1 Domande sulle banche dati

1. Cos'è Uniprot? Oltre ad una descrizione generale si descrivano i suoi livelli e si elenchino almeno 5 sezioni che si possono trovare nelle pagine delle singole proteine (entry). Si descrivano inoltre brevemente i database PDB e ExPASy. Cosa hanno in comune queste tre risorse?

   - Uniprot è un database rinomato che contiene informazioni sulle sequenze proteiche; si divide in tre livelli:

   (a) UKB (uniprot knowledge base) che contiene le informazioni revisionate e conosciute dalla letteratura biomedica

   (b) Uniprot reference che contiene informazioni su famiglie di proteine raggruppate tramite clustering

   (c) Uniprot Archive che contiene sequenze proteiche non ridondanti stabili

   5 sezioni per una entry sono:

   - Function = riassume informazioni sulla entry ricercata. Elenca i processi molecolari e biologici a cui prende parte la proteina.
   - Structure = contiene informazioni circa la sua struttura secondaria. Sono elencati gli spettri risolti a raggi X e/o NMR contenuti in PDB.
   - Sequence = contiene la sequenza amminoacidica con eventuali isoforme. Sono esplicitate anche le eventuali varianti naturali.
   - Pathology = contiene le informazioni sulle malattie legate a tale proteina
   - Interaction = elenca le interazioni tra subunità con altre proteine

   PDB sta per protein data bank, è un database che contiene files .pdb, più informazioni su struttura 3D di proteine, visulazzione dei ligandi e analisi delle caratteristiche di una proteina.
   ExPASy è un altro database proteico multifunzione. Uno dei tools più usati è translate tool che serve per tradurre dna in proteina.
   Questi database hanno in comune il fatto che contengono informazioni su proteine, e Uniprot possiede link a PDB nella sezione structure (strutture risolte della entry).

2. Si citino e discutano brevemente 4 banche dati presenti in NCBI. Se possibile escludere NCBI protein.

   - I quattro database sono Gene, OMIM e Homologene.

   - Gene: raccoglie sequenze nucleotidiche con particolare enfasi sulla descrizione genica. È orienta to ai geni e ai loci.
   - OMIM: contiene in informazioni sui cosiddetti "disordini genetici" nell'uomo. Raggruppa le informazioni sulle malattie mendeliane umane.

- Homologene: contiene sequenze di geni omologhi nell'uomo.
- PubChem: dedicato ai composti chimici, come inibitori, stimolatori, ecc. come farmaci che agiscono su una proteina o su un gene.

3. Descrivere Uniprot, Pfam, Prosite, CATH e PDB, con particolare attenzione alla più importante tra di esse (qual è?)

   - Possiamo subito dire che sono tutte banche dati proteiche, nello specifico:

   - Uniprot è un database rinomato che contiene informazioni sulle sequenze proteiche; si divide in tre livelli: UKB (uniprot knowledge base) che contiene le informazioni revisionate e conosciute dalla letteratura biomedica; Uniprot reference che contiene informazioni su famiglie di proteine raggruppate tramite clustering; infine Uniprot Archive che contiene sequenze proteiche non ridondanti stabili.
   - Pfam e Prosite sono utili per studiare e catalogare le strutture proteiche, per dedurne proprietà strutturali e di similitudine con altre classi di proteine, per confrontare una proteina di interesse con una classe di proteine che si sospetta essere funzionalmente simile.

     Pfam suddivide le proteine in famiglie strutturali e ne descrive le caratteristiche in base a metodi statistici come allineamenti e Hidden Markov Models, utili per dedurre proprietà simili nelle classi di oggetti (criteri per decidere se una proteina appartiene ad una certa famiglia strutturale oppure no).

     Prosite individua, data una sequenza query, le possibili famiglie di appartenenza e le informazioni relative ai siti conservati e funzionali per poterli confrontare. Determina possibili caratteristiche funzionali, domini, cofattori, siti attivi funzionali per enzimi, amminoacidi strutturalmente importanti, livello di conservazione.
   - CATH è un database che definisce famiglie strutturali, sfruttando un criterio gerarchico di classificazione in famiglie che svolgono una funzione biologica comune. Aiutano a predire le strutture e a caratterizzarlo. L'acronimo individua i 4 elementi gerarchici: CLASSE (contenuto e tipo di strutture secondarie), ARCHITETTURA (descrizzazione dell'orientamento delle strutture secondarie senza tener conto delle connessioni), TOPOLOGIA (tiene conto delle connessioni che caratterizzano le strutture secondarie) e HOMOLOGIA (raggruppa proteine con strutture e funzioni simili).

4. Cos'è Pubmed? Descrivere la banca dati, i suoi contenuti, e gli strumenti messi a disposizione degli utenti (illustrati a lezione), sia per la ricerca che per la gestione dei risultati.

   - Pubmed, accessibile tramite l'interfaccia di accesso ai database ENTREZ, è la banca dati per la letteratura biomedica più completa. Contiene

articoli scientifici peer-reviewed e quindi contiene informazioni solide. Durante le ricerche è possibile filtrare i risultati per anno di pubblicazioni, autori, parole chiave (tramite cui effettuiamo le ricerche), abstract, citazioni ecc. Ogni articolo è identificato dal PMID (ID PubMed).

5. Si descrivano il formato FASTA e il formato XML nell'ambito della bioinformatica.

   Il formato FASTA e il formato XML sono diversi formati di rappresentazione delle sequenze:

   - Il formato FASTA rappresenta mediante testo sequenze nucleotidiche o peptidiche, con sequenze maiuscole. La prima riga è sempre di commento (preceduta da ">") e le linee successive, ciascuna di 80 caratteri, rappresentano la sequenza.
   - Il formato XML (eXtensible Markup Language) replica la struttura logica del record nella banca dati, i tag permettono di delimitare e definire campi e sottocampi, per permetterne una facile lettura da software diversi che lavorano con le sequenze.

2 Domande sugli allineamenti di sequenza

Tutte nella simulazione d'esame.

3 Domande sul confronto di sequenze e Matrici di Sostituzione

Le domande tipiche si trovano nella simulazione d'esame. Aggiungiamo:

- Cosa indica il valore atteso E? Cosa indica il p-value?

  - Il valore atteso E (o indice di incertezza) indica il numero di allineamenti con punteggio $\geq S$ ottenuti per caso sulla ricerca nel database. Esso viene stimato dalla seguente equazione:

$$E = KMNe^{-\lambda S}$$

  Dove: $K, \lambda =$ parametri
  $M =$ lunghezza della query
  $N =$ lunghezza della sequenza nel database
  $S =$score
  Più alto è il valore di E più è probabile che un allineamento sia poco significativo; più basso è E più la probabilità che l'allineamento sia casuale è bassa e l'allimeamento è significativo (non casuale), quindi il valore di E decresce esponenzialmente con l'aumentare di S. Ottenere un allineamento con $E = 1$ significa che esiste un altro allineamento con lo stesso score S che è risultato per caso. La stessa ricerca su un database più

piccolo o più grande, anche se restituisce lo stesso allineamento deve avere un valore E diverso (ciò dipende da k).
Per stimare la probabilità che un certo allineamento sia causale si utilizza il $p - VALUE$ pari a $p = 1 - e^{-E}$. Ovviamente quando $E - VALUE$ è elevato (ci sono molte sequenze allineate casualmente) il $p - VALUE$ sarà elevato quindi casualità elevata. La soglia critica è dell'1%. Se la probabilità supera l'1% abbiamo che la probabilità di ottenere tali allineamenti casualmente è troppo elevata.

4 Domande su BLAST e PSI-BLAST

- Cos'è BLAST? Descrivere il suo algoritmo.

  - BLAST è un tool web-accessibile che permette un confronto rapido tra una sequenza query e il contenuto di un database. È fondamentale per capire la relazione di una sequenza query con altre proteine o sequenze di DNA note. L'algoritmo BLAST (come FASTA) è un approssimazione euristica per l'allineamento locale (trovano soluzioni approssimate ma vicine a quelle ottimali in tempi brevi).
  L'algoritmo si divide in 3 fasi:

  1. Compila una lista di parole di lunghezza W con un punteggio oltre la soglia T.
  2. Scansiona il database per le voci della lista appena compilata.
  3. Quando riesco a trovare una corrispondenza si deve estendere l'allineamento, ricalcolando il punteggio e fermandosi quando diventa inferiore a una certa soglia. Nella versione originale di BLAST ciascun hit è esteso in entrambe le direzioni, nella versione migliorata sono necessari due hit vicini entro una distanza A.

- In BLAST a cosa servono i parametri W,T e A?

  - Il parametro T è la soglia di punteggio nella lista di parole corrispondenti alla query, mentre W è la lunghezza (o dimensione) della parola query. Scegliere w piccoli mi permetterà di avere un numero di hits maggiore ma più sensibilità. Viceversa scegliere w grandi velocizzano la ricerca a scapito della sensibilità. Il parametro A, utilizzato nella versione migliorata di BLAST, definisce la distanza entro cui è necessario trovare due hit perché avvengano le estensioni, questo fa in modo che avvengano meno frequentemente.

- Si discuta Psi-BLAST: principi di base e applicazioni. Quali limiti di BLAST supera? Come?

  - Psi-BLAST è un algoritmo euristico di allineamento locale che permette di effettuare una ricerca più in profondità, rispetto a BLAST, utilizzando una matrice di punteggio adattata dinamicamente. L'algoritmo è diviso in 5 fasi: costruisce un allineamneto multiplo di sequenze a partire

dagli hit migliori e crea quindi un "profilo" detto Matrice di Calcolo Posizione Specifica (PSSM), usato come query nel database per l'iterazione successiva. Il PSSM catturna il pattern di conservazione nell'allineamento multiplo ottenuto dai migliori hits di BLASTP e lo immagazzina sottoforma di matrice di score (dove le posizioni più conservate hanno punteggi più alti e le regioni poco conservate, punteggi più bassi). Il profilo è quindi una specie di nuova query in cui ogni posizione ha un "peso" differenziato: questa informazione può essere utilizzata per estendere la ricerca.
Psi-BLAST supera il limite di BLAST di non trovare proteine omologhe se queste sono troppo distanti dall'ancestrale comune, inoltre riesce a risolvere il problema delle query lunghe (che BLAST classico invece non trova).

5 Allineamenti Multipli di Sequenze

- Cos'è Clustal-W?

  - Clustal-W è la versione moderna di Clustal Omega, un metodo progressivo per determinare come combinare uno per uno allineamenti a coppie, per creare un allineamento multiplo, usando un albero guida. Si compone di 3 fasi:

    1. Realizzare una serie di allineamenti a coppie globali, usando NW, di cui si calcola la distanza in una matrice delle distanze.
    2. Creare un albero guida a partire dalla matrice delle distanze.
    3. Allineare progressivamente le sequenze, prima le più vicine, ovvero più simili, e poi le più distanti.

  Clustal-W (come Clustal Omega) per molte sequenze è parecchio lento perché ha una crescita esponenziale, per cui è preferibile per sequenze lunghe utilizzare altri metodi, per esempio MUSCLE.

- Citare due metodi alternativi a Clustal-W. In cosa differiscono? In che modo lo migliorano?

  - Metodi alternativi a Clustal-W sono i metodi iterativi, come MUSCLE e Praline: differiscono perché consistono nel calcolare una soluzione subottimale e modificarla ripetutamente, con metodi di programmazione dinamica (o altri) fino a quando la soluzione converge.
  Un altro metodo alternativo è TCoffee (che ha un procedimento simile a Clustal-W) che determina un allineamento multiplo, il più poessibile coerente con vincoli esterni (differisce per questo da Clustal-W, garantendo la soddisfazione della necessità di coerenza).

- Cos'è il benchmarking negli allineamenti multipli?

  - Il benchmarking consiste nella valutazione degli algoritmi di allineamento multiplo, tramite l'utilizzo di dataset di test, quindi allineamenti

noti e affidabili. Posso così confrontare le risposte di diversi algoritmi per un determinato set di interesse. Il database più utilizzato per questa funzione è BAliBASE.

6 Reti Neurali

- Come funziona una rete neurale?
  -

Databases

Esercizio 1

Traccia

Scaricare il fasta della sequenza genomica di human hemoglobin subunit beta (NM_000518.5).

1. Visitare il sito di ExPASy (expasy.org):
2. Provare il tools TRANSLATE (resources A..Z) per tradurre automaticamente una sequenza genica in una proteica.
3. Sottomettere la sequenza genomica scaricata
4. Quale frame è corretto (confrontare la sequenza predetta con quella reale `NP_000509.1`)?
5. Perché ci sono 6 frames?

Svolgimento

4. Il frame corretto è il terzo.

```
5'3' Frame 3
ICF-HNCVH-QPQTDTMVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQR
FFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH
VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH-ARFLAVQ
FLLKVPLFPKSNY-TGGYYEGP-ASGFCLIKNIYFHC
```

5. Ci sono 6 frames perché ogni regione del DNA ha sei possibili frame di lettura, tre in ciascuna direzione. Il frame di lettura utilizzato determina quali amminoacidi verranno codificati da un gene.

Esercizio 2

Traccia

Cercare la sequenza nucleotidica e amminoacidica della rodopsina (rhodopsin), il pigmento visivo che innesca la visione nei vertebrati

1. Cominciamo dal database Nucleotide. Quante sequenze ci sono per la ricerca "rhodopsin"?
2. Limitare la ricerca al database RefSeq. Quanti record ci sono?
3. Limitiare la ricerca ad homo sapiens (human), usando l'opzione advanced search. Quante sequenze nucleotidiche trova?
4. Visualizziamo l'entry "`Homo sapiens rhodopsin (RHO), RefSeqGene on chromosome 3`". Quante bp ci sono nella sequenza?
5. Ci sono malattie genetiche associate a questa entry? Di tipo solo autosomico dominante? (OMIM)
6. Scaricare il file della sequenza nucloetidica del gene di rhodopsin

Svolgimento

1. Nel database Nucleotide ci sono 56577 entries per la ricerca "rhodopsin".
2. Selezionando il database RefSeq nella colonna sulla sinistra, da "Reference Database": ci sono 7977 entries.
3. In Advances Search inseriamo con "All Fields" la ricerca "rhodopsin" e nell'AND selezioniamo "Organism" e poi cerchiamo: Homo Sapiens (human), trova 20 sequenze nucleotidiche.
4. Selezionando la prima entry ("`Homo sapiens rhodopsin (RHO), RefSeqGene on chromosome 3`") ci sono 198295559 bp.
5. Apriamo il collegamento in "gene" nella sezione `"FEATURES": /db_xref = "MIM:180380"`
   Arriviamo nella pagina OMIM correlata, ci sono diverse malattie genetiche associate a questa entry, ci sono anche delle malattie autosomiche recessive (`In affected members of 2 Indonesian families segregating autosomal recessive retinitis pigmentosa-4 (RP4; 613731), Kartasasmita et`

al. (2011) identified homozygosity for a 482G-A transition in exon 2 of the RHO gene, resulting in trp161-to-ter (W161X) substitution. Haplotype analysis suggested that this is a founder mutation.)

6. Scarichiamo il file della sequenza nucleotidica dall'apposito link in NCBI.

Esercizio 3

Traccia

Ricercare la proteina "Hemoglobin subunit beta" di Homo sapiens. Filtrare solo i record con RefSeq selezionare il risultato con codice RefSeq `NP_000509.1` (accession).

1. Individuare
   - lunghezza
   - il refseq del trascritto
2. Salvare localmente la sequenza FASTA della PROTEINA
3. Salvare localmente la sequenza FASTA del TRASCRITTO
4. Ci sono SNP? Cos'è un SNP?
5. Ci sono malattie mendeliane note legate a questa proteina?
6. Ci sono strutture legate a questa proteina?
7. Quante risolte per NMR e quante mediante Cristallografia (X-Ray)?

Se vogliamo adesso scaricare la sequenza amminoacidica, della rodopsina (rhodopsin) per l'uomo su quale database dobbiamo andare e quali filtri utilizzare ?

1. Scaricare il FASTA della proteina e salvarlo in una directory locale.
2. Collegarsi ad OMIM sfruttando il link sulla destra. Quanti records si ottengono? Trovare almeno due mutazioni puntiformi associate a retinite pigmentosa.

Svolgimento

1. Ricerchiamo nel database Protein di NCBI, con Advanced Search, `"(Hemoglobin subunit beta[Protein Name] AND "Homo sapiens"[Organism] AND refseq[filter]"` (oppure selezioniamo RefSeq dopo aver effettuato la prima parte della ricerca, nella colonna a sinistra). Selezioniamo la prima (codice RefSeq `NP_000509.1`)

- Ha una lunghezza di 147 aa
- Il refseq del trascritto è NM_000518.5, lo troviamo, nella sezione FEATURES, sotto "gene", nella stringa `"/coded_by="NM_000518.5:51..494""`

4. Nella colonna laterale destra, troviamo il link `SNP` che ci porta agli SNP correlati (identificati dalla ricerca `SNP Links for Protein (Select 4504349)`) Le entry trovate sono 481. Un SNP è un polimorfismo a singolo nucleotide (ovvero un polimorfismo, cioè una variazione, del materiale genico a carico di un unico nucleotide, tale per cui l'allele polimorfico risulta presente nella popolazione in una proporzione superiore all'1%)

5. Nel database OMIM per la stessa entry, come malattie mendeliane note troviamo: Delta-beta thalassemia, Erythrocytosis, Heinz body anemia, Hereditary persistence of fetal hemoglobin, Methemoglobinemia beta type, Sickle cell anemia, Thalassemia, Thalassemia-beta, dominant inclusion-body, resistenza alla Malaria.

6. Collegandosi al link Structure nella colonna destra troviamo 369 strutture legate a questa proteina, dalla sezione filtri possiamo vedere che 313 risolte per X-Ray e una sola per NMR.

Per cercare la sequenza amminoacidica della proteina Rodopsina possiamo sempre collegarci a NCBI Protein, o anche a UniProt, il filtro da usare è `"Homo Sapiens"[Organism]`

2. Ci sono solo 2 record, in quello della retina pigmentosa troviamo le mutazioni puntiformi: T58R, 180380.0004 e T17M, 180380.0006)

Esercizio 4

Traccia

Nel database Uniprot si cerchi la proteina Transferrin receptor (TFR1) per l'uomo (P02786).

1. Quante isoforme ha ?
2. Ha la struttura risolta ? Se si, a partire da quale aminoacido è risolta.
3. Quale è il nome del gene che la codifica (entrare in HGNC)

Esercizio 5

Traccia

Nel database Uniprot si cerchi la proteina Transferrin receptor 2 (TFR2) per l'uomo (Q9UP52).

1. Quante isoforme ha, se ne ha più di una perche ?
2. Ha la struttura risolta ? Se si, a partire da quale aminoacido è risolta.
3. Quale è il nome del gene che la codifica (entrare in HGNC)

Esercizio 6

Traccia

Scaricare le sequenze proteiche del recettore della transferrina (TFR1), ma che abbiano la struttura 3D risolta e formino un complesso con un qualsiasi ligando.

1. Utilizzare il database Protein.
2. Limitare la ricerca solo al database PDB (quelli con struttura risolta).
3. In ricerca avanzata cercare "TFR1" e "complex" in tutti i campi
4. Scegliere una entry specifica
5. In "Display Settings" selezionare "FASTA"
6. In "Send" selezionare "Complete Record" e "File"

Matrici di Punteggio

Traccia

Esercizio 1

Allineare con i 2 algoritmi le sequenze GAATTCAGTTA GGATCGA Per l'allineamento globale (NW) usare la seguente opzione End Gap Penality, settata su True

1. Quale dei 2 algoritmi restituisce l'allineamento con il punteggio maggiore? Perché?

Svolgimento

Con WS troviamo che:

```
# Length: 4
# Identity:       3/4 (75.0%)
# Similarity:     3/4 (75.0%)
# Gaps:           0/4 ( 0.0%)
# Score: 15.0
```

Con NW troviamo che:

```
# Length: 11
# Identity:       4/11 (36.4%)
# Similarity:     4/11 (36.4%)
# Gaps:           4/11 (36.4%)
# Score: -3.5
```

Esercizio 2

Traccia

Allineare con i 2 algoritmi le sequenze GAATTCAGTTA GGATCGA

1. Settando la penalità per apertura (e chiusura) dei gap a 1 con i due algoritimi (NW e WS) cosa cambia? Perché?

Esercizio 3

Traccia

Utilizzando l'algoritmo NW disponibile su: [

`https://www.ebi.ac.uk/Tools/psa/emboss_needle/]`
`(https://www.ebi.ac.uk/Tools/psa/emboss_needle/`

) Allineare la sequenza di calmodulina umana (CALM1) con quella di:

1. Bos taurus (bovina)
2. Arabidopsis thaliana (pianta) (ottenere le sequenze da opportuni database. . .).

   Mantenere i settaggi di default per i gaps, e utilizzare la matrice di score BLOSUM 62)
3. Qual è l'identità di sequenza? Quali reisidui differiscono pur restando simili per prorietà chimico-fisiche?

Esercizio 4

Traccia

Utilizzando l'algoritmo WS per allineamenti locali disponibile su:

[https://www.ebi.ac.uk/
Tools/psa/emboss_water/](https://www.ebi.ac.uk/Tools/psa/emboss_water/)

Allineare la sequenza delle due proteine con codice Uniprot P46065 e P21457

1. Di quali proteine si tratta? Cosa hanno in comune?
2. Qual e' l'identita' di sequenza? E la similitudine? Si puo' trattare di proteine omologhe? Perche'?
3. Identificare una zona in cui l'identita' e' estesa a 8 residui. Che struttura secondaria ha la seconda proteina in quella zona?
4. Se si allinea la prima proteina con P51177 quali sono i punteggi di allineamento?Allineare LOCALMENTE la prima lunga regione senza gaps. Qual e'? E quali sono i nuovi punteggi? Di quali zone di SII si tratta?

Blast

Esercizio 1

Traccia

Eseguire una ricerca tramite blastp su NCBI usando la seguente sequenza di 12 aminoacidi: PNLHGLFGRKTG

1. Metterla in formato FASTA. I parametri di ricerca saranno automaticamente adattati per sequenze corte.
2. Resettare l'interfaccia
3. Attivare l'opzione "Show results in a new window" per poter confrontare i parametri di default con quelli modificati automaticamente.
4. Osservare la sezione "search summary":
5. Qual è il valore di cut-off dell'E-value?
6. Come è cambiata la "word size"?
7. Qual è la matrice di punteggio?
8. Come sono variati i parametri rispetto al defualt e perchè?

Esercizio 2

Traccia

PSI-BLAST - proteina sconosciuta

1. Un campione biologico ha rivelato la presenza della sequenza proteica di origine sconosciuta riportata in:

   `[http://goo.gl/siebf5](http://goo.gl/siebf5)`

2. Si ritiene che debba appartenere alla specie Danio Rerio (zebrafish).

3. Utilizzare PSI-BLAST con i seguenti parametri: RefSeq come database, escludendo i modelli dagli output, limitandosi all'organismo Danio rerio, PAM30 come matrice di score.ù

4. Di che tipo di proteina si tratta? (Guardare se ci sono domini conservati!) Quanti hits ci sono alla prima iterazione? Qual e' l'hit con score piu' basso ed E-value piu' alto? Segnarsi il codice RefSeq. Quante hits hanno score >200

5. Alla seconda ietrazione, qual e' l'hit con score minore? Che E-value ha? E che score ha la proteina con peggior score alla iterazione precedente?Perché?

6. Quante nuove hit compaiono alla terza iterazione?

7. A quale iterazione non vengono piu' aggiunte hits?

Esercizio 3

Traccia

Entrare in BLASTX di NCBI e copiare la sequenza di "dinosauro" "Lost World" come input. [ftp://ftp.ncbi.nlm.nih.gov/pub/FieldGuide/lostworld.txt](ftp://ftp.ncbi.nlm.nih.gov/pub/FieldG
Resettare la pagina prima di impostare i parametri Assicuratevi di includere l'intera sequenza. Ricercare sul database "nr". Escludere i modelli (XM/XP).

1. A quale proteine appartiene probabilmente questa sequenza nucleotidica?

2. Nella pagina dei risultati, guardare i risultati degli allineamenti.

3. La pagina risultante mostrerà la sequenza query scritta come proteina (utilizzando le 20 lettere corrispondenti agli amminoacidi). Il Dr. Mark Boguski che ha creato la sequenza ha lasciato un messaggio nascosto nella sequenza query in posizioni corrispondenti ai 4 gap della sequenza allineata. Qual è il suo messaggio?

MSA

Esercizio 1

Traccia

Nel sito Homologene scaricare le sequenze fasta che ci sono nell'entry relativa alla proteina NP_000940.1 ed allinearle con muscle EBI :

`[https://www.ebi.ac.uk/Tools/msa/muscle/](https://www.ebi.ac.uk/Tools/msa/muscle/)`

(attenzione! Selezionare come output il formato Clustal!)

1. Quante sequenze si stanno allineando?
2. Cosa permette di dire che le sequenze sono in formato FASTA?
3. Quali due delle sequenze non conservano la stringa "ICLI"?
4. Quante e quali inserzioni di un singolo aminoacido sono avvenute e in quali sequenze?
5. Aprire l'allineamento in Jalview dopo averlo esportato in formato FASTA da MUSCLE. Selezionare la regione che si estende da "GQSPPE. . . " a ". . . VRDVQ" della sequenza NP_990185.1, tramite il tab Web Service lanciare JPRED. Qual è l'elemento di struttura secondaria più ricorrente, secondo la predizione di JPRED? Quante alfa eliche sono predette? Suggerimento: Usare HTML format per l'output
6. (se non fosse disponibile dal tab, collegarsi a:

   `[http://www.compbio.dundee.ac.uk/jpred/](http://www.compbio.dundee.ac.uk/jpred/)`

). ATTENZIONE: JPRED può essere lento!!!

Esercizio 2

Traccia

Cerchiamo l'entry 1EBM nel database PDB

1. Quali macromolecole contiene la struttura?
2. Quante catene? Cosa rappresenta la catena A? E' mutata?
3. È una proteina intera? Mancano residui? Perchè?

4. Cliccare sul tab Sequence. Che informazioni troviamo?

Scarichiamo il file PDB e visualizziamolo con un editor di testo (attenti a dove lo salvate!)

1. Chi sono gli autori del lavoro strutturale?
2. Si tratta di cristallografia a raggi X o di NMR?
3. Qual è la risoluzione della struttura?
4. Cosa si trova al REMARK 200? Hanno usato luce di sincrotrone per risolvere la struttura?
5. Cosa si trova al REMARK 470? Spiegate i residui mancanti
6. Cosa si tova nel campo SEQRES?
7. Quante α-eliche e β-sheets ci sono?
8. Trovare le coordinate del carbonio alfa di Asp174

SystemsBiology

Esercizio 1

Traccia

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 cost6ante 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

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

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

*****************MODEL STATE INFORMATION
P1(0)=25e-6   %M (moli/I)
P2(0)=24e-6   %M (moli/I)

*****************MODEL PARAMETERS
k_conf=7e3 \%M^-1 s^-1
k_as=\item4e5 \%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 Sono sufficienti per la prima reazione sono sufficienti, 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 non sono sufficienti per stabilire l'equilibrio dell'intero sistema.

Domanda 2

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

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

Dobbiamo ora calcolare la costante di equilibrio empirica:

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

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

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

$$k_{eq} = \frac{P_3(20)}{P_{1a}(20) * P_2(20)} = \frac{2.3 \times 10^{-5}}{1.13 \times 10^{-5} * .08 \times 10^{-7}} == 6.8 \times 10^6$$

Domanda 3 Per capirlo aggiungiamo la reazione 3, con una nuova costante k_{deg}.
Dopo circa 10^5 secondi (quindi circa 28 ore) abbiamo raggiunto lo zero (più o meno) per P_3.
Spiegazione: dopo la prima reazione, di dissociazione di P_1, P_{1a} tende a dissiparsi, quindi non è possibile dopo le 28 ore che si formi P_3 e quindi poi tende a sparire.

Esercizio 2

Traccia

La proteina P, è sintetizzata dai ribosomi con una costante cinetica kl pari a.$5 \times 10^{-7} M^{-1} s^{-1}$. È noto che la proteina P_1 dimerizza (formando il dimero P_2) con una $K_{Dim}. = 5nM$ e che il dimero può reversibilmente dissociare con una costante cinetica di dissociazione $kdim\_d = 5 \times 10^{-5} s^{-1}$.
Nella stessa cellula, un enzima E lega irreversibilmente un cofattore C con una costante cinetica $kcof = 8.4 \times 10^{4} M^{-1} s^{-1}$ a formare l'enzima attivo E_a. Quest'ultimo catalizza l'attivazione del dimero P_2, trasformandolo quindi in P_{2a}, con una costante di catalisi $kcat = 1.2 \times 10^{-2} s^{-1}$ ed una costante di Michaelis $K_M = 5\mu M$. Il dimero attivato lega poi un recettore intracellulare R in modo reversibile a formare il complesso $P_{2a}R$ con costanti cinetiche di associazione e dissociazione rispettivamente $kRa = 1.3 \times 10^{5} M^{-1} s^{-1}$ e $kRd = 10^{-2} s^{-1}$. Il complesso $P_{2a}R$ dissocia poi in modo irreversibile nel dimeno P_2 inattivo e nel recettore attivo R. con costante cinetica $kRact = 4 \times 10^{-1} s^{-1}$.
Sapendo che le concentrazioni iniziali delle specie molecolari presenti nel sistema sono: $P_1(0) = 1.5\mu M, E(0) = 10.5\mu M, C(0) = 5.3\mu M, R(0) = 320\mu M$, simulare dinamicamente il sistema e rispondere ai seguenti quesiti:

1. Dopo quanto tempo il recettore R è totalmente saturato da P_{2a}? Qual è il rispettivo valore massimo di produzione di R_a, la sua forma attiva?

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

```
*****************MODEL NAME
Esercizio 2
*****************MODEL NOTES

*****************MODEL STATE INFORMATION
P1(0)=1\item5e-6    %M (moli/I)
E(0)=10.5e-6     %M (moli/I)
C(0)=5.3e-6      %M (moli/I)
R(0)=320e-6      %M (moli/I)

*****************MODEL PARAMETERS
```

```
k_1=\item5e-7     %M^-1 s^-1
k_Dim=5e-9     %M^-1 s^-1
kdim_d=5e-5    %s^-1
k_cof=8.4e4   %M^-1s^-1 %alla terza domanda cambia
k_cat=1.2e-2   %s^-1
KM=5e-6       %M
k_Ra=1.3e5
k_Rd=1e-2
k_Ract=4e-1

******************MODEL VARIABLES
kdim_a=kdim_d/k_Dim
VMax=k_cat*Ea

******************MODEL REACTIONS
P1 => P1 : r1 % biosintesi della proteina P1
   vf=k_1
P1 + P1 <=> P2: r2 %dimerizzazione della proteina P1
   vf=kdim_a * P1^2
   vr=kdim_d * P2
E + C => Ea : r3 %enzima lefa cofattore irreversibilmente e diventa attiva
   vf=kcof*E*C
P2 => P2a :r4 %attivazione enzimatica di P2
   vf=VMax*P2)/(P2+KM)
P2a + R <=> P2aR :r5  %il dimero attivato
   vf=kRa * P2a * R
   vr=kRd*P2aR
P2aE => P2 + Ra :r6 %di
   vf=kRact*P2aR

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

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

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

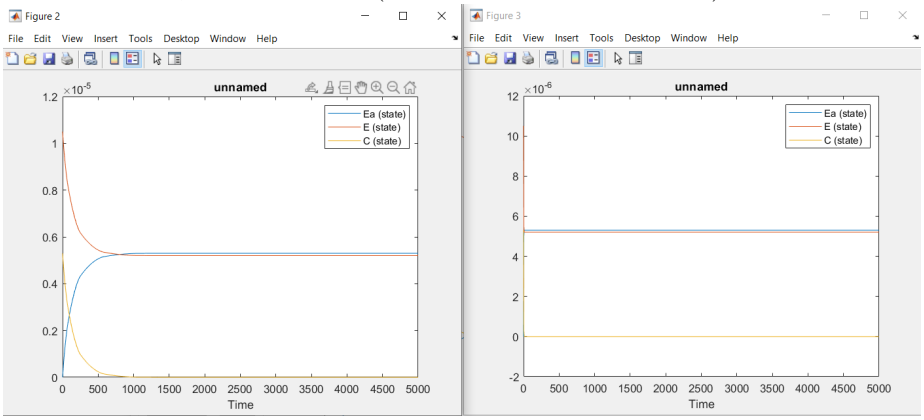
Risposte

Domanda 1 Dopo i 5500 secondi il reagente R è completamente saturato.

Domanda 2 Decresce più velocemente E.

Domanda 3 Le due figure differiscono, perchè con 10^2 su si vedono bene le curve con cui aumentano e diminuiscono i reagenti sui 5000 secondi; inizialmente con 10^4 non si nota quasi, perché troppo veloce per essere visualizzata bene.

La misura di E, C e Ea cambia (ma non abbiamo visto come).



Domanda 4 Per vederlo abbiamo runnato per 400 secondi:
La concentrazione di Ra è circa 2.1e-5, mentre quella mutata è 14.5e-6.
Calcoliamo la percentuale di Ra mutata relativamente alla quantità normalmente prodotta: dopo 400 secondi la proteina mutata è il 70% di quella non mutata.

Domanda 5 Raddoppiamo R(0) per verificare la differenza con Wild Type in caso di overespressione e poi lo dimezziamo per vedere una down-regolazione:
In 2000 secondi si nota sui grafici la differenza: