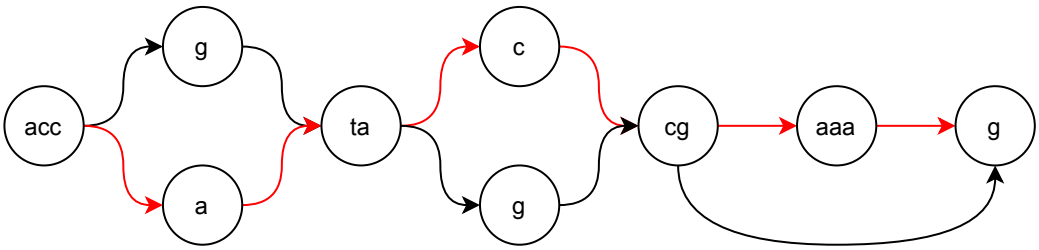
$$S_2 = acc\mathbf{a}ta\mathbf{g}cg\mathbf{aaa}g$$

$$S_3 = acc\mathbf{g}ta\mathbf{c}cg\mathbf{---}g$$

E ottengo un grafo del tipo:



Studiando i cammini dei grafi ottengo tutte le rappresentazioni.
Questa rappresentazione però ha dei difetti, in quanto potrei avere cammini che non rappresentano nessuna sequenza di partenza. Pensando all'esempio sopra potrei avere il cammino in rosso che non rappresenta nessuna delle tre sequenze:



Rappresento quindi più di quello che voglio rappresentare.
Un pangenoma è un grafo che rappresenta una popolazione senza fare grandi distinzioni, avendo percorsi che non sono riscontrabili in nessun individuo della popolazione. Si ha comunque che il concetto di sequenza non è più adeguato. Il grafo di una popolazione è enorme e comunque, tramite colori, si possono distinguere i vari percorsi della popolazione (distinguendo facilmente "tracce comuni"). Parlando quindi di **genoma di riferimento** o si parla di quello specifico di un individuo o si parla del pangenoma di una popolazione, con le varianti.
Dal punto di vista di *file* le varianti vengono date in un file **Variant Call Format (*VCF*)**. L'input classico dei software è quindi spesso un VCF, così come l'output.

2.5 Progetti attuali

Vediamo ora quali sono i grandi progetti su larga scala attualmente in corso:

- **The Cancer Genome Atlas Pan-Cancer Analysis Project (*TCGA*)**, che cerca di costruire un catalogo delle caratteristiche

genomiche dei tumori, ovvero un catalogo delle mutazioni genomiche associate a tumori (ad esempio quello del seno si sa che è legato alla mutazione del gene BRCA che si sa bene dov'è)

- **The 1000 Genomes Project Consortium: A global reference for human genetic variation**, che cerca di ricostruire e raffinare un sequenziamento di diversi genomi per costruire un genoma di riferimento per una popolazione, nel dettaglio umana, (in formato VCF)
- **Trans-Omics for Precision Medicine**, il progetto per la medicina translazionale
- **The Computational Pangenome Consortium**, che mira a studiare nuovi strumenti software che possano trattare il grafo del pangenoma visto che la maggioranza del software attuale ancora funziona su sequenze e non su grafi

2.6 Sequenziamento del DNA

Il sequenziamento (che letteralmente significa "produrre la sequenza") solitamente si svolge concatenando diverse operazioni:

1. estrazione del DNA
2. si ha una "libreria preparatoria" dove si mette del materiale genetico su un materiale preparatorio
3. si ha un meccanismo di "copie" tramite PCR o simili
4. si mettono i sample genomici in una macchina di sequenziamento che produce in output i dati

Un genoma non può essere letto "nucleotide per nucleotide" e i biologi, con la tecnologia attuale producono le cosiddette **read** del DNA originali. Si hanno due tipi di read:

- **read**, dette anche **short read**, lunghe circa 100 basi. Illumina produce tendenzialmente 100 o al più 150 basi
- **long read**, lunghe circa 10000 basi (se non di più, anche 20000)

Per ottenere il sequenziamento si ha un processo in cui:

- si divide il genoma in due parti, "aprendo" il filamento di DNA per permetterne la lettura
- si ha la **generazione delle read** da copie multiple del genoma tramite un processo biologico svolto dai macchinari, che sfruttano processi chimici
- si ha poi **l'assemblaggio dei frammenti**, ovvero un processo computazionale dove tramite algoritmi si assemblano le varie read per ottenere il genoma di partenza, avendo che le read hanno pezzi in *overlap*

Il problema del sequenziamento risale alla fine degli anni settanta con Sander e Gilbert che avevano studiato un processo di replicazione dando le basi allo studio del sequenziamento.
Dopo il sequenziamento dell'uomo si è passati a sequenziare molti altri organismi.
Oggi il sequenziamento è reso semplice dalla tecnologia. Un esempio è la tecnologia MinION, così piccola sta stare in una mano, che produce *long read* (anche se comunque con diversi errori). MinION è una tecnologia di *Oxford Nanopore*. MinION è USB ed è fatta per biologi che devono sequenziare in situazioni d'emergenza (esempio banale un biologo in Africa in piena emergenza Ebola). L'elaborazione dati viene fatta da un server.
Il primo sequenziamento è costato 3 miliardi di dollari per diversi anni, ora si fa in meno di 40 ore a 5000 dollari. Di recente si è passati addirittura a poche ore per un costo di circa 1000 dollari. Tornando alla *legge di Moore* si ha che il costo è collassato rispetto alla legge e quindi la capacità delle tecnologie di sequenziamento è molto maggiore della capacità di processare i dati, per quanto visto ad inizio capitolo. Si hanno quindi tanti dati ma non si è in grado di elaborarli.
Si tratterà anche il **confronto di genomi** per studiare poi gli aspetti evoluzionistici, tramite **alberi evolutivi**, anche **alberi evolutivi tumorali**. Il **confronto tra sequenze** permette di studiare le evoluzioni, anche quelle tumorali, dove si hanno mutazioni radicali di DNA. Approfondiremo anche tali mutazioni e il loro effetto (basta il cambio di una base per portare, ad esempio, all'anemia falciforme). Studieremo quindi anche come fare gli **allineamenti**. Approfondiremo il discorso della **filogenesi** e della **filogenesi tumorale**.
Tutto questo, in questo ultimo anno, è stato applicato allo studio di **sars-cov-2**, avendo lo studio delle variazioni.
Verrà approfondito anche il discorso del **riarrangiamento**.

Capitolo 3

Grafi di assemblaggio

La prima tematica che affrontiamo è l'assemblaggio delle read tramite grafi. Per questo problema abbiamo quindi:

- **input**: collezioni di read (short read e/o long read)
- **output**: grafo di assemblaggio da cui estrarre un cammino o un'unica sequenza

Si hanno principalmente due tipi di grafo:

- **grafo di De Brujin (*DBG*)** (*si legge "grafo di de broin"*), che si prestano più per *short read* (da 100 o 150 basi)
- **grafo di overlap**, più comodo in caso di *long read*

Si useranno per questi scopi varie nozioni, tra cui:

- relazione di prefisso/suffisso tra k-mers
- relazione di prefisso/suffisso tra read
- Longest Common Prefix tra sequenze
- estrazione di cammino di Eulero dal grafo
- estrazione di cammino Hamiltoniano dal grafo
- Maximal Exact Matches (*SMEMs*)
- Burrows Wheeler Transform (*BWT*)
- indici succinti (come FM-Index)

- suffix tree e suffix array
- bloom filters, nati in ambito fisico e usati ora in ambito BigData
- min-hash e min-sketch, usati anche nelle reti neurali e nel Deep Learning quando si ha a che fare con grandi moli di dati

Studiare i grafi di assemblaggio può essere utile anche in ottica di applicare procedimenti simili ad altri problemi posti dai biologi.