Laboratorio di Bioinformatica. Modulo 2

Domande in stile esame

Chiara Solito

Corso di Laurea in Bioinformatica
Università degli studi di Verona
A.A. 2021/22

Definizioni preliminari di ripasso

Annotazione Genica

Annotare un genoma significa conoscere la localizzazione, la struttura, la funzionalità di tutti gli elementi che compongono l'intero genoma. In pratica l'annotazione è quello che si fa dopo aver sequenziato un genoma, gli elementi che vengono annotati sono:

- Geni codificanti proteine
- Geni non codificanti proteine
- Elementi regolatori
- Elementi ripetuti
- Pseudogeni = geni che hanno perso la funzione codifica
- Altri elementi

L'annotazione può essere funzionale (consiste nel caratterizzare ogni singolo gene assegnando una funzione biologica ad ogni proteina da esso codificata) oppure genica (che definisce all'interno del genoma la localizzazione e struttura di ogni gene ed eventuali trascritti alternativi).

Genome Browser

Con il termine Genome Browser si intendono server con dei tools per la notazione automatica dei genomi, NON si intende una banca dati che raccoglie semplicemente informazioni.

Predizione Genica

La predizione genica è volta a trovare i geni di un genoma appena sequenziato. In realtà non si tratta di una vera e propria ricerca dal momento che non abbiamo una certezza assoluta, quindi occorrono dei confronti con le evidenze scientifiche. La predizione genica è alla base del processo di annotazione, poiché non posso definire il ruolo biologico di una molecola in tutta la sua complessità mappandola nel genoma se prima non conosco quale gene potrebbe essere. Parliamo però sempre di ipotesi che vanno confermate.

HMM

Gli Hidden Markov Models sono modelli probabilistici per dati sequenziali (temporali e non). Sono stati utilizzati molto a partire dal rionoscimento del parlato fino ad arrrivare ad una serie di applicazioni, in cui gli stati sequenziali potevano non essere così evidenti.
Si introducono a partire dai Modelli di Markov, definiti tramite 5 assunzioni.

1. Il sistema evolve in passi discreti.
2. Il sistema è in uno stato ad ogni istante di tempo.
3. Markovianità del primo ordine: il sistema non ha memoria, lo stato successivo dipende solo da quello corrente.
4. Modellazione probabilistica, ovvero la transizione tra gli stati è descritta in modo probabilistico.
5. Tutti gli stati sono osservabili.

La transizione tra stati viene definita tramite una matrice e le probabilità iniziali mi dicono come rimango negli stati o come passo da uno stato all'altro. La caratteristica principale però è che li stati siano osservabili e questo alle volte è limitante: per passare a un modello a stati nascosti bisogna rimuovere l'ulitma assunzione.
Negli Hidden Markov Models quindi intuisco lo stato dalle condizioni che contornano la situazione, dato che quello che osservo dipende dallo stato in cui mi trovo. Questo aggiunge un livello di incertezza: aggiungo quindi la probabilità che determinate osservazioni accadano in determinati stati.
Tecnicamente: in un modello di Markov se il sistema entra in uno stato si ha l'emissione di un solo simbolo; in un HMM se il sistema entra in uno stato si ha una distribuzione di probabilità che descrive la probabilità di osservare un determinato simbolo.
Queste caratteristiche fanno si che gli HMM possano essere utilizzati, nella "ricerca" di un gene, come "generatori di sequenze". Gli esoni e gli introni di una sequenza da modellare e poi da generare sono identificati da uno stato. La catena di acquisizione degli stati parte dal 5' fino al 3' in cui ogni base è generata grazie ad una matrice di emissione condizionata solo dallo stato corrente.

Variant Calling e Variant Analysis

Con Variant Calling (o Variant Analysis) identifichiamo il processo con cui una variante viene identificata a partire dalla sequenza e la successiva analisi della stessa, andando a stabilirne ad esempio la criticità. L'obiettivo è spesso identificare la variante responsabile di una malattia o di un certo fenotipo.

Protein Folding

Il protein folding è un ri-arrangiamento globulare e compatto della catena polipeptidica. Il modo in cui si ripiega la proteina è determinato dalla sua sequenza primaria. Tale orientamento punta a impacchettare i residui idrofobici all'interno del core della proteina.
L'equazione che regola tale funzionamento è data dalla seconda legge della termodinamica: $\Delta G = \Delta H - T\Delta S$ cioè l'energia libera di Gibbs è pari all'entalpia (sommatoria delle energie interne, ovvero le energie di legame) meno la temperatura moltiplicata per l'entropia (gradi di disordine del sistema). In questa

equazione occorre considerare anche l'effetto idrofobico (a favore del ripiegamento: ovvero il fatto che i residui idrofobici tendono a diminuire le interazioni con l'acqua), questo va a diinuire l'entropia dell'acqua, che forma delle gabbie ordinate attorno ai residui. Il risultato complessivo è che il folding di una proteina è marginalmente stabile poiché il grado di ΔG è dell'ordine di poche kcal/mol (l'energia di qualche legame idrogeno), il che costituisce il problema principale del cercare di simulare questo ripiegamento.
Ad oggi infatti non siamo ancora in grado di simulare il folding di una proteina al PC (solo un pc al mondo è in grado di simulare il ripiegamento di proteine molto piccole e compatte) poiché richiede un enorme potenza e precisione di calcolo. Quello che è possibile fare è ricreare il modello strutturale di una proteina tramite il Protein Modeling.

Homology Modeling

Con Modellazione per omologia intendiamo una tecnica di modellazione proteica che usi proteine omologhe a quella query. Per essere omologhe ovviamente le proteine devono avere un antenato comune: la proteina con struttura nota verrà chiamata proteina template e sarà usata per predire la struttura della proteina chiamata target.

Docking

Con Docking intendiamo lo studio delle interazioni ligando-proteina. Questa area della Bioinformatica è di interesse per la farmaceutica ad esempio, perché permette di simulare l'interazione tra il ligando e la proteina, migliorando eventualmente il ligando o trovando ligandi affini alla proteina.
L'obiettivo è, data una struttura proteica (cristallografica o modellata), predire quali ligandi essa lega e dove lega tali ligandi. La modellazione è necessaria nella maggiornaza di casi, e ha varie applicazioni:

- Predizione funzionale
- Disegno di farmaci, sostituendo ad un approccio brute force, un approccio razionale
- Studio dei meccanismo di interazione

Per arrivare all'obiettivo dobbiamo trovare una conformazione ligando-proteina tale che minimizzi l'energia totale del complesso. Tutto questo viene riassunto nel concetto di docking molecolare, ovvero tecniche di "attracco molecolare".

Domande tratte dai vecchi esami scritti - Teoria

Domanda 1

Descrivere il ruolo dell'entropia nell'interazione ligando-proteina.

Domanda 2

Descrivere il folding delle proteine. Si può provare a ripiegare una proteina al PC? Perché?

Domanda 3

Descrivere la predizione di geni in modo indiretto. Quali sono i "comunity experiment" della Gene Prediction?

Domanda 4

Descrivere i metodi ab initio della predizione della struttura delle proteine.

Domanda 5

Descrivere il termine di van der Waals per i campi di forza.

Domanda 6

Cos'è ENSEMBL? Descriverne il funzionamento.

Domanda 7

Descrivere gli elementi di struttura supersecondaria, fornendo alcuni esempi.

Domanda 8

Differenza tra profilo e profilo HMM nell'ambito della predizione strutturale di proteine.

Domanda 9

Definire l'annotazione genomica e descrivere i passi necessari all'annotazione di un genoma.

Domanda 10

Descrivere i passi necessari per costruire modelli per omologia della struttura delle proteine.

Domanda 11

Come funziona Modeller?

Domanda 12

Cos'è CASP?

Domanda 13

Cos'è un genome browser? Fornire un esempio, con la descrizione del suo funzionamento.

Domanda 14

Come avviene la validazione della predizione genica e dell'annotazione dei genomi?

Domanda 15

Descrivere i potenziali statistici utlizzati nei programmi di validazione della qualità della struttura di proteine.

Svolgimento

Dopo essere arrivati alla creazione di modelli di strutture proteiche, ho bisogno di validarne la qualità. Quello che voglio fare è eseguire più modelli per una stessa proteina e valutare dove le strutture si posizionano nel grafico dell'energia, la percentuale maggiore di modelli nel grafico, corrisponde all'energia minima di modello più probabile.
Per stimare la validità dei modelli utilizziamo i potenziali statistici: si parte da una valutazione dell'ambiente per ogni amminoacido, come essi interagiscono con i vicini e in generale il loro intorno. Per produrre un modello statistico devo tenere conto delle proprietà chimico-fisiche dell'amminoacido e le sue interazioni, confrontandole con quelle dei PDB da cui abbiamo tratto i template. Ad esempio possiamo prendere una specifica interazione, contare quante volte avviene e verificare che per quella famiglia sia statisticamente probabile che sia avvenuta e quante volte. Esistono diversi programmi che si occupano di questa cosa, uno famoso è QMEAN. Questi programmi confrontano sia graficamente che statisticamente i modelli, calcolando anche la root mean square deviation, ovvero lo scarto quadratico medio ottenuto dalla sovrapposizione tra le due strutture, ottimizzata, per poter calcolare la distanza minima tra i carboni α: questa infatti è una misura universale di similarità strutturale.
Successivamente al sueperamento della validazione statistica, procedo con un afunzionale e sperimentale: devo cioè verificare che la struttura che ho trovato risponda alle domande biologiche e alle funzioni che la proteina svolge.

Domanda 16

Descrizione dell'utilizzo degli algoritmi HMM. Caratteristiche, elementi necessari, ecc. (Focalizzarsi sul programma HHPred).

Domanda 17

Descrizione della "Modellazione per omologia". Descrivere il metodo "Modeling by satisfaction of spatial restraints".

Svolgimento

Per "modellazione per omologia" intendiamo una modellazione proteica (quindi della struttura sconosciuta di una proteina di cui conosciamo la sequenza) che si basa sulla conoscenza della struttura di una proteina ad essa omologa.
Per essere omologhe le proteine come sappiamo devono avere un antenato comune: se tali proteine condividono un passato evolutivo allora avranno probabilemente almeno il core proteico simile. La struttura nota viene definita in questo caso template, mentre quella da modellare è il target.
Velocemente abbiamo tre fasi generali per modellare una proteina per omologia:

1. Ho la proteina target e conosco la sequenza
2. Faccio il blast in alcuni database alla ricerca di proteine simili nella famiglia, che abbiano struttura nota.
3. Uso la proteina per modellare la struttura

Nello specifico però cosa faccio? Vediamo tutto passo-passo:

1. Cerco stampi nei vari database tramite BLAST e successivamente procedo con lo studio delle sequenze template
2. Procedo a un allineamento multiplo di template con il mio target, perché voglio modellarre le mie coordinate x,y e z e inserirle in una matrice di sostituzione. Qui si ha la maggiore fonte di errori: Estraggo il miglior allineamento a coppie trovato dopo l'allineamento multiplo (rimuovendo tutti i membri della famiglia - 1 ogni volta)
3. Prelevo le strutture dal PDB del template e procedo a costruire il modello Il modello può essere costruito in più modi, il più banale è quello di mettere aa solo nei match e nulla nei mismatch, uno molto usato è quello di mettere aa della query ma con coordinate random, che rispettino però i vincoli spaziali. Per rispettare tutti i vincoli è importante costruire un campo di forza: le strutture proteiche minimizzano l'energia libera, è necessario quindi scrivere una funzione matematica, un'equazione matematica, che è appunto il campo di forza, da minimizzare, che rappresenta l'energia libera del sistema, dalla quale dovrò estrapolare il minimo assoluto: essa rappresenta tutte le interazioni del sistema.
   La funzione nello specifico è: $V(R) = E_{(bonded)} + E_{(non-bonded)}$ dove, per i bonded intendiamo i legami covalenti: stretching, bending e rotation, per i non-bonded parliamo di legami non covalenti, ovvero potenziale di van der Waals e potenziale elettrostatico.
   Una volta ottenuta questa funzione dobbiamo estrapolarne il minimo: abbiamo tre metodi per farlo: uno è l'algoritmo di Montecarlo, che non calcola le derivate, ma è in grado di mantenere i cambiamenti che ci fanno arrivare al minimo della funzione, senza trovare l'ottimo ma una buona approssimazione. Gli altri due metodi invece calcolano le derivate: Steepest Descent ha un problema, che è quello di potersi potenzialmente perdere

nei minimi locali, perché non accetta passaggi per sezioni positive della derivata; al contrario Conjugate Gradient accetta passaggi positivi ed è in grado di trovare minimi non solo locali.

4. Passo infine alla fase di valutazione, che vedremo in seguito.

Un web server che usa questo tipo di procedura è Modeller, nello specifico Modeller effettua una tecnica chiamata "Satisfaction of Spacial Restraints"

1. Allinea la sequenza con le strutture di una famiglia
2. Crea una mappa di Vincoli Spaziali: questa si basa su più aspetti rispetto a quelli della modellazione che non usa questa tecnica: In generale dobbiamo tenere conto delle distanze Cα dal templato, degli angoli diedtri, delle catene laterali, delle interazioni tra i residui, della stereochimica del campo di forza e dei dati sperimentali.
3. Soddisfa i vincoli: in termini probabilistici crea o delle gaussiane o delle funzioni pseudo-molla.

Come risultato ho anche qui una funzione obiettivo, da minimizzare, che contiene le coordinate e tutti i vincoli. Gli approcci per minimizzare rimangono montecarlo o quelli che calcolano la derivata.
Modeller fa una modellizzazione naif ovvero fa un copia-incolla di tutte le coordinate uguali, lì dove non sono conservate e lì dove non sono conservati mette un copia-incolla del "Bubble" e negli indel mette coordinate casuali. L'output è un modello, con la valutazione stereochimica e i potenziali statistici che servono a validare il modello.
È importante invece tenere conto dei primi programmi sviluppati allo scopo di modellare, che sfruttavano Metodi AB Initio: Fragfold e Rosetta.
Fragfold cerca porzioni della sequenza della proteina che hanno le stesse strutture secondarie, valutando separatamente ogni frammento e poi modellandolo, facendo ottimizzazione e assemblamento dei frammenti.
Rosetta invece divide la proteina in frammenti da 9 aminoacidi, creando una banca dati con PDB da 9 amminoacidi. Da questo database trae i 25 frammenti con la sequenza amminoacidica più simile e li combina, anch'esso ottimizzando e minimizzando la funzione (modificando nello specifico gli angoli diedri per trovare l'energia minima). Rosetta ha rappresentato la prima rivoluzione nel mondo della modellazione di strutture proteiche.

Domanda 18

Cos'è un Campo di Forza? Spiegarne i modelli matematici e dove vengono utilizzati.

Domanda 19

Descrivere il metodo utilizzato nella validazione periodica nel campo della modellazione proteica.

Domanda 20

Descrivere il docking lingando-proteina: algoritmi, sfide, metodo della griglia, ecc.

Svolgimento

Con docking ligando-proteina intendiamo lo studio delle interazioni ligando-proteina, questa area della bioinformatica è di grande interesse ad esempio per la farmaceutica. L'obiettivo che ci si pone, data una struttura proteica, è di predire quali ligandi essa lega e dove, nella maniera più veloce possibile. Per farlo devo trovare la conformazione ligando-proteina che minimizza l'energia del complesso.
Sappiamo che ci sono delle configurazioni permesse e alcune che non sono permesse:

- Posizioni relative: 3 gradi
- Posizioni di orientazione: 3 gradi
- Legami ruotabili dei ligandi: N gradi
- Legami della proteina intera in soluzione: M gradi

Quando una proteina lega un ligando possono succedere due cose:

- Induced Fit: il ligando promuove un cambiamento conformazionale nella proteina.
- Conformational Section: la proteina campiona lo spazio conformazionale ovvero esplora diverse configurazioni.

Per cui quello che devo fare è:

1. Trovare metodi per campionare lo spazio conformazionale: un metodo famoso è quello della griglia, ma è molto lento, un algoritmo molto usato è il genetic search. Vediamo entrambi nel dettaglio.
2. Usare una funzione di scoring: basata su empirical function, che usa dati sperimentali, oppure knowledge-based ovvero basata su conoscienze a priori.
3. Infine effettuo un cluster dei risultati.

Il metodo della griglia si basa sulla discretizzazione dello spazio conformazionale in una griglia. Se conosco la posizione della cavità di binding posso limitare la superficie della griglia ad essa, prendo ognuno degli atomi del ligando e lo faccio passare per tutti i punti della griglia. Calcolo di ogni punto l'energia e creo delle mappe energetiche, individuando così i punti ad energia favorevole. Ricostruisco quindi il ligando nella conformazione migliore a partire da queste mappe, costruendo per ogni conformazione possibile, la somma delle energie dei punti

della griglia.
Come detto, il metodo della griglia è piuttosto lento, per questo si preferisce una tecnica di ricerca casuale di campionamento dello spazio: il genetic search algorithm. Esso crea tanti stati iniziali random, comincia ad incrociarli e da questi prende posizioni e features casuali. Da ogni stato posso produrre un discendente e valutarne l'energia, come per l'evoluzione. Incrocio e produco discendenti finchè non trovo un discendente che migliora l'energia. Ripeto questo procedimento iterativamemte.
Per verificare quali conformazioni sono possibili bisogna tenere conto delle forze che agiscono sul binding e creare delle scoring function: intra e intermolecolari: se ne creo di empiriche mi baso su lunghezze di legame, angoli di legame, angoli diedri, forze elettrostatiche, interazioni dipolari, ecc. I ligandi sono soggetti a forze intermolecolari deboli, tra le più importanti ci sono le forze columbiche di attrazione e repulsione, particolare attenzione viene riposta ad esempio sull'istidina che a seconda della protonazione ha carica negativa o positiva. Altre forze deboli importanti sono i legami idrogeno, i ponti salini, gli ioni metallici e le interazioni pi-greco (in cui distinguiamo tra end-to-face e face-to-face). Il campo di forza che vado a creare è un campo empirico, ovvero conosco tutte le foze ottenute da dati sperimentali.
Infine clusterizzo i risultati, in generale tenendo conto di almeno 100 conformazioni, per trovare la migliore.