Tema d'esame - Teoria

Rispondere in modo sintetico ai seguenti quesiti:

1) Descrivere brevemente le caratteristiche del Protein Data Bank ed il contenuto di un file .pdb

2) Spiegare la differenza tra omologia e similitudine. È possibile che due sequenze abbiano un'identità di sequenza del 57% e una similitudine del 21%? Perché?

3) Siano date le due sequenze di amminoacidi:
   s_1 =

   `MVALMGTKAADL`

   s_2 =

   `MVAIMASRAGEI`

   Allinearle senza inserire gaps e senza costruire alcuna matrice e rispondere ai seguenti quesiti.

   a) Stimare quantitativamente il grado di identita.
   b) Quantificare la similitudine utilizzando la matrice di punteggio data.

   Illustrare in ambo i casi il calcolo svolto.
   Cosa si può concludere circa l'omologia?

4) Che tipo di allineamento di sequenza si ottiene applicando gli algoritmi di Needleman-Wunsch e Waterman-Smith? Spiegare brevemente in cosa differiscono tra loro, e qual è la sostanziale differenza rispetto ai metodi euristici.

5) Che differenza c'è tra allineamento globale e locale di sequenza? Illustrare come si può quantitativamente valutare la significatività di un allineamento globale.

6) Che cosa s'intende per "twilight zone" nell'allineamento di sequenze proteiche?

Svolgimento

1) Il Protein Data Bank (PDB) è la principale risorsa, di riferimento mondiale, per le struttore 3D di macromolecole (proteine, super complessi e acidi nucleici). Le strutture all'interno del Database sono determinate tramite cristallografia, NMR e cryoEM (microscopia crioelettronica).
   I dati depositati non sono dati modellati, ma dedotti sperimentalmente,

ma le informazioni sperimentali possono essere utilizzate per modellare strutture non presenti nel database, ma affidabili. Ad oggi sono presenti più di 18000 strutture depositate. Il file di struttura (.pdb) è un formato testuale, identificato da 4 caratteri alfanumerici e se aperto con un visualizzatore molecolare permette di visualizzare la struttura 3D della macromolecola. Contiene la descrezione e l'annotazione di struttura di proteine e acidi nucleici tra cui: coordinate atomiche, rotameri di catene laterali osservati, assegnazione a particolari strutture secondarie e connettività atomica. Altre molecole comee acqua, ioni, acidi nucleici, ligandi e così via possono a loro volta essere descritti nel formato.
Vediamo la struttura del file che è composto da: intestazione e titolo, tecnica utilizzata e autori, remarks (osservazioni), sequenza, atomi con le loro coordinate atomiche (inclusi gli eteroatomi, essenziali per la cristallizzazione). Oltre ai file di struttura, nel database sono presenti informazioni sulle sequenze e molti strumenti per analisi di struttura, visualizzazione di ligandi e determinazione delle similitudini.

2) L'omologia è un carattere qualitativo, che indica che due entità (ad esempio due sequenze) hanno una stessa origine filogenetica, ovvero derivano da un antenato comune. Possiamo parlare di probabilità di omologia (non di grado) ma non stimarla del tutto quantitativamente. La similitudine è un carattere quantitativo, che indica la misura in cui due entità (ad esempio due sequenze) sono correlate, si basa su identità più conservazione. In questo caso posso parlare di grado di similitudine.
No, non è possibile che due sequenze abbiano un'identità del 57 % e una similitudine del 21 % perché essendo la similitudine basata su identità e conservazione, non è possibile che la similitudine sia inferiore all'identità: La percentuale di similitudine conta i residui "simili", oltre a quelli identici. La somiglianza tra gli amminoacidi può essere definita o dalle loro proprietà chimiche o basata su una matrice PAM, questo conferma che non può esserci un'identità maggiore della similitudine.

3) Le due sequenze hanno uno score di identità del 42% (abbiamo semplicemente calcolato il numero di matches fratto la lunghezza della sequenza più lunga) $\frac{5}{12} * 100 = 42\%$ mentre il similarity score è calcolato tramite la somma dei punteggi attribuiti tramite la matrice data (considerando le due sequenze), fratto lo score di identità della prima sequenza con se stessa (sempre calcolata con la matrice data) $\frac{38}{55} * 100 = 69\%$. Possiamo affermare che le due sequenze sono omologhe poiché hanno una s.i. molto maggiore del 20 % (soglia minima per stabilire una probabile relazione di omologia).

4) Entrambi gli algoritmi (Needlman-Wunsch e Waterman-Smith) sono utilizzati per ottenere un allineamento globale ottimo. La principale differenza tra i due è che nel caso di WS si può effettuare una ricerca di sottoallineamenti ottimi (quindi allineamenti locali) e si considerano negativamente gli indel (utilizzo un sistema di pesatura, in cui il peso w di

una indel di lunghezza k dipende dalla penalizzazione per l'apertura di una singola indel g e dalla penalizzazione per l'allungamento e). Utilizzando l'algoritmo di WS posso anche tenere conto della similitudine, perché la matrice è costruita a partire da PAM o BLOSUM (non hanno infatti solo zeri o uni, come in NW).
Questi algoritmi trovano sempre allineamenti ottimali, quindi sono precisi ma molto lenti, per questo sono stati definiti nuovi algoritmi di allineamento (metodi euristici) che hanno come punto di forza la velocità a discapito della precisione: questi trovano soluzioni approssimate ma vicine a quella ottimale, in tempi più brevi.

5) L'allineamento globale si estende da un capo all'altro delle due sequenze, considerandone l'intera lunghezza; quello locale trova le regioni di due sequenze che si allineano in modo ottimale: usando l'allineamento locale colgo la similitudine tra due sequenze senza tener conto di dove sono nella struttura primaria globale, in questo modo noto quantitativamente la similitudine. L'allineamento locale è infatti più utilizzato per le ricerche sui database, trovando domini o regioni limitate di omologia.
Per significatività statistica di un allineamento intendiamo la significatività di punteggi S trovati "casualmente": cerco, nel caso di omologie remote con punteggi di ordine di grandezza di quelle che posso trovare casualmente tra due sequenze, se sono significative o meno. Quantitativamente viene calcolato tramite Z-score (allineamento globale): $Z - score = \frac{S-\mu}{\sigma}$. Se Z-score è uguale o vicino allo zero, vuol dire che la somiglianza osservata non è migliore rispetto alla media di permutazioni casuali della sequenza. Più lo Z-score è alto, più non deriva dalla casualità ma da una reale situazione di omologia.

6) La twilight-zone è una zona, definita tramite PAM250 (matrice Pointed Accepted Mutation, in cui 250 sostituzioni si sono verificate tra due proteine su una lunghezza di 100 amminoacidi), che delimita una percentuale di identita (del 20-25 %) sotto la quale non è più possibile distinguere una similitudine, quindi non posso stabilire discendenza. Non sappiamo se la variazione è dovuta al caso oppure all'evoluzione.

Tema d'esame - Laboratorio

Esercizio 1

Utilizzando NCBI Gene, individuare l'entry relativa al gene che codifica per la proteina 'frequenin'. Rispondere alle seguenti domande, spiegando il procedimento seguito.

a) Qual è il nome del gene e il relativo id in *homo sapiens* e *drosophila melanogaster*?

b) Su quali cromosomi si trova nei due organismi di cui sopra?

c) Qual è il codice Uniprot della proteina espressa da questo gene nell'uomo?

d) Quali sono le principali funzioni molecolari della proteina espressa?

e) È possibile affermare che la proteina si trova nella membrana post-sinaptica? Perchè?

f) Quanti articoli PubMed sono collegati a questo gene nella specie *bos taurus* dalla pagina del database Gene?

Svolgimento Esercizio 1

Ci colleghiamo al sito di NCBI, e selezioniamo dal menù a tendina nella barra di ricerca il Database Gene: cerchiamo 'frequenin'.

a) Nell'"Advanced Research" aggiungiamo per il field "Organism" *Homo Sapiens*: il gene si chiama NCS1 - Neuronal Calcium Sensor 1, con ID: 23413.
Cerchiamo nello stesso modo per l'organismo *Drosophila melanogaster* in cui trova Frq1 (Frequenin 1 - con Gene ID: 32797) e Frq2 (Frequenin 2 - con Gene ID: 32799).

b) Per trovare su quali cromosomi si trova il gene nei due organismi guardiamo la sezione "Genomi Context": sia frquenina 1 che 2 nella Drosophila si trovano sul cromosoma X, nell'uomo si trova (NCS1) sul cromosoma 9.

c) Per trovare il codice Uniprot della proteina espressa nell'uomo, guardiamo la sezione "NCBI Reference Sequences" sotto la sezione "mRNA and Proteins": la proteina prodotta ha codice UniProt P62166.

d) Ci spostiamo ora su UniProt, visualizzando la entry trovata precedentemente, e guardiamo la sezione "Function": I sensori neuronali per il calcio, sono regolatori di fosforilazione di recettori accoppiati a proteine G, in maniera dipendente dal calcio. Regola direttamente GRK1 e può sostituire la calmodulina. Inoltre stimola l'attività chinasica di PI4KB ed è coinvolto nella plasticità sinaptica a lungo termine attraverso la sua interazione con PICK1.

e) Nella sezione "Subcellular Location", troviamo tra le note che la proteina è associata all'apparato di Golgi e si trova nelle densità post-sinaptiche dei dendriti, quindi possiamo affermare questa cosa solo per questo tessuto, infatti nelle giunzioni neuromuscolari, troviamo la proteina nel terminale nervoso pre-sinaptico.

f) Torniamo ad NCBI, ed effettuiamo la ricerca '(frequenin) AND "Bos Taurus"[Organism]'. Nella sezione "bibliography" troviamo 4 articoli principali, correlati nella sottosezione "Related articles in PubMed"

Esercizio 2

Ricercare in BLASTP le sequenze simili a NP_000781 nell'organismo *Danio rerio* (7955), ricercando nel dataabase di sequenze con codice RefSeq escludendo le sequenze modellate. Rispondere alle seguenti domande:

a) A quale organismo appartiene la sequenza di input? Che proteina è?

b) Quante sono le hits trovate e a quale superfamiglia appartengono?

c) Quante hits hanno score compreso tra 80 e 200? Qual è il loro codice RefSeq?

d) Qual è la hit (nome proteina e codice RefSeq) che rappresenta l'allineamento locale che ricopre la porzione minore rispetto alla lunghezza della query? Qual è l'E-value? Qual è l'identità di sequenza?

Svolgimento Esercizio 2

Ci colleghiamo a "Protein BLAST" di NCBI, nella sezione query inseriamo il codice RefSeq (NP000781), ricercando nella sezione "Choose Search Set" l'organismo *Danio rerio*, nella stessa sezione selezioniamo il database "RefSeq_Protein", con opzione flaggata "Models".

a) Nel Summary del risultato (sezione Description) troviamo che la sequenza di input appartiene all'organismo *Homo Sapiens* - anche cliccando sul codice RefSeq si apre la pagina Gene (NCBI) corrispondente. La proteina è la decarbossilasi-L-amminoacido aromatica (isoforma 1).

b) Nella sezione "Description" sotto il Summary troviamo che ci sono 6 hits, nella superfamiglia AAT I Superfamily (che troviamo guardando la sezione Graphic Summary, collegata alla pagina NCBI dei domini conservati).

c) Sempre nella sezione Description è indicato lo Score degli hits 4 dei 6 trovati hanno score compreso tra 80 e 200:

- NP_001017708.2 con score 153
- NP_001083039.2 con score 132
- NP_919400.1 con score 133
- NP_001007349.1 con score 125

d) La proteina che ricopre la sezione minore rispetto alla lunghezza della query (identificata tramite la sezione "Query Cover" sempre in Description) è uncharacterized pprotein LOC100038790[Danio rerio], (NP_001083039.2 con score 132) con coverage del 69%. L'e-value è $9 * e^{-34}$ e l'identità di sequenza è $26,19\%$

Esercizio 3

Cercare sul database UNIPROT la proteina OUTER MEMBRANE PHOSPHOLIPASE A.

a) Quale entries ci sono relative a *Escherichia coli* (qualunque ceppo)? Qual è la lunghezza della catena polipeptidica più frequente?

b) Selezionare la entry relativa al ceppo K12 (PA1_ECOLI). Con quale metodo sono state risolte le strutture tridimensionali? Qual è il file PDB a maggior risoluzione che risolve solo la catena A? Cosa si intende per catena? Qual è la percentuale di sequenza risolta in questa struttura?

Aprire il relativo file PDB con Pymol e rispondere a questi ulteriori quesiti.

c) Quante sono le α-eliche? Elencare i residui che complessivamente formano la struttura ad α-elica più lunga.

d) Sono presenti ioni Ca^{2+} o Mg^{2+}?

e) Che tipo di interazione stabilizza la coppia di residui Ile41 e Lys68? Indicare i gruppi chimici coinvolti e la loro distanza.

Svolgimento Esercizio 3

a) Cerchiamo su UniProt la proteina OUTER MEMBRANE PHOSPHOLIPASE A. Nella sezione di filtraggio (Filter by) c'è una barra di ricerca (Other Organism) dove cerchiamo: *Escherichia Coli*. Troviamo 161 entries in qualunque ceppo. La catena polipeptidica più frequente (fosfolipasi A1), ovvero la prima entry, è 289 AA.

b) La entry relativa al ceppo K12 (PA1_ECOLI) è proprio la prima: la selezioniamo. Nella sezione Structure, possiamo vedere i metodi con cui sono state risolte le strutture, che sono: X-Ray e Predicted con Alpha Fold. Ricordiamo che la risoluzione è tanto più alta, tanto più è basso il valore in Armstrong (errore associato agli atomi), in questo caso è 1QD5 con una risoluzione di 2.17 Armstrong. La percentuale di sequenza risolta in questa struttura è di 269 su 289 (è risolta da 21 a 289) quindi del 93%.

c) Apriamo ora il file della entry con Pymol (scaricabile dal link apposito laterale): Le α-eliche sono 4: la posizione di quella più lunga è in posizione 17-24 (colorati tramite color ss nei comandi principali): SIIANMLQ.

d) Proviamo a selezionare i due ioni (tra cui Ca^{2+} e Mg^{2+} tramite il comando `select elem Mg` e `select elem Ca`) ma la selezione non produce alcun risultato: possiamo quindi asserire che non ci sono tali ioni nella molecola.

e) Selezioniamo dalla sequenza Ile41 (I nella posizione 41 che è indicata) e Lys68 (K nella posizione 68 che è indicata), dopo di che nascondiamo tutto (Hide) e visualizziamo come sticks solo i due residui.
L'interazione che stabilizza sarà un legame idrogeno, i gruppi chimici coinvolti (che evidenziamo tramite il comando Label → atom name): sono CO-NH e NH-CO. Hanno entrambi distanza 2.8 Armstrong.

Esercizio 4

In seguito ad una variazione locale di pH la proteina P_1 suisce un cambiamento conformazionale che la rende attiva. La forma attiva della proteina (P_{1a}) è in grado di legare la proteina P_2 in modo reversibile con le costanti cinetiche $k_{as} = 1.4 \times 10^5 M^{-1}s^{-1}$ e $k_{dis} = 5 \times 10^{-2}s^{-1}$.
Sapendo che le concentrazioni iniziali della protina P_1 e della proteina P_2 sono rispettivamente $35\mu M$ e $24\mu M$ e che la costante cinetica di variazione conformazionale della proteina P_1 è $kconf = 7 \times 10^3 M^{-1}s^{-1}$ simulare il sistema dinamicamente e rispondere ai seguenti quesiti:

1. Sono sufficienti 0.2 secondi (200 ms) per stabilire l'equilibrio?
2. Assumendo che a t=20s il sistema sia all'equilibrio, determinare empiricamente la costante di equilibrio e confrontarla con la costante di equilibrio teorica.
3. Supponendo ora che la proteina P_{1a} sia soggetta a degradazione con una costante cinetica $kdeg = 1.2 \times 10^{-1} M^{-1}s^{-1}$, per quanto tempo nel sistema si può rilevare la presenza di P_3?

Soluzione

Dobbiamo usare il tool `IQM` - ogni volta lui ci chiede di installarlo: quindi usiamo il comando `installIQMtoolsInitial`. Per iniziare lanciamo il comando `IQMeditBC`. Ci si apre la finstra Complete Model View, con il modello delle varie parti inizializzato.
Facciamo uno schema:

```
*****************MODEL NAME

*****************MODEL NOTES

*****************MODEL STATE INFORMATION
P1(0)=35e-6   %M (moli/I)
P2(0)=24e-6   %M (moli/I)
```

```
*****************MODEL PARAMETERS
k_conf=7e3 \%M^-1 s^-1
k_as=1.44e5 \%M^-1 s^-1
k_dis=5e-2 \%s^-1
k_deg=1.2e-2 %aggiunto successivamente
                %all'esercizio 3

*****************MODEL VARIABLES

*****************MODEL REACTIONS
P1 => P1a : r1
   vf=k_conf*P1
P1a + P2 <=> P3: r2
   vf=k_as*P1a*P2
   vr=k_dis*P3
P1a => :r3     %aggiunto successivamente
   vf=k_deg*P1a    %all'esercizio 3

*****************MODEL FUNCTIONS

*****************MODEL EVENTS

*****************MODEL MATLAB FUNCTIONS
```

Risposte

Domanda 1 Per sapere se per la prima reazione sono sufficienti 0.2 secondi (200 ms) per stabilire l'equilibrio, basta guardare il plot simulato per il tempo necessario. Per la prima specie è sufficiente, ma per la seconda no. Né P_3 né P_{1a} hanno raggiunto l'equilibrio. Già dopo 10 secondi invece è visibile l'equilibrio raggiunto da tutte le specie.
Quindi 0.2 secondi non sono sufficienti per stabilire l'equilibrio dell'intero sistema.

Domanda 2 Dal grafico traiamo empiricamente i valori di $P_{1a}(20)$, $P_2(20)$ e $P_3(20)$ (riportate sotto)

$$k_{eq} = \frac{P_3}{P_{1a} * P_2} = 6.8 \times 10^6$$

$$K_{as} * P_{1a} * P_2 = k_{dis} * P_3$$

Dobbiamo ora calcolare la costante di equilibrio empirica:

$$P_3(20) c.a. = 2.3x10^{-5}$$

$$P_{1a}(20)c.a. = 1.13 \times 10^{-5}$$

$$P_2(20)c.a. = 1.08 \times 10^{-7}$$

$$k_{eq} = \frac{P_3(20)}{P_{1a}(20) * P_2(20)} = \frac{2.3 \times 10^{-5}}{1.13 \times 10^{-5} * 1.08 \times 10^{-7}} == 6.8 \times 10^6$$