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.

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. E qui parte tutto il discorso sulla funzione obiettivo ecc. ecc. Da questa funzione, minimizzando con gli algoritmi (Montecarlo oppure Steepest Descent o Conjugate Gradient) mi ricavo le coordinate e la configurazione con la quale la mia sequenza query raggiunge il minimo energetico.

4. Passo 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 (che anche qui ho un dubbio: ma se noi anche prima stiamo ai vincoli spaziali, perché non chiamare anche quella tecnica allo stesso modo? BOH. Poi: modeller è sempre modellazione per omologia quindi che differenza c'è esattamente?) 1. Allinea la sequenza con le strutture di una famiglia 2. Crea una mappa di Vincoli Spaziali 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. L'unica differenza di cui sono sicura è che: 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 (che a me sembra la stessa roba di prima anche qui, ma okay)). L'output è un modello, con la valutazione stereochimica e i potenziali statistici che servono a validare il modello.

POI Lui ha detto che non vengono più usate queste cose, ma una roba molto usata sono gli HMM. Questo perché un HMM può essere usato come modello che caratterizza sequenze di una certa famiglia proteica e quindi possono essere coinvolti nei processi generativi di predizione della struttura, massimizzando la likelihood. Infatti programmi come PFAM, HHSearch, SWISSModel e AlphaFold sono basati sugli HMM (anche se poi spiegando AlphaFold non è che abbia proprio parlato di HMM, perché dice che crea delle matrici di interazione e contatto tra amminoacidi, usando il templato per raffinare la mappa e poi inserendo i vincoli della funzione nella mappa stessa)

Vabbé poi ha parlato un po' dei Metodi AB Initio (Fragfold e Rosetta) ma che siamo sicuri che non si usino più tanto.

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.